POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bayesian distance-based clustering of areal data

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Filippo Galli, 10725182

Advisor:
Prof. Alessandra Guglielmi

Co-advisors:
Matteo Gianella

Academic year:
2025-2026

Abstract: This thesis develops and empirically validates a scalable Bayesian nonparametric framework for clustering areal data based on pairwise distances between kernel density estimates. Motivated by the computational intractability of state-of-the-art spatial Bayesian models — which require hours of computation even for moderate geographic scales — we propose a strategy that transforms the inferential problem from tens of thousands of individual observations to a small number of area-specific density representations and their pairwise distances. Building upon the distance-based cohesion-repulsion likelihood of Natarajan et al. (2024), we combine a Normalized Generalized Gamma (NGG) process induced prior for fully data-driven inference on the number of clusters with extensions that incorporate spatial adjacency structure and auxiliary demographic covariates. An efficient MCMC sampling scheme synthesizes the Sequential Allocated Merge-Split sampler, a distance-based adaptation of Locality-Sensitive Sampling for observation selection, a Smart-Dumb-Dumb-Smart move-type selection strategy, periodic Gibbs scans, and shuffle moves. The framework is validated on simulated data and applied to two real datasets: California census income data aggregated over 93 Public Use Microdata Areas, where inference completes in under two seconds compared to ten hours for existing approaches in the field of boundary detection, and Italian municipalities income data comprising 7,903 areal units. Results demonstrate competitive clustering accuracy, interpretable socioeconomic groupings, and substantial computational gains across all model configurations. The entire framework is implemented as a high-performance, modular C++ library with R bindings, publicly available at https://github.com/Filippo-Galli/BNPClust.

Key-words: Bayesian nonparametric clustering, areal data, distance-based likelihood, Normalized Generalized Gamma Process, kernel density estimation, spatial clustering, MCMC, Product Partition Model

1. Introduction

The explosion of data across different disciplines has fundamentally reshaped research methodologies. From sociology to chemistry, from physics to economics, the ability to extract meaningful insights from massive datasets has become essential. However, the fundamental challenge remains unchanged: analyzing increasingly large datasets within practically acceptable computational timeframes. This temporal bottleneck has become a critical limitation, restricting the applicability of sophisticated statistical models to real-world problems where timely decision-making is crucial.

In the Bayesian statistical framework, computational scalability represents one of the most pressing challenges. As sample sizes grow, the runtime requirements for inference—particularly for Markov chain Monte Carlo (MCMC) methods—often grow prohibitively, rendering theoretically sound approaches impractical for real applications. This thesis directly addresses the scalability problem in the specific setting of clustering complex areal data by developing and empirically validating a computational framework designed to efficiently compute cluster estimates for large-scale spatial data within a computationally manageable time frame.

The motivation for this work emerges from a concrete policy application: identifying spatial patterns in socioeconomic data to inform targeted interventions at the local level. The practical scenario involves clustering areal population units, specifically Public Use Microdata Areas (PUMAs), to detect regional boundaries and homogeneous regions that share similar socioeconomic characteristics. Understanding these patterns allows policymakers to identify where interventions are most needed and to design geographically-informed policies.

The computational barrier becomes evident when examining Bayesian state-of-the-art models for to address this question. The SPMIX framework of Gianella, Beraha, et al. (2023) is within the current standard methodology of spatial boundary detection for areal data. In this work, the authors develop a sophisticated Bayesian model to detect boundaries between adjacent areal units if corresponding income densities show significant differences. Their approach models income densities as a finite Gaussian mixture with a random number of components using a logistic multivariate conditionally autoregressive (CAR) prior on the weights of the mixture to induce spatial dependencies.

Yet even with this sophisticated approach, analyzing the 80,000 observations grouped into 93 PUMAs covering Los Angeles required ten hours of computation on a machine equipped with an Intel i7-1255U @ 4.700 GHz processor and 32 GB of RAM. This computational demand severely limits the method's practical utility: repeated analyses under different model specifications, sensitivity analyses, or applications to larger geographic regions become prohibitively expensive. Such constraints effectively prevent Bayesian researchers from rapidly exploring different scenarios or scaling the analysis to state or national levels.

This thesis bridges the gap between statistical methodology and practical feasibility by exploring how strategic data aggregation, from individual observations to more complex objects, can reduce computational complexity while preserving statistical interpretability and correctness. Rather than accepting computational limitations as constraints, we propose to describe observations using their kernel density estimations that substantially reduce data numerosity and, consequently, computational burden.

Once decided to work with density data, to perform clustering of areas we opt to analyze pairwise distances between area-specific densities instead of working with the densities directly. This choice is motivated by two main reasons: first, distances are often easier to compute and store than full density objects, especially in high-dimensional settings; second, distance-based approaches naturally accommodate complex data structures and can leverage a wide range of distance metrics tailored to specific applications. This philosophy aligns with foundational work in Bayesian clustering that emphasizes pairwise information over raw high-dimensional coordinates (Lau et al. 2007).

A foundational reference for the proposed methodology is the work of Duan et al. (2021). These authors introduce a distance-based likelihood framework for clustering arbitrary objects using pairwise distances, modeling each cluster as a cohesive group defined by small intra-cluster distances. Their primary contribution lies in formulating a likelihood that features an intra-cluster cohesion term, which promotes compactness within clusters while appropriately addressing the dependencies inherent in pairwise distances.

Specifically, the authors employ a Gamma distribution to model the likelihood of distances, with a calibration power of $\frac{1}{n_j}$, where n_j denotes the number of members in cluster j. This calibration correct the discrepancy between the number of pairwise distances $\frac{n_j(n_j-1)}{2}$ and the number of effective independent distances $n_j - 1$. This principled calibration strategy, grounded in information-theoretic analysis via Bregman divergences, ensures exchangeability and provides a formal link between distance-based and model-based clustering approaches.

The connection between distance-based methods and probabilistic modeling has been formalized

more recently in the decision-theoretic framework. Rigon et al. (2023) demonstrate that algorithmic clustering methods such as k-means can be recovered as maximum a posteriori (MAP) estimates within probabilistic models where the likelihood is defined on distances rather than on the raw data. Their generalized Bayes framework replaces the traditional log-likelihood with a loss function scaled by a temperature parameter, enabling uncertainty quantification in loss-based clustering. However, their approach—like many earlier methods—requires the number of clusters K to be pre-specified, limiting its applicability in settings where cluster structure is unknown a priori.

Building upon this foundation, Natarajan et al. (2024) extend the distance-based likelihood framework by introducing a repulsion term that penalizes small distances between observations in different clusters. This dual structure—combining cohesion (within-cluster similarity) and repulsion (between-cluster separation)—enhances cluster separation and improves clustering performance, particularly in challenging scenarios with overlapping clusters. Critically, the repulsion mechanism enables K to become an object of Bayesian inference rather than a fixed hyperparameter, addressing a fundamental limitation that persists across both classical distance-based methods and recent probabilistic reformulations.

Our contribution builds directly upon the methodological advances of Natarajan et al. (2024), adapting and extending their distance-based clustering framework to the specific challenges of areal data analysis. We apply their approach to cluster kernel density estimations of personal income distributions at the Public Use Microdata Area (PUMA) level, an aggregation strategy that dramatically reduces the inferential problem dimensionality from approximately 80,000 individual observations to 93 kernel density estimates and their pairwise distances. This dimensionality reduction enables efficient Bayesian inference on manageable sub-problems while preserving the essential distributional information needed for spatial clustering.

To enhance the foundational distance-based framework, we incorporate three key methodological extensions. First, we adopt a flexible nonparametric prior on the random partition of areas induced by a sample from the Normalized Generalized Gamma (NGG) process, allowing for fully data-driven inference on the number of clusters K without pre-specification. Then, we integrate spatial adjacency information and auxiliary covariates directly into the prior specification, strengthening the model's ability to capture relevant dependencies inherent in areal data contexts. Finally, we develop an efficient MCMC sampling scheme that combines advanced split-merge proposals with locality-sensitive observation selection and data-aware move type selection, enhancing computational efficiency and mixing behavior. These methodological innovations collectively enable scalable Bayesian clustering of areal data based on pairwise distances between kernel density estimates.

The structure of the thesis proceeds as follows: Section 2 presents three datasets—income data from the California census by the US Census Bureau, its extension to all U.S. states, and income data for Italian municipalities by ISTAT. Section 3 describes the methodological framework, extending the distance-based clustering model of Natarajan et al. (2024) to incorporate areal data structures, a brief description of the Normalized Generalized Gamma Process prior, spatial and covariate information and a brief introduction to the MCMC sampling scheme used. In Section 4 we present a simulation study to validate the model's ability to recover known cluster structures under controlled conditions, demonstrating the robustness of the clustering performance and the computational efficiency of the proposed MCMC algorithm. Section 5 and Section 6 present the posterior inference on the California census income and Italian municipalities datasets, respectively, demonstrating the computational gains and clustering robustness within different models. Finally, Section 7 discusses the limitations, and directions for future work.

All the code as C++ library with the R bindings, technical documentation, and examples are available at: https://github.com/Filippo-Galli/BNPClust.

2. Motivating examples

We apply our method to three datasets that span different scales of complexity and data dimensionality. This section introduces the datasets, each picked up to test specific aspects of scalability and methodological robustness. The first two datasets contain data from of the American Community

Survey (ACS) census, while the third one is composed of income data on the Italian municipalities census data provided by ISTAT.

2.1. California census income dataset: the benchmark case

The Los Angeles area serves as our primary motivating example and benchmark for algorithm development. This dataset includes 93 Public Use Microdata Areas (PUMAs) as shown in the panel of Figure 1. The area spans Los Angeles, Ventura, and Orange counties, with individual-level personal income data collected through the American Community Survey (ACS) in the 2020. Personal income values are log-transformed to account for the right-skewed nature of income distributions.

Within each PUMA, we compute kernel density estimations (KDE) using the R command `density` to obtain smooth, area-specific income densities. These densities serve as the fundamental objects for clustering: rather than operating on 80,000 raw observations, our framework aggregates the data into 93 kernel density estimations, each representing the economic profile of a PUMA. To obtain pairwise distances between these density objects, we compute the Wasserstein distance (Villani (2003) and Panaretos et al. (2019)) and the Jeffrey divergence (Jeffreys 1946), providing complementary measures of distributional dissimilarity that capture both overall shape and tail behavior.

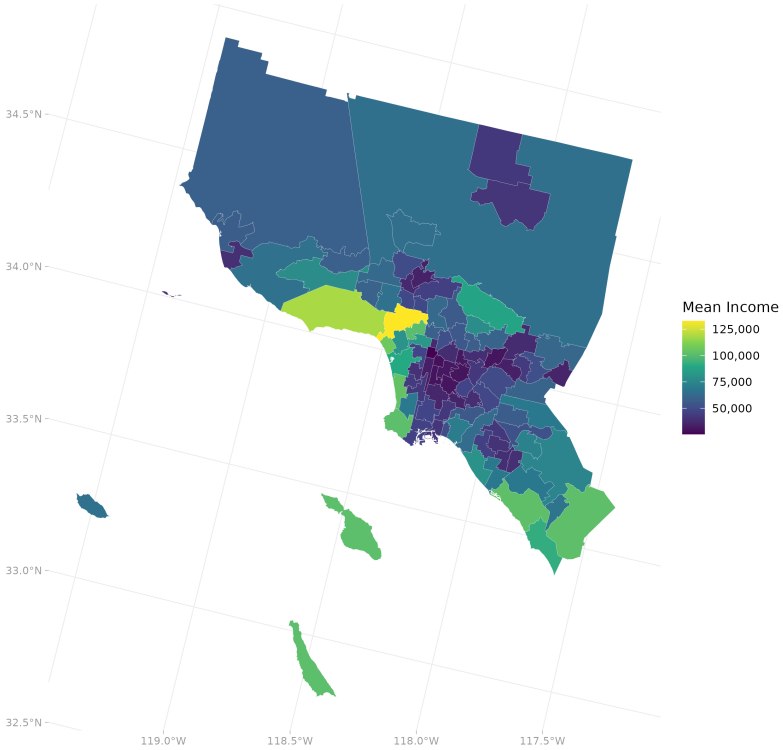


Figure 1: Average income for each areal unit across the dataset described in Section 2.1.

2.2. United States census income dataset: scaling to national coverage

As a second dataset, we extend the analysis of the previous dataset from California to the entire United States by extracting income data for all 2,351 PUMAs as shown in Appendix A Figure 24. This national-scale analysis serves two purposes: first, it allows us to evaluate whether our proposed framework can handle realistic policy-relevant scales without prohibitive computational cost; and second, it provides a validation opportunity across geographically diverse regions with varying demographic and economic characteristics. Heterogeneity across different states and metropolitan areas tests the robustness of the clustering methodology to diverse regional economic structures.

2.3. Italian municipalities

To push the scalability limits of our framework, we analyze data from all Italian municipalities from the ISTAT 2020 census, comprising approximately 7,903 areal units as detailed in Appendix A Figure 25. The Italian data present a different structure compared to the ACS: rather than individual-level income observations, the ISTAT source provides aggregated income information at the municipality level in the form of histograms with seven income ranges. This shift from individual to aggregated data necessitates a slightly modified preprocessing pipeline: instead of kernel density estimation applied to individual values, we work directly with the provided income histograms as our density representations. Distances are computed between these municipality-level income distributions using the discrete version of both the Jeffrey divergence described in Jeffreys (1946) and the 1-Wasserstein distance of Villani (2003).

3. Bayesian distance-based clustering model

This section presents the Bayesian nonparametric model that underpins our scalable clustering framework. We build an areal Product Partition Model with Covariates (aPPMx) where clustering of the areas, and more in general posterior inference, is obtained from data which are pairwise distances between kernel density estimations of incomes. Our model use the distance-based likelihood proposed by Natarajan et al. (2024) combined with a PPM prior derived from a sample of a Normalized Generalized Gamma (NGG) process for flexible partitioning. Covariates and spatial adjacency information are incorporated through prior modifications to enhance clustering performance in areal data contexts.

3.1. Distance-based pseudo-likelihood

The likelihood formulation of our model builds directly upon the distance-based likelihood framework introduced by Natarajan et al. (2024). While Duan et al. (2021) established the partial likelihood paradigm for distance-based clustering, Natarajan et al. (2024) extended it through the addition of a repulsion term that penalises small inter-cluster distances. This innovation mitigates a critical identifiability limitation in the antecedent approach: the challenges in estimating the number of clusters K based solely on cohesion. This results in a dual-component likelihood that combines a likelihood cohesion factor (LCF) to promote intra-cluster similarity with a likelihood repulsion factor (LRF) to enforce inter-cluster separation, thereby improving cluster recovery and inference on K. The likelihood is formally specified as

$$\mathcal{L}(D \mid \boldsymbol{\theta}, \boldsymbol{\lambda}, \rho_n) = \left[\prod_{k=1}^{K} \prod_{\substack{i,j \in C_k \\ i<j}} f(D_{ij} \mid \lambda_k) \right] \cdot \left[\prod_{(k,t)\in A} \prod_{\substack{i \in C_k \\ j \in C_t}} g(D_{ij} \mid \theta_{kt}) \right], \tag{1}$$

where $D = \{D_{ij}\}_{i,j=1,\ldots,n}$ denotes the $n \times n$ pairwise distance matrix derived from the n areal kernel density estimates, $\rho_n = C_1, \ldots, C_K$ represents a partition of the n areal units into K clusters, $f(\cdot \mid \lambda_k)$ and $g(\cdot \mid \theta_{kt})$ are the respective probability densities for intra- and inter-cluster distances, and $A = \{(k,t) : 1 \leq k < t \leq K\}$ enumerates all distinct cluster pairs.

This likelihood encompasses two complementary components. The LCF, the initial product in (1), favours small intra-cluster distances by assigning elevated density to compact groupings, thereby fostering homogeneous clusters. Conversely, the LRF, the second product, imposes penalties on small inter-cluster distances to promote distinct boundaries between clusters. Collectively, these elements yield a likelihood that harmonises internal cohesion and external repulsion, facilitating simultaneous inference on cluster assignments and K.

For computational convenience and interpretability, we specify f and g as gamma distributions:

$$f(d \mid \lambda_k) = \text{Gamma}(\delta_1, \lambda_k), \tag{2}$$
$$g(d \mid \theta_{kt}) = \text{Gamma}(\delta_2, \theta_{kt}), \tag{3}$$

where $\delta_1, \delta_2 > 0$ are fixed shape parameters. The shape parameters control the cluster separation behavior: setting $\delta_1 < 1$ produces a decreasing density that encourages small intra-cluster distances, while $\delta_2 > 1$ produces a density favoring larger inter-cluster distances. This asymmetry in shape parameters naturally encodes the clustering objective.

We complete the likelihood specification with conjugate gamma priors on the rate parameters:

$$\lambda_k \sim \text{Gamma}(\alpha, \beta), \quad k = 1, \ldots, K, \tag{4}$$

$$\theta_{kt} \sim \text{Gamma}(\zeta, \gamma), \quad (k,t) \in A = \{(k,t) : 1 \leq k < t \leq K\}, \tag{5}$$

where $(\alpha, \beta, \zeta, \gamma)$ are hyperparameters specified by the user. The choice of conjugate gamma priors for the rate parameters ensures analytical tractability: the gamma priors are conjugate to the gamma likelihoods, allowing us to analytically marginalize over λ_k and θ_{kt}. This yields an integrated likelihood that depends only on the distance matrix D and the partition ρ_n, substantially simplifying posterior computation. This marginalization strategy is crucial for computational efficiency when iterating over partition proposals during MCMC sampling.

3.2. The Normalized Generalized Gamma Process

To enable clustering with a data-driven number of components, we place a normalized generalized gamma (NGG) process on a random probability measure P over the space of cluster parameters. The marginal prior on the random partition ρ_n is then induced by the ties of samples $\theta_1, \ldots, \theta_n \mid P \overset{\text{i.i.d.}}{\sim} P$ with $P \sim \text{NGG}(a, \tau, \sigma)$. The NGG process is a normalized random measure (NRM), obtained by normalizing the total mass T of an underlying generalized gamma completely random measure (CRM) $\mu = \sum_{j=1}^{\infty} J_j \delta_{X_j}$ with Lévy measure of the form $\rho_{a,\sigma,\tau}(ds, dy)\mu_0(dy) = \frac{a}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-\tau s}\, ds\, \mu_0(dy)$, where $a > 0$ is a mass parameter, $\tau > 0$ a rate parameter, $\sigma \in (0,1)$ a discount parameter, and μ_0 a base probability measure.

James et al. (2009) provide a posterior characterization of NRMs via an auxiliary latent variable U, which facilitates tractable inference. Specifically, conditional on the total mass T, $U \mid T \sim \text{Gamma}(n, T)$; marginalizing over T and U yields the prior partition probabilities. Conditional on U, the predictive probability of a new cluster is proportional to $a(U+\tau)^{\sigma}$, interpreting U as a random concentration parameter that controls the induced number of clusters.

The exchangeable partition probability function (EPPF) of the NGG process is

$$P[\rho = \rho_n, U \in du] = \frac{a^{|\rho_n|} u^{n-1}}{\Gamma(n)(u+\tau)^{n-\sigma|\rho_n|}} \mathrm{e}^{-(\mathrm{a}/\sigma)[(\mathrm{u}+\tau)^{\sigma}-\tau^{\sigma}]}\, \mathrm{du} \cdot \prod_{\mathrm{c}\in\rho_{\mathrm{n}}} \frac{\Gamma(|\mathrm{c}|-\sigma)}{\Gamma(1-\sigma)}, \tag{6}$$

where ρ_n denotes a partition of $\{1, \ldots, n\}$, $|\rho_n|$ the number of clusters, and $|c|$ the size of cluster $c \in \rho_n$ as described in Favaro et al. (2013).

The parameter $a > 0$ is a mass parameter controlling the average number of clusters, with larger values favoring more numerous clusters. The parameter $\tau > 0$ is a rate parameter that scales the intensity of the underlying Lévy measure as described in Favaro et al. (2013), controlling the clustering behavior of the NGG process in a manner that is redundant with the parameter a. The parameter $\sigma \in (0,1)$ is the discount parameter controlling the variance of the number of clusters; larger values of σ induce higher variance in the cluster count and weaker rich-get-richer effects, where clusters with larger sizes are increasingly favored to grow. An important special case arises when $\sigma \to 0$ and $\tau \to 1$, under which the NGGP reduces to the Dirichlet Process with mass parameter a. This relationship establishes the Dirichlet Process as a limiting case of the more flexible NGGP family. A detailed exposition of the NGGP, including theoretical properties and posterior inference, is provided in Lijoi et al. (2007), Favaro et al. (2013) and Argiento et al. (2016).

3.3. Covariates and spatial information

In applications involving areal data, neighboring regions often exhibit similarity in the outcomes of interest. Furthermore, auxiliary covariates frequently provide valuable information for clustering. We

incorporate both types of information through prior modifications.

3.3.1 Spatial adjacency

Geographic adjacency structures are incorporated as a correction to the base partition prior. Following the approach in Cremaschi et al. (2023) and detailed in Gianella, Quintana, et al. (2026), the adjacency-weighted prior takes the form:

$$\pi(\rho_n \mid W) \propto \pi_0(\rho_n) \cdot \exp\left(\lambda_s \sum_{h=1}^{K} \|W_h\|\right) \tag{7}$$
$$= \int \frac{a^{|\rho_n|} u^{n-1}}{\Gamma(n)(u+\tau)^{n-\sigma|\rho_n|}} \exp\left(-\frac{a}{\sigma}\left[(u+\tau)^\sigma - \tau^\sigma\right]\right) du \cdot \prod_{c \in \rho_n} \frac{\Gamma(|c|-\sigma)}{\Gamma(1-\sigma)} \cdot \exp\left(\lambda_s \sum_{h=1}^{K} \|W_h\|\right)$$

where $\rho_n = \{C_1, \ldots, C_K\}$, W is the spatial adjacency binary matrix, $\|W_h\|$ counts the number of adjacencies within cluster h, and $\lambda_s > 0$ is a tuning parameter controlling spatial similarity strength.

The exponential term in (7) induces a preference for spatially cohesive clusters. Specifically, larger values of λ_s encourage the formation of clusters composed of spatially contiguous areal units by assigning higher prior probability to partitions where geographically adjacent areas are grouped together. This modification is motivated by the empirical observation that spatial proximity frequently correlates with outcome homogeneity in areal data applications.

3.3.2 General covariates

Auxiliary covariates are incorporated through a similarity-weighted prior extension as in the PPMx prior in Müller et al. (2011). If auxiliary covariates $\underline{x} = \{x_1, \ldots, x_n\}$ are available, the covariate-informed prior becomes:

$$\pi(\rho_n \mid \underline{x}) \propto \pi(\rho_n \mid W) \cdot \prod_{h=1}^{K} g(x_h^*), \tag{8}$$

where $x_h^* = \{x_i : i \in C_h\}$ collects the covariates for all observations in cluster h.

The similarity function $g(x_h^*)$ represents the marginal likelihood under an auxiliary probability model $q(\cdot \mid \cdot)$, which treats the covariates as informative without necessarily assuming they are random at the observation level. Formally,

$$g(x_h^*) = \int \prod_{i \in C_h} q(X_i \mid \zeta_h)\, q(\zeta_h)\, d\zeta_h, \tag{9}$$

where ζ_h are cluster-specific latent parameters governing the covariate distribution within cluster h. The specification of the auxiliary model q depends on the nature and structure of the available covariates. By marginalizing over ζ_h, the function $g(x_h^*)$ captures the likelihood of observing the cluster-specific covariate configuration and automatically upweights partitions where observations within clusters have coherent covariate patterns.

3.4. MCMC sampling scheme

Since our model could marginalize out the component parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ thanks to the conjugacy structure, we design an MCMC algorithm that directly targets the posterior distribution of the partition ρ_n of the n areas given the observed distance matrix D. So, we need to sample the partition of ρ_n and also the parameter U needed by (6).

For what concern the parameter U we rely on a Random Walk Metropolis-Hastings with adaptive proposal variance based on the Robbins-Monro process described in Garthwaite et al. (2016). To sample allocations i.e., to sample ρ_n, the MCMC algorithm combines split-merge proposals for global moves across the partition space with periodic Gibbs updates for local refinement. The Gibbs component

employs Algorithm 3 from Neal (2000), which marginalizes out component parameters to focus inference solely on cluster assignments. Every 25 split-merge iterations, a full Gibbs scan using this algorithm ensures local exploration and mixing. The split-merge sampler represents an advanced synthesis of recent methodological developments. Rather than the Restricted Gibbs Merge-Split sampler of Jain et al. (2004), we adopt the Sequential Allocated Merge-Split (SAMS) sampler of Dahl et al. (2022), which employs sequential allocation during split proposals for improved computational efficiency and mixing behavior. Observation selection for split-merge proposals follows a modified version of the Locality-Sensitive Sampling (LSS) strategy of Luo et al. (2018), which assigns probability weights to observations based on distance-based proximity structure as described in Algorithm 1. By selecting observations stochastically according to these weights, LSS improves proposal quality and acceptance rates substantially compared to uniform random selection, while maintaining that all observations retain positive selection probability, ensuring irreducibility.

Algorithm 1 Local-Sensitive index selection

Require: Distance matrix D, current partition ρ_n, boolean indicating similarity or dissimilarity choice

1: Select initial observation i uniformly at random from $\{1, \ldots, n\}$
2: Compute weights $w_{i,j} = f(D_{i,j})$ for all $j \neq i$, where f yields similarities or dissimilarities
3: Normalize: $p_{i,j} = w_{i,j} / \sum_{k \neq i} w_{i,k}$ for $j \neq i$
4: Sample $j \neq i$ from $\{p_{i,j}\}_{j \neq i}$

Ensure: Indices i and j

As described in Luo et al. (2018), there is freedom in selecting the observation probability weighting scheme. In their work, they employed the likelihood to weight observations. However, since our likelihood scales as $\mathcal{O}(n^2)$—which is computationally expensive—we leverage the distance-based proximity measures already computed within our framework. Our approach proceeds as follows: the initial observation i is selected uniformly at random; subsequently, we employ two alternative transformations to convert pairwise distances into similarity or dissimilarity measures. In the first approach, we define the similarity measure as $1/d_{i,j}$ and the dissimilarity measure as $d_{i,j}$. This approach demonstrating some limitations as numerical stability and sensitivity to the scale of the distance which, combined, reduced the ESS. The second approach applies a logarithmic transformation: the similarity measure is computed as $\log(\max_{j \neq i} d_{i,j}) - \log(d_{i,j})$, while the dissimilarity measure is $\log(d_{i,j})$. Both transformations enable stochastic observation selection without explicitly computing the likelihood, thereby preserving the computational efficiency advantages of the distance-based framework.

Move type selection employs the Smart-Dumb-Dumb-Smart (SDDS) scheme of Wang et al. (2015), detailed in Algorithm 2. There are two variants each of split and merge moves: dumb and smart. For splits, the dumb variant random partitions with equal probabilities of the observations between the two clusters, whereas the smart variant employs SAMS sampling (Dahl et al. 2022). For merges, the distinction lies in the reverse probability computation: the dumb merge assumes a SAMS split of the merged cluster, while the smart merge uses a uniform split probability.

We additionally incorporate the shuffle move of Martínez et al. (2014), which reassigns observations within existing clusters without altering cluster cardinality, enabling fine-grained refinement of cluster compositions that complements the coarser global split-merge transitions.

4. Simulated data

To validate the correctness of the proposed model, we first test on three simulated datasets introduced in Natarajan et al. (2024). Since the simulated data do not include auxiliary covariates, we evaluate the model under two configurations using the prior (6) with and without spatial adjacency information to assess the impact of spatial structure on clustering performance in a synthetic setting.

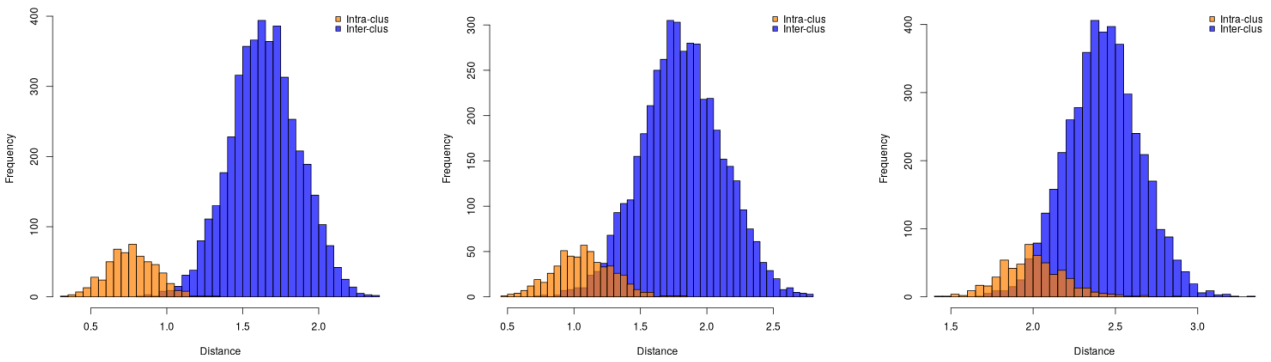
Each simulated dataset contains $n = 100$ observations drawn from a mixture of $K = 10$ multivariate Gaussian kernels. The cluster centers are fixed at the vertices of the standard d-dimensional simplex, and observations within each cluster are generated with identity covariance matrices scaled by $\sigma^2 I_d$. We evaluate three increasing difficulty scenarios, parameterized by $(\sigma, d) \in \{(0.18, 10), (0.25, 10), (0.2, 50)\}$:

Algorithm 2 Smart-Dumb-Dumb-Smart move selection

With probability 0.5, select Smart-split-Dumb-merge; else Dumb-split-Smart-merge
if Smart-split-Dumb-merge **then**
    Select i, j via dissimilarities
    if i, j same cluster **then**
        SmartSplit
    else
        DumbMerge
    end if
else
    Select i, j via similarities
    if i, j same cluster **then**
        DumbSplit
    else
        SmartMerge
    end if
end if

higher values of σ and d increase the overlap between intra- and inter-cluster distance distributions, creating progressively more challenging clustering scenarios. To evaluate the spatial extension, we construct an artificial adjacency matrix W using the eight-nearest neighbors of each observation in the Euclidean space. This allows us to test whether the spatial homogeneity factore (see (7)) improves clustering performance when spatial proximity information is available, even in this synthetic setting where spatial structure is artificially imposed. All experiments use the parameter initialization from Natarajan et al. (2024). For the prior (6) inducing the PPM, we fix $a = 1$, $\tau = 1$, $\sigma = 0.1$, and the spatial coefficient $\lambda_s = 1$. Each MCMC chain is run for 20,000 iterations with the first 5,000 iterations discarded as burn-in.

In Figure 2, we visualize the distance distributions for each scenario, plotting histograms of intra-cluster and inter-cluster distances. This distances will be used as data by our model to perform clustering. The leftmost panel corresponds to the easiest scenario $(\sigma, d) = (0.18, 10)$, where intra- and inter-cluster distance distributions are well-separated. The middle panel shows the moderate-difficulty scenario $(\sigma, d) = (0.25, 10)$, where the distributions begin to overlap. The rightmost panel depicts the most challenging scenario $(\sigma, d) = (0.2, 50)$, where the distributions are heavily overlapping, illustrating the increasing difficulty of clustering as σ and d increase.



$(\sigma, d) = (0.18, 10)$ $(\sigma, d) = (0.25, 10)$ $(\sigma, d) = (0.2, 50)$

Figure 2: Histograms of distances used by our model for the simulated datasets under three scenarios of increasing difficulty. Each panel shows the distribution of intra-cluster (blue) and inter-cluster (orange) distances, illustrating the increasing overlap as σ and d increase.

Table 1 reports clustering performance measured by the Adjusted Rand Index (ARI) and Normal-

ized Mutual Information (NMI), which quantify agreement between the estimated and true cluster assignments for the three simulated datasets, higher values indicate better agreement between the posterior inference and the ground truth. The NGG process derived prior achieves perfect recovery in the easiest scenario ($(\sigma, d) = (0.18, 10)$) but degrades as difficulty increases. Incorporating spatial adjacency information substantially improves performance in the challenging high-dimensional scenario ($(\sigma, d) = (0.2, 50)$, ARI = 0.82). The reference Natarajan et al. model achieves superior performance in the highest-dimensional scenario (ARI = 0.90), suggesting that its distance-based approach may better exploit the specific geometry of simplex-centered clusters. These comparisons establish that our framework achieves competitive clustering accuracy while enabling flexible incorporation of spatial structure.

Configuration	$(\sigma, d) = (0.18, 10)$	$(\sigma, d) = (0.25, 10)$	$(\sigma, d) = (0.2, 50)$
NGG only	1.00 (1.00)	0.90 (0.94)	0.65 (0.69)
NGG + Spatial adjacency	1.00 (1.00)	0.90 (0.94)	0.82 (0.81)
Natarajan et al. (2024)	1.00 (1.00)	0.85 (0.90)	0.90 (0.94)

Table 1: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) in parentheses between estimated and true cluster assignments for three simulated datasets under different model configurations.

5. California census income dataset

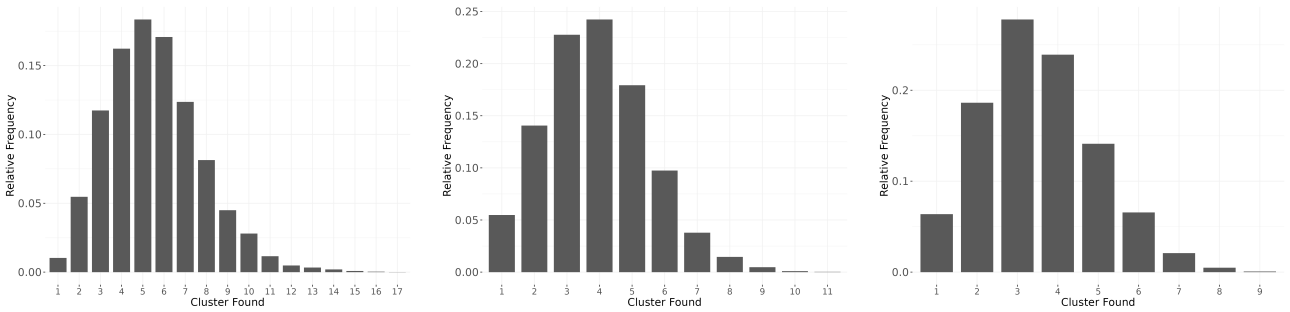
This section presents the empirical results obtained from the analysis of the California census income dataset, previously introduced in Section 2.1. The first part of the study investigates how the prior distribution of the number of clusters varies when spatial and demographic covariates are incorporated into the prior (6). Subsequently, we examine the a priori behavior of the fully specified model under different values of the parameter σ. Three model configurations based on the prior (6) are considered: (i) a baseline specification, (ii) an extension incorporating spatial adjacency information, and (iii) a further augmentation including demographic covariates.

5.1. Prior sensitivity to covariates

Using the hyperparameter initialisation adopted in the simulated-data experiments (Section 4)—namely $a = 1$, $\tau = 1$, and $\sigma = 0.1$—we illustrate the induced prior distribution of the number of clusters under three model specifications. The baseline model, denoted NGGP, relies solely on the prior in (6) and serves as a reference that isolates the effect of incorporating geographic and demographic information. The NGGPW specification extends the baseline by including spatial adjacency through the modification in (7), which penalises disconnected components and thus encourages spatially contiguous clusters. Finally, NGGPWx further enriches the prior by introducing demographic-similarity weights derived from American Community Survey (ACS) covariates for the 93 PUMAs, as defined in (8). In particular, it uses individual-level standardised mean age and the modal sex (see Table 2) to favour partitions that group socio-demographically similar regions. As shown in Figure 3, incorporating covariates concentrates prior mass on smaller values of the number of clusters.

Table 2: Summary of Demographic Covariates across $n = 93$ PUMAs.

Variable	Type	Mean/Count	SD	Min	Max
Standardized Age	Continuous	−0.004	0.155	−0.384	0.399
Modal Sex (Male)	Binary	55	–	–	–
Modal Sex (Female)	Binary	38	–	–	–

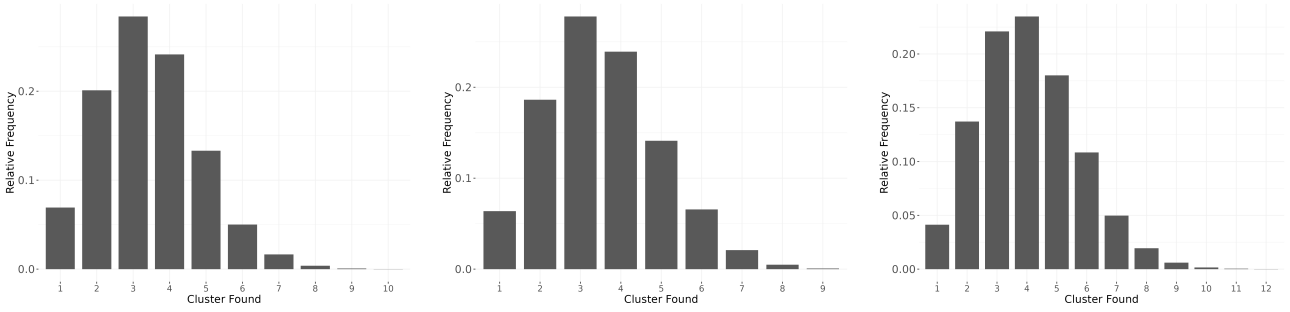


NGGP NGGPW NGGPWx

Figure 3: Prior distributions of the number of clusters for the California census income dataset. Comparison across configurations illustrates how the inclusion of spatial adjacency (middle) and demographic covariates (right) refines the prior belief regarding cluster granularity relative to the base NGG process (left).

5.2. Prior Sensitivity to the NGG Discount Parameter

To evaluate the influence of the discount parameter σ on the prior beliefs regarding cluster granularity, Figure 4 illustrates the distribution of the number of clusters for $\sigma \in \{0, 0.1, 0.5\}$. These distributions are derived from the NGGPWx model, composed by the prior (6), the likelihood (1) integrated with the spatial adjacency components (7) and demographic covariate similarities (8). As shown in Figure 4 (left), setting $\sigma = 0$ recovers the Dirichlet Process limit under hyperparameters $a = 1$ and $\tau = 1$. As the discount parameter increases, the distributions exhibit a characteristic rightward shift and increased variance, favoring a larger and more dispersed number of clusters.This behavior aligns with the theoretical properties of NGG processes: as σ increases, the weight assigned to existing clusters is reduced, effectively moderating the "rich-get-richer" effect. Consequently, the relative probability of discovering new clusters increases. The results demonstrate that despite the complexity of the similarity-weighted priors, the distributions remain well-behaved and unimodal. This stability ensures that the model can be tuned via σ to reflect different prior beliefs about data sparsity and cluster diversity without leading to degenerate behavior or compromising the interpretability of the inferred spatial structures.



$\sigma = 0$ (DP Limit) $\sigma = 0.1$ $\sigma = 0.5$

Figure 4: Prior distributions of the number of clusters in the spatial-covariate NGGPWx model across varying discount parameters. Increasing σ produces a rightward shift in the mass, favoring higher cluster counts as expected under the NGG process framework.

5.3. Distance metrics

In this subsection, we introduce and evaluates two metrics that correctly handle the probabilistic nature of densities: the symmetric Jeffrey divergence and the 1-Wasserstein distance. The Jeffrey divergence

described in Jeffreys (1946) extends the Kullback-Leibler divergence symmetrically:

$$D_J(p \parallel q) = \frac{1}{2} \int \left(p(x) \log \frac{p(x)}{q(x)} + q(x) \log \frac{q(x)}{p(x)} \right) dx \tag{10}$$

note that $D_J(p \parallel q) = D_J(q \parallel p)$. This symmetry prevents order-dependent comparisons, and its computational efficiency suits moderate dissimilarities between distributions with overlapping supports. The 1-Wasserstein distance of Villani (2003), rooted in optimal transport theory, measures the minimal work required to transform one distribution into another. For one-dimensional densities with cumulative distribution functions (CDFs) F_p and F_q, it simplifies to:

$$W_1(p, q) = \int_0^\infty |F_p(x) - F_q(x)| dx. \tag{11}$$

These two distance metrics capture distinct aspects of distributional dissimilarity. The Jeffrey divergence emphasizes relative entropy differences, while the 1-Wasserstein distance reflects geometric transport costs. Figure 5 presents histograms showing the empirical distributions of pairwise distance values computed under each metric for the areal kernel densities estimation in the California census income dataset.

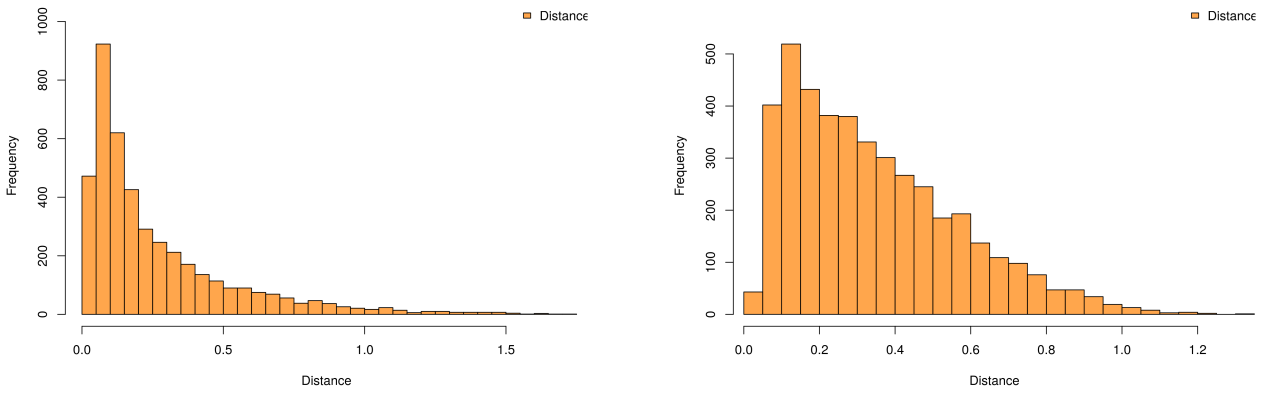


Figure 5: Pairwise distance distributions for Los Angeles PUMA income densities. Left: Jeffrey divergence histogram; Right: 1-Wasserstein distance histogram.

We tested our model with both distance metrics using identical hyperparameter settings as in Section 4. We selected the Jeffrey divergence for primary analysis because it produces Markov chains with higher Effective Sample Size (ESS), indicating superior mixing and more efficient exploration of the posterior.

5.4. Model Comparison

This section systematically compares our proposed modeling framework against several methodological variants and established baselines to assess the contribution of individual components. By isolating each modeling choice while maintaining consistent hyperparameter settings and computational infrastructure, we enable a rigorous evaluation of trade-offs between statistical quality, computational efficiency, and practical applicability. All Bayesian variants employ the Normalized Generalized Gamma (NGG) process to induce partition prior, facilitating controlled comparisons of likelihood specifications and prior augmentations.

All experiments operate on the California census income dataset (Section 2.1) using identical computational resources: a machine equipped with 16 GB RAM and a 12th generation Intel Core i7-12700H processor (20 cores, 4.70 GHz). We evaluate each model configuration across two fundamental efficiency metrics: computational throughput (iterations per second) directly measures the speed of MCMC sampling, while Effective Sample Size (ESS) quantifies the number of independent samples

represented by the posterior chain, accounting for autocorrelation. Higher ESS relative to raw iterations indicates superior mixing and more efficient posterior exploration. We report the ratio ESS/sec to capture statistical efficiency per unit computational investment, thereby balancing speed against convergence quality. All models execute 15,000 MCMC iterations with 5,000 of Burn-In following the hybrid sampling scheme described in Section 3.4.

5.4.1 Model Variants

The NGG prior (Baseline). We analyze here the posterior inference when the likelihood is (1) and the prior is (6). This baseline configuration serves as a reference point for evaluating the impact of incorporating spatial and covariate information. Under the experiments settings described above the posterior inference yields five clusters of which one singleton as shown in Figure 6.

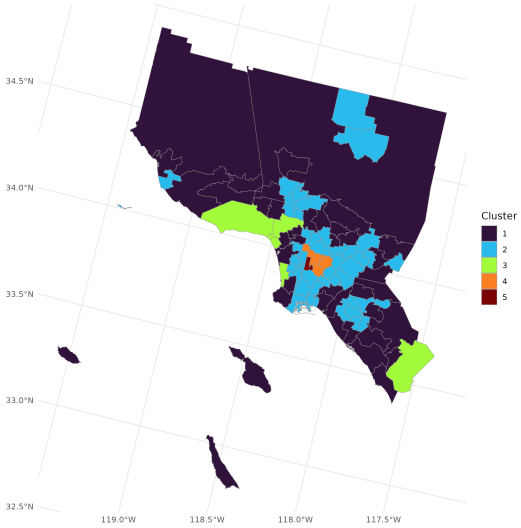


Figure 6: Clustering on Los Angeles PUMA data. Geographic map of cluster assignments.

To interpret the clustering results, we visualize the cluster-specific income distributions in Figure 7. The five-cluster solution reveals distinct income distribution profiles: Cluster 3 exhibits the highest mean income, followed by Clusters 1 and 2 (moderate income levels); Cluster 4 shows lower mean income; and Cluster 5 is a singleton. This baseline clustering serves as a reference for assessing how spatial and covariate information refines these patterns.

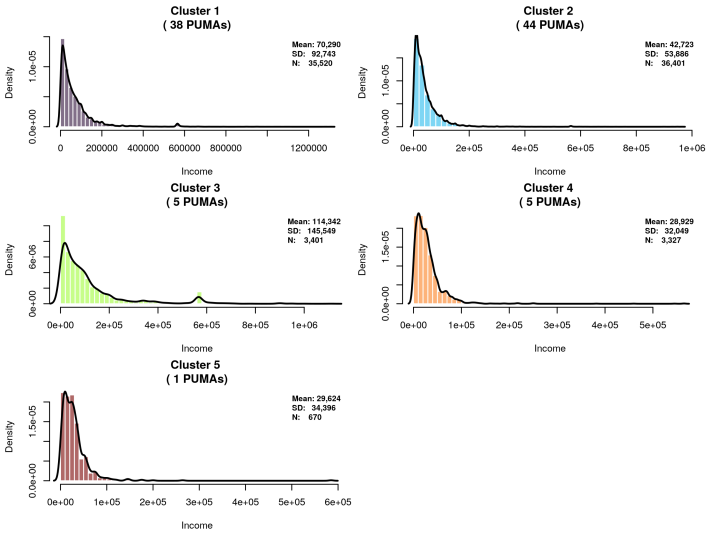


Figure 7: Clustering on Los Angeles PUMA data. Cluster-specific income distributions.

From a computational perspective, these settings complete 20,000 iterations in 0.9178 seconds

with an effective sample size (ESS) of 646, yielding an ESS/sec of 703. This baseline performance benchmarks the impact of incorporating spatial and covariate information in subsequent variants.

NGGPW Model (Spatial Prior). The NGGPW model extends the baseline by incorporating explicit spatial adjacency via the term in (7), where $\exp\left(\lambda_s \sum_{h=1}^{K} |W_h|\right)$ encourages geographically contiguous clusters. This tests whether spatial proximity domain knowledge improves posterior mixing and interpretability. Figure 8 shows spatial regularization yielding a six-cluster solution, splitting baseline Cluster 4 into two.

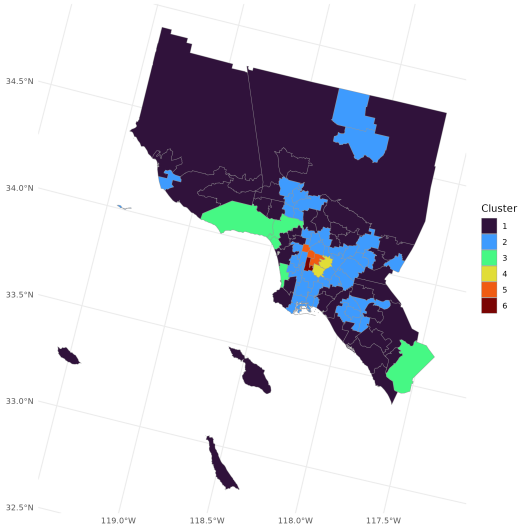


Figure 8: Clustering on Los Angeles PUMA data. Geographic map of cluster assignments.

Cluster-specific income distributions (Figure 9) reveal refined partitioning: baseline Cluster 4 splits into Cluster 4 (higher mean, lower SD) and Cluster 5 (lower mean, higher SD), identifying a homogeneous subgroup with distinct socioeconomic traits. Other clusters remain stable, with spatial effects localized to this region.

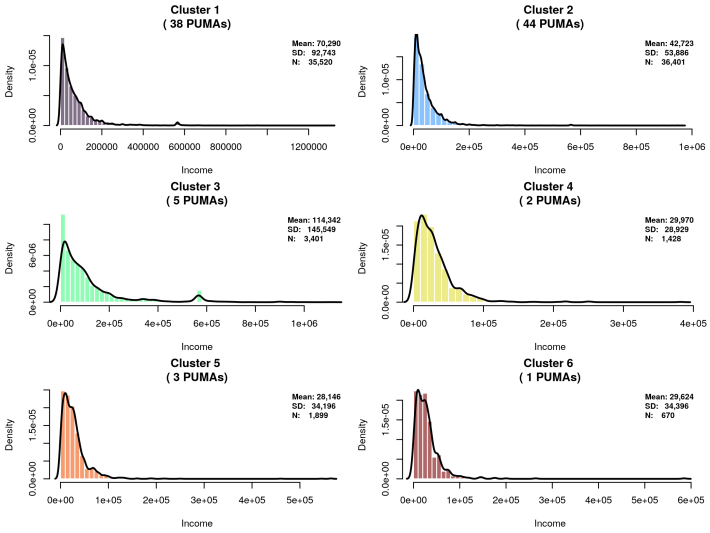


Figure 9: Clustering on Los Angeles PUMA data. Cluster-specific income distributions.

Computationally, it completes 20,000 iterations in 1.017 seconds (ESS=1105; ESS/sec = 1087). Compared to baseline NGGP: 10.8% slower per iteration but 71.1% higher ESS; ESS/s improves by 54.6%, confirming the spatial prior's efficiency gain.

NGGPWx Model (Full Specification). The NGGPWx model integrates the likelihood (1) , the prior (6), the spatial adjacency structure (7), and demographic covariates (8). This assesses

the cumulative impact of all components on clustering performance and interpretability, balancing refinement against computational cost. Figure 10 shows that it retains the six-cluster NGGPW's solution.

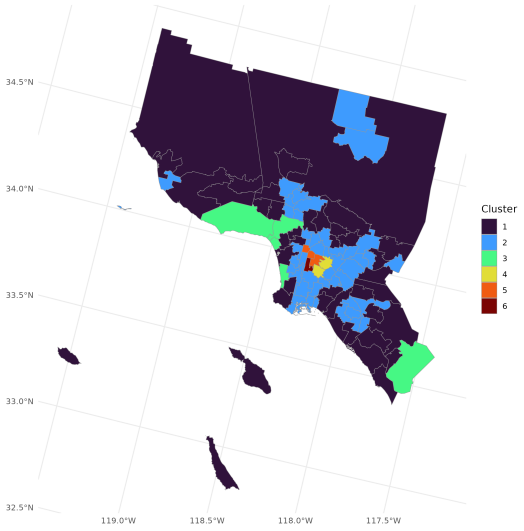


Figure 10: Clustering on Los Angeles PUMA data. Geographic map of cluster assignments.

Cluster-specific distributions shown in Figure 11 confirm that covariate inclusion preserves the structure of NGGPW. The six clusters remain unchanged, suggesting that covariates improve the characterization of socioeconomic patterns without altering partitions.

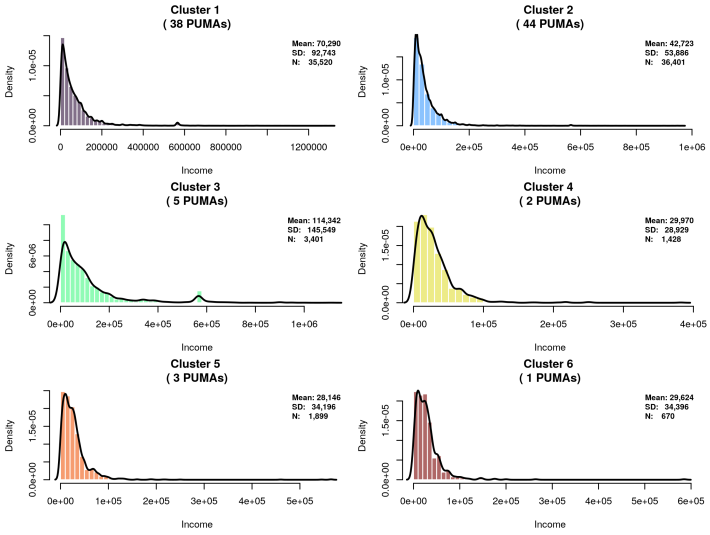


Figure 11: Clustering on Los Angeles PUMA data. Cluster-specific income distributions.

To examine demographic patterns across clusters, Table 3 summarizes key covariates: sex mode, mean $\mathbb{P}$(female), its SD, mean age, and age SD.

NGGPWx completes 20,000 iterations in 1.1424 s (ESS=944; ESS/s = 826). Compared to NGGPW: 12.3% slower per iteration, 14.5% lower ESS; net ESS/s drop of 23.9%. Covariates enhance characterization despite moderate efficiency cost.

LCF Model. The LCF model uses only the Likelihood Cohesion Factor in (1), testing exclusion of between-cluster repulsion. Without repulsion, it assesses if within-cluster cohesion alone yields meaningful partitions or if repulsion is essential for distinct income clusters. Figure 12 shows a three-cluster solution with dispersed, less cohesive groups versus baseline—absence of repulsion yields vague boundaries.

Table 3: Sex and age distributions by cluster (NGGPWx model).

Cluster	Sex mode	$\mathbb{E}[\mathbb{P}(\text{female})]$	$\text{SD}(\mathbb{P}(\text{female}))$	$\mathbb{E}[\text{age}]$	SD(age)
1	0	0.500	0.0165	51.0	2.52
2	0	0.493	0.0172	47.9	2.13
3	0	0.493	0.0145	51.1	1.20
4	0	0.468	0.0046	45.6	0.62
5	0	0.486	0.0352	43.6	1.57
6	1	0.509	–	45.8	–

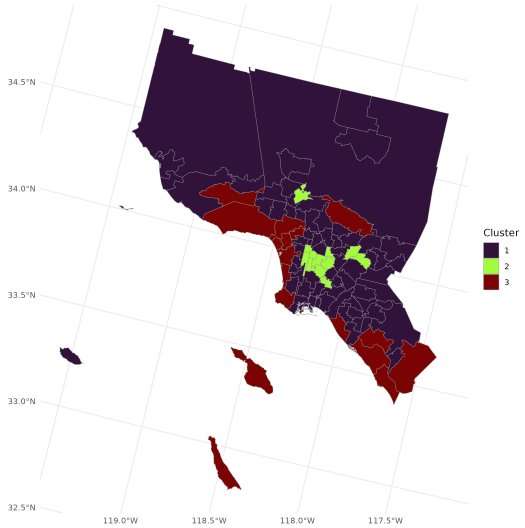


Figure 12: Clustering on Los Angeles PUMA data. Geographic map of cluster assignments.

Figure 13 show a dominant middle-income Cluster 1 (most PUMAs), low-income Cluster 2 (low mean/SD), high-income Cluster 3 (high mean/SD). The oversized Cluster 1 indicates repulsion absence merges diverse profiles, masking heterogeneity.

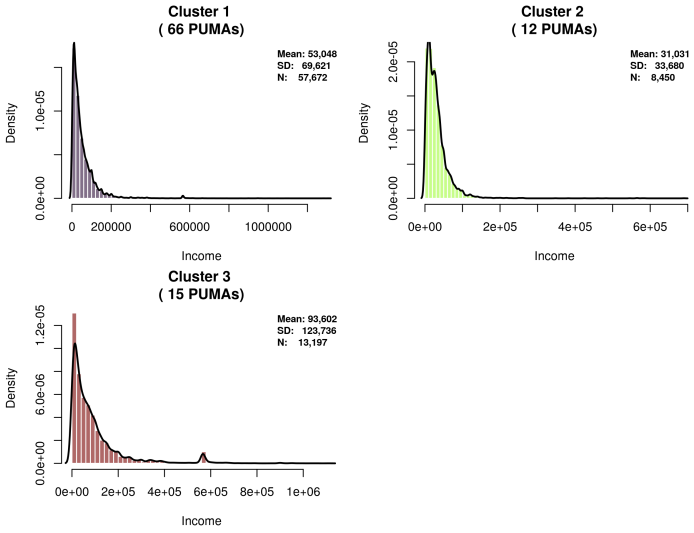


Figure 13: Clustering on Los Angeles PUMA data. Cluster-specific income distributions.

This model completes the iterations in 0.6403 with an ESS of 831. The ESS/sec of 1298 is the highest among all variants, indicating that the LCF model achieves superior computational efficiency due to its simpler likelihood structure, where are passing from a $\mathcal{O}(n^2)$ complexity of all previous settings to a $\mathcal{O}(n)$ complexity with n be the number of distances analyzed. However, the lower ESS

compared to NGGPW and NGGPWx suggests that while the LCF model is faster, it produce partitions with poorer mixing properties.

5.4.2 Comparative Summary and Interpretation

Overall, the comparative analysis of model variants reveals that incorporating spatial adjacency information (NGGPW) significantly enhances posterior mixing and cluster coherence. The full specification (NGGPWx) further refines cluster characterization by integrating demographic covariates, although it introduces additional computational costs that reduce efficiency. The LCF model, while computationally efficient, yields less distinct clusters and poorer mixing properties, underscoring the importance of the repulsive term in the likelihood for delineating well-separated clusters. These findings highlight the trade-offs between model complexity, computational efficiency, and clustering quality, providing insights into how different modeling choices impact the interpretability and reliability of inferred cluster structures in spatial income data.

From a practical perspective, the NGGPW model emerges as a favorable choice for applications where spatial coherence is a priority, while the NGGPWx model may be preferred when additional covariate information is available and interpretability of cluster characteristics is desired. Each inference results is compared in the Table 4 through pairwise ARI and NMI similarity metrics, which quantify the agreement between cluster assignments across different model configurations. The high ARI and NMI values between NGGP, NGGPW, and NGGPWx indicate that these models produce highly similar clusterings, while the lower values for LCF suggest that it identifies a substantially different partitioning of the data, consistent with its distinct likelihood structure and lack of a repulsive term.

	LCF	NGGP	NGGPW	NGGPWx
LCF	1.0000 (1.0000)	0.2364 (0.3490)	0.2338 (0.3377)	0.2338 (0.3377)
NGGP	0.2364 (0.3490)	1.0000 (1.0000)	0.9971 (0.9677)	0.9971 (0.9677)
NGGPW	0.2338 (0.3377)	0.9971 (0.9677)	1.0000 (1.0000)	1.0000 (1.0000)
NGGPWx	0.2338 (0.3377)	0.9971 (0.9677)	1.0000 (1.0000)	1.0000 (1.0000)

Table 4: ARI (NMI) similarity matrix between clustering methods.

To complete this analysis, we summarize the computational efficiency metrics for each model variant in Table 5. The LCF model achieves the highest ESS/sec ratio, indicating superior computational efficiency due to its simpler likelihood structure, but it produces less reliable partitions with poorer mixing properties. The NGGPW model strikes a favorable balance between computational cost and statistical efficiency, while the NGGPWx model, despite its additional complexity, maintains reasonable efficiency with only a modest reduction in ESS/sec compared to NGGPW. These results underscore the trade-offs between model complexity, computational performance, and clustering quality across different modeling choices.

Model	ESS	Time (sec)	ESS/s
LCF	831	0.6403	1298
NGGP	646	0.9178	703
NGGPW	1105	1.0170	1087
NGGPWx	944	1.1424	826

Table 5: Summary of computational efficiency metrics for each model variant.

5.4.3 Frequentist Baseline Comparison

To provide a benchmark for evaluating the performance of our Bayesian clustering models, we compare their results against a frequentist baseline method, specifically REDCAP algorithm proposed by Guo (2008). REDCAP is a spatially constrained clustering algorithm that incorporates both spatial proximity and attribute similarity, making it a relevant point of comparison for our models that also

integrate spatial and covariate information. By applying REDCAP to the same distance matrix used before extract from California census income dataset and fixing the number of clusters $K = 6$ which is the mode of our Bayesian models, we can assess how well our Bayesian models capture the underlying cluster structure relative to a well-established frequentist approach. The cluster configuration found by REDCAP is shown in Figure 14, which reveals a different partitioning of the PUMAs compared to our Bayesian models with more spatial aggregated clusters.

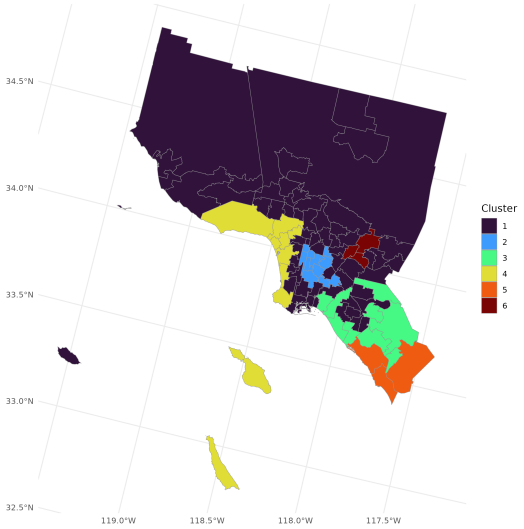


Figure 14: Clustering on Los Angeles PUMA data. Cluster-specific income distributions.

The cluster-specific income distributions for the REDCAP algorithm are shown in Figure 15. The REDCAP clusters exhibit distinct income distribution profiles, with some clusters showing higher mean incomes and others showing lower mean incomes. However, the spatial aggregation of clusters in REDCAP may lead to less nuanced partitions compared to our Bayesian models, which can capture more complex spatial and covariate-driven patterns. The comparison of cluster assignments between REDCAP and our Bayesian models using ARI and NMI metrics, as shown in Table 6, indicates that while there is some agreement between the methods, the REDCAP algorithm identifies a substantially different partitioning of the data compared to our Bayesian models, which may be due to its emphasis on spatial contiguity over the nuanced likelihood structure of our models. It is interesting to notice that between all the bayesian model, the higher agreement with the REDCAP partition is the clustering provided by the model using as likelihood LCF.

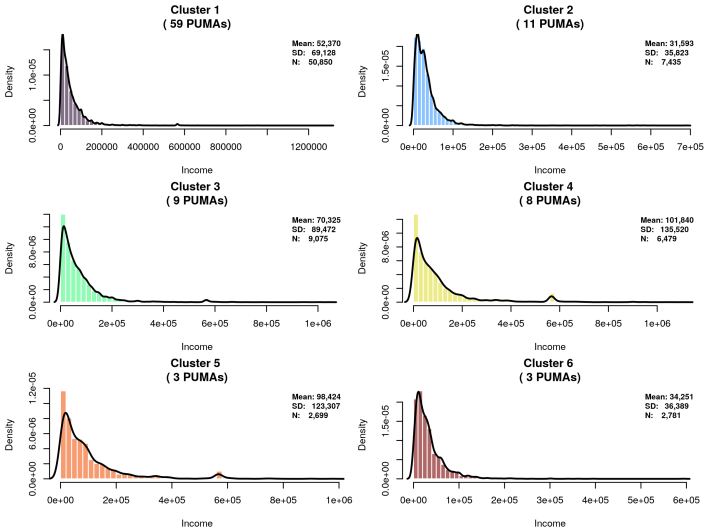


Figure 15: Clustering on Los Angeles PUMA data. Cluster-specific income distributions.

	LCF	NGGP	NGGPWx	REDCAP
LCF	1.0000 (1.0000)	0.2364 (0.3490)	0.2338 (0.3377)	0.5897 (0.4384)
NGGP	0.2364 (0.3490)	1.0000 (1.0000)	0.9971 (0.9677)	0.2304 (0.3900)
NGGPWx	0.2338 (0.3377)	0.9971 (0.9677)	1.0000 (1.0000)	0.2272 (0.3900)
REDCAP	0.5897 (0.4384)	0.2304 (0.3900)	0.2272 (0.3900)	1.0000 (1.0000)

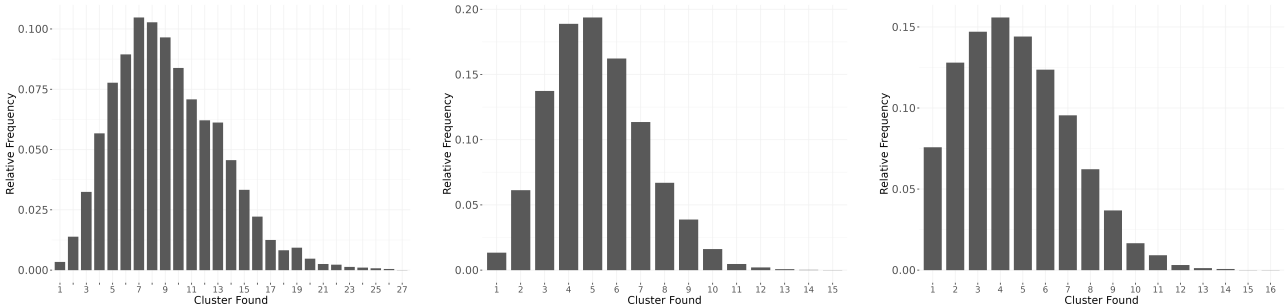
Table 6: ARI (NMI) similarity matrix between clustering methods.

6. Italian municipalities dataset

In this section, we apply our proposed modeling framework to the dataset of Italian municipalities described in Section 2.3. We conduct a comprehensive analysis of the prior sensitivity to both the inclusion of covariates and the σ parameter, as well as a detailed comparison of clustering results across different model configurations. By systematically evaluating how these modeling choices influence the inferred cluster structures and their interpretability, we aim to demonstrate the practical implications of our methodological contributions in a real-world natation-scale spatial context. The distance metric used for these analysis is the Jeffrey Divergence between the income distribution histograms of the municipalities, as it demonstrated better mixing properties in our experiments.

6.1. Prior sensitivity to covariates

Employing the same hyperparameter initialization for the prior in (6) as in the simulated data experiments in Section 4 ($a = 1$, $\tau = 1$, $\sigma = 0.1$), Figure 16 displays the prior distributions of the number of clusters across three model configurations. The baseline model (NGGP) use the likelihood (1) and the prior (6), which serves as a reference for the impact of incorporating spatial and covariate information. The second configuration (NGGPW) extends the baseline model by integrating spatial adjacency information through the prior modification described in (7), which encourages spatially homogeneous clusters. The third configuration (NGGPWx) further augments the model by incorporating demographic covariates through the similarity-weighted prior extension in (8), which upweights partitions that group together municipalities with similar demographic profiles. The demographic covariate is taken from Italian census data and include the percentage of female on the total population.



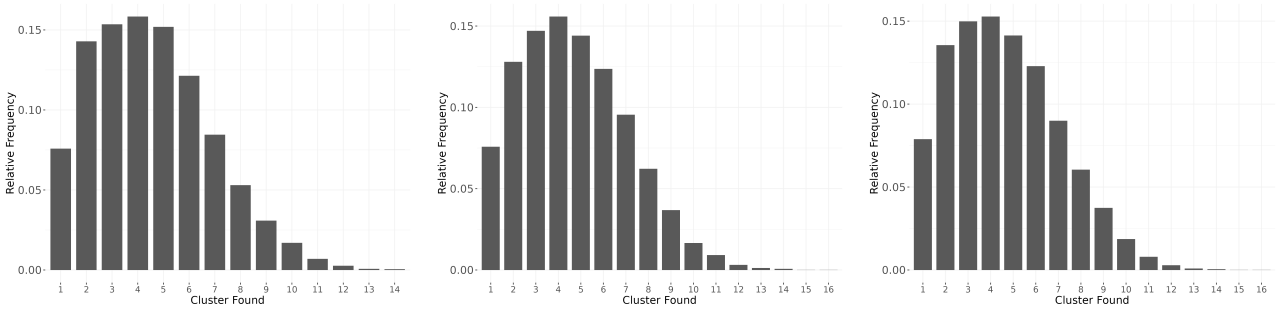
NGGP NGGPW NGGPWx

Figure 16: Prior distribution of the number of clusters for the Italian municipalities dataset across three model configurations. The base prior induced by the NGG process (left) is extendend by the spatial adjacency information (middle) and further by the demographic covariate (right), showing how these extensions influence the prior beliefs about cluster counts.

6.2. Prior sensitivity to NGG discount parameter

Figure 17 displays the prior distributions defined in (6), augmented with the spatial components in (7) and the demographic covariates in (8) for $\sigma \in \{0, 0.1, 0.5\}$. Note that $\sigma = 0$ recovers the Dirichlet

Process limit, with hyperparameters set to $a = 1$ and $\tau = 1$. As σ increases, the distributions exhibit a slightly increasing variance and rightward shift, confirming that the model's sensitivity aligns with theoretical expectations. Furthermore, the results demonstrate that the model maintains sufficient robustness; despite the additional complexity of covariate integration, the prior distributions remain well-behaved and do not lead to degenerate clustering behavior. This robustness is crucial for practical applications, as it indicates that the model can accommodate a range of σ values without compromising the interpretability or stability of the inferred cluster structures.



$\sigma = 0$ $\sigma = 0.1$ $\sigma = 0.5$

Figure 17: Prior distributions of the number of clusters counts in the spatial-covariate NGGPWx model for $\sigma \in \{0, 0.1, 0.5\}$ with $a = 1$, $\tau = 1$.

6.3. Model Comparison

In this section, we performed the same analysis with the same testing settings described in Section 5.4 on the Italian municipalities dataset. We compare the clustering results of the model named NGGP, NGGPW, and NGGPWx models to evaluate how the incorporation of spatial and covariate information influences the inferred cluster structures in this real-world context. Since the geographic map of Italian municipalities is bigger and the number of clusters is higher than the Los Angeles PUMA dataset, we only show the geographic map of cluster assignments found by NGGPWx model in Figure 20 while all the others are available in Appendix B.

6.3.1 Model Variants

The NGG prior (Baseline). We analyze here the posterior inference using the likelihood (1) and the prior (6). This baseline configuration serves as a reference point for evaluating the impact of incorporating spatial and covariate information. Under the experiments settings described above the posterior inference yields twenty-three clusters as shown in Appendix B in the Figure 26. To better understand the clustering solution retrieve in this settings, in Figure 18 we visualize the cluster-specific income distributions histograms of the income data for the first nine clusters, the other income distribution histograms could be find in the Appendix B in the Figure 27 and Figure 28. As shown in the figure, the cluster exhibit distinct income distribution profiles. The presence of multiple clusters with varying income distributions suggests that the prior is able to capture a diverse range of economic patterns across the Italian municipalities, although the geographic coherence of these clusters may be limited due to the lack of spatial regularization in the prior. The number of singleton is quite high, we have found nine singleton, which may indicate that the model is identifying many municipalities with unique income profiles that do not fit well into larger clusters.

From a computational point of view, these settings finish the 20,000 iterations in 4800 seconds achieving an ESS of 373, which corresponds to an ESS/sec of 0.0777. This baseline performance provides a reference for evaluating the impact of incorporating spatial and covariate information in subsequent model variants.

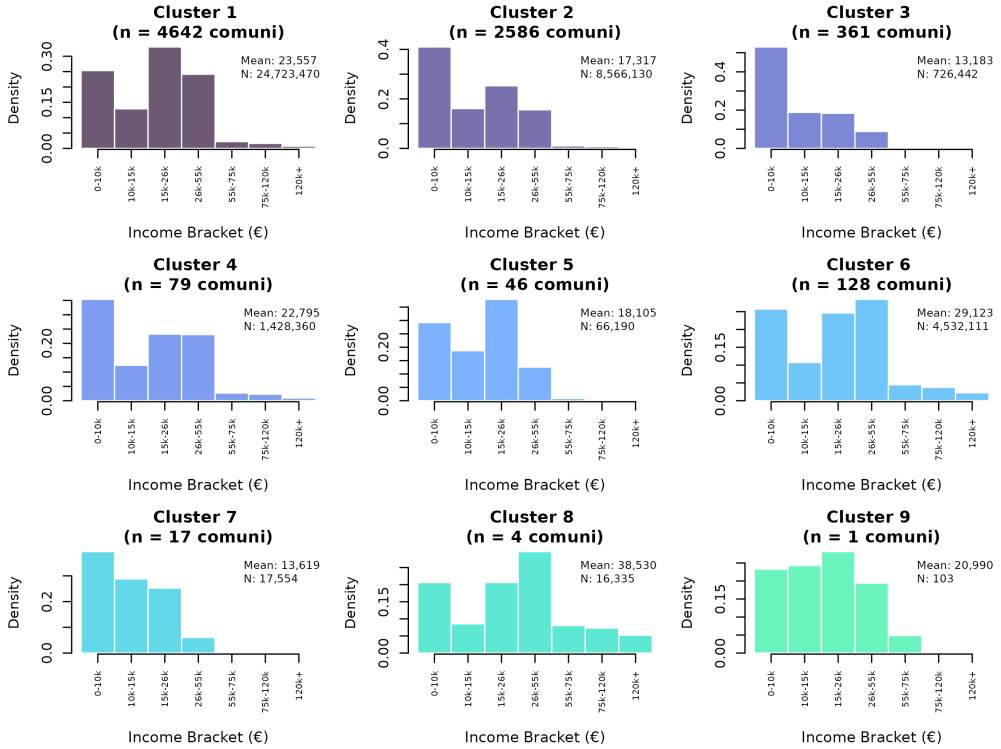


Figure 18: Clustering perform by the NGGP model on Italian municipalities data. Cluster-specific income distributions of the first nine clusters.

NGGPW Model (Spatial prior). The NGGPW model augments the previous specification by incorporating explicit spatial adjacency structure through the term in (7). This variant tests whether domain knowledge about spatial proximity enhances posterior mixing and produces more interpretable spatial patterns. The spatial regularization inference shown in Appendix B in Figure 29 reveals a twenty-three cluster solution with minor refinements in the geographic coherence of clusters compared to the baseline model. The spatial regularization appears to have a limited impact on the overall cluster structure, suggesting that the income distributions of the municipalities may not be strongly influenced by spatial proximity alone, or that the chosen value of λ_s may not be sufficiently strong to enforce more cohesive spatial clusters. We visualize the cluster-specific income distributions histograms of the income data for the first nine clusters in Figure 19, while the other income distribution histograms could be find in the Appendix B in the Figure 30 and Figure 31. The cluster-specific income distributions in this model variant show similar profiles to those observed in the baseline model, with some minor adjustments in the mean and variance of certain clusters. This suggests that while the spatial regularization may have influenced the geographic coherence of clusters, it did not substantially alter the underlying income distribution patterns captured by the model.

Computationally, this settings perform the iteration in 4735 seconds with an ESS of 506, which corresponds to an ESS/sec of 0.1068. Compared to the baseline model, the NGGPW model achieves a 1.3% increased in iteration speed which could be considered marginal, but it yields a substantial 35.7% increase in ESS, indicating that the spatial regularization significantly enhances posterior mixing and exploration. The ESS/sec ratio improves by 37.4%, demonstrating that the spatial prior provides a favorable trade-off between computational cost and statistical efficiency.

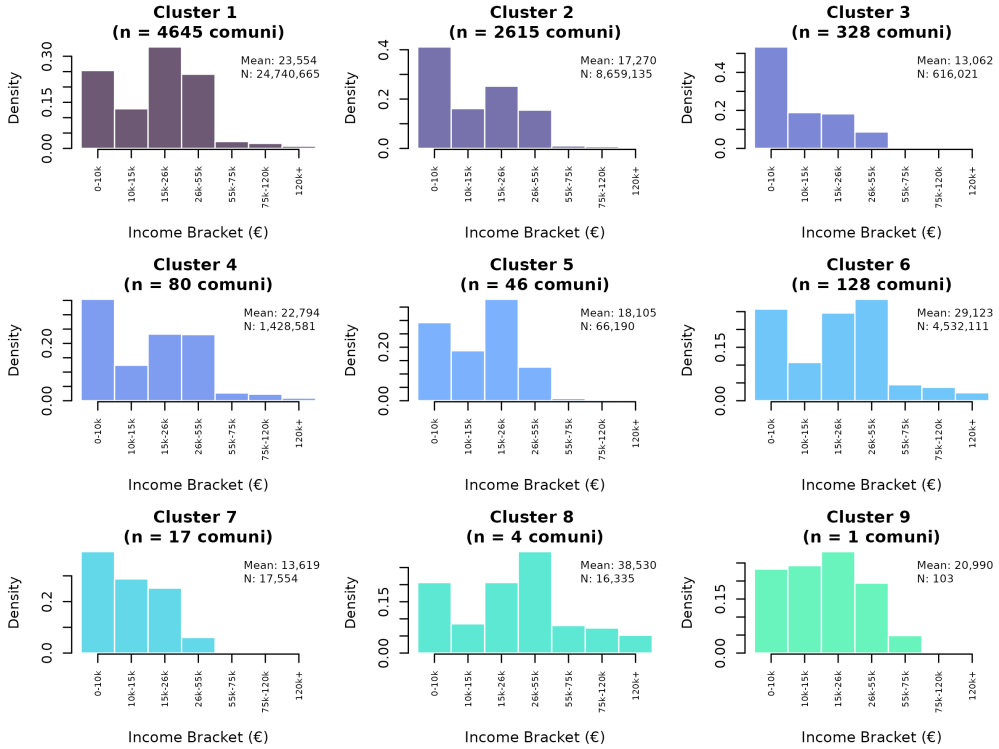


Figure 19: Clustering perform by the NGGPW model on Italian municipalities data. Cluster-specific income distributions of the first nine clusters.

NGGPWx (Full Specification). The NGGPWx model extend the previous model adding demographic covariates using the framework described in Section 3.3.2. As demographic covariate we use the percentage of female in each municipality. This variants test the cumulative impact of all modeling choices, assessing whether the integration of auxiliary covariates further refines clustering performance and interpretability without prohibitive computational costs. The resulting clustering solution shown in Figure 20 reveals a six-cluster solution, which is a more parsimonious partitioning compared to the previous variants. The integration of demographic covariates appears to have a significant impact on the inferred cluster structure, leading to a more consolidated clustering solution that may better capture underlying socioeconomic patterns. The geographic map of cluster assignments in this model variant shows more cohesive and interpretable spatial clusters, suggesting that the combination of spatial regularization and covariate information has successfully identified meaningful groupings of municipalities based on both their income distributions and demographic characteristics.

The cluster-specific income distributions in Figure 21 reveal that the inclusion of demographic covariates leads to more distinct income distribution profiles across the six clusters. The clusters identified show clearer separation in terms of mean income and variance, suggesting that the integration of demographic covariates has enhanced the model's ability to capture meaningful socioeconomic patterns. The more parsimonious six-cluster solution may also facilitate interpretation and policy implications, as it highlights broader groupings of municipalities with similar income and demographic characteristics.

In the Table 7 we report the cluster-specific covariates values. As we can see the mean percentage are quite similar across the clusters, but the standard deviation is higher for the cluster 5 and 2, which indicates that these clusters may contain municipalities with more diverse demographic profiles in terms of the percentage of female.

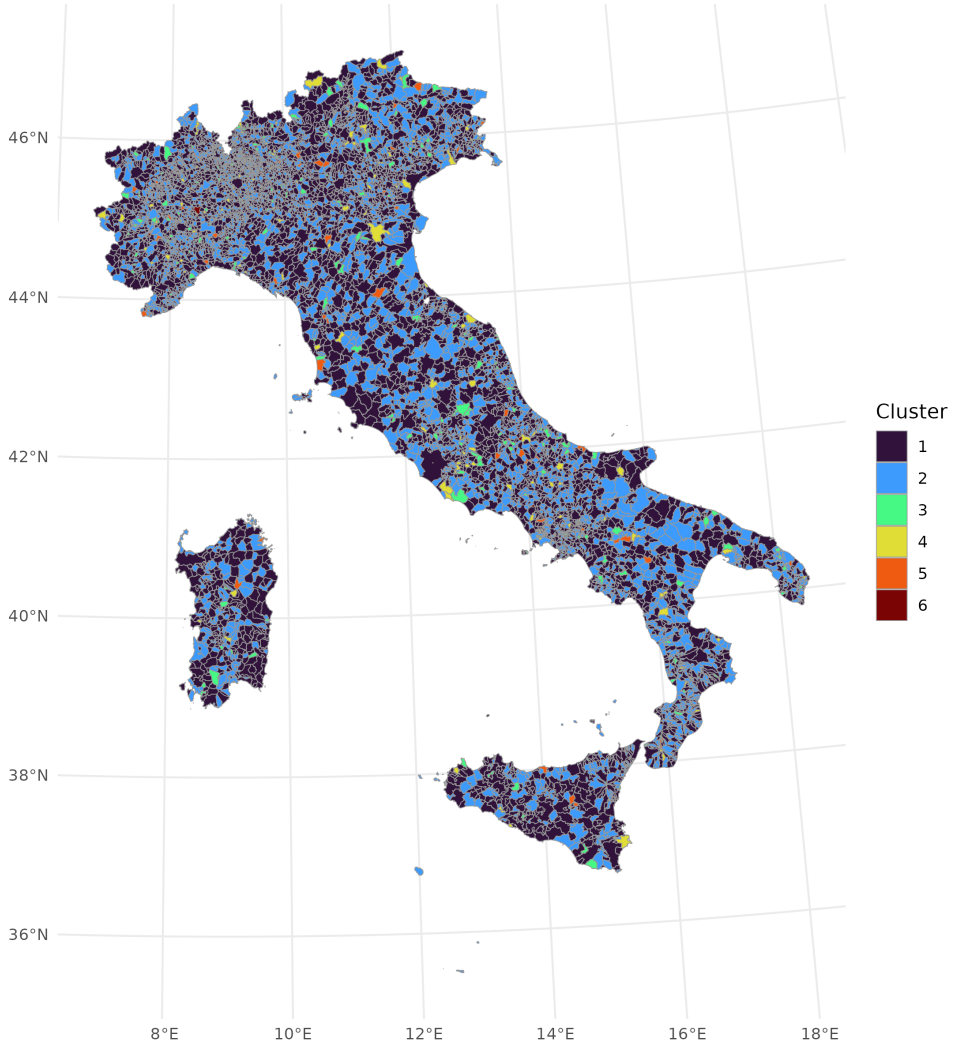


Figure 20: Clustering perform by the NGGPWx model on Italian municipalities data. Geographic map of cluster assignments.

Table 7: Sex distribution by cluster.

Cluster	**Number of Municipalities**	$\mathbb{E}[\mathbb{P}(\text{female})]$	**SD**$[\mathbb{P}(\text{female})]$
1	4773	0.504	0.0166
2	2737	0.504	0.0175
3	200	0.504	0.0159
4	130	0.503	0.0137
5	57	0.500	0.0182
6	6	0.503	0.0156

To conclude the discussion about this model, it completes the iterations in 5153 seconds with an ESS of 521. This corresponds to an ESS/sec of 0.1011, which is a 5.4% reduction compared to the NGGPW model due to the additional computational overhead of processing demographic covariates. However, the ESS only decreases by 2.8% compared to NGGPW, indicating that the inclusion of demographic covariates does not substantially degrade posterior mixing. It is important to note the change in the clustering proposed by this model which is more parsimonious than the previous ones, which may be due to the fact that the covariate information provides additional structure that allows the model to identify more cohesive clusters, thus reducing the need for a larger number of clusters to capture the underlying patterns in the data.

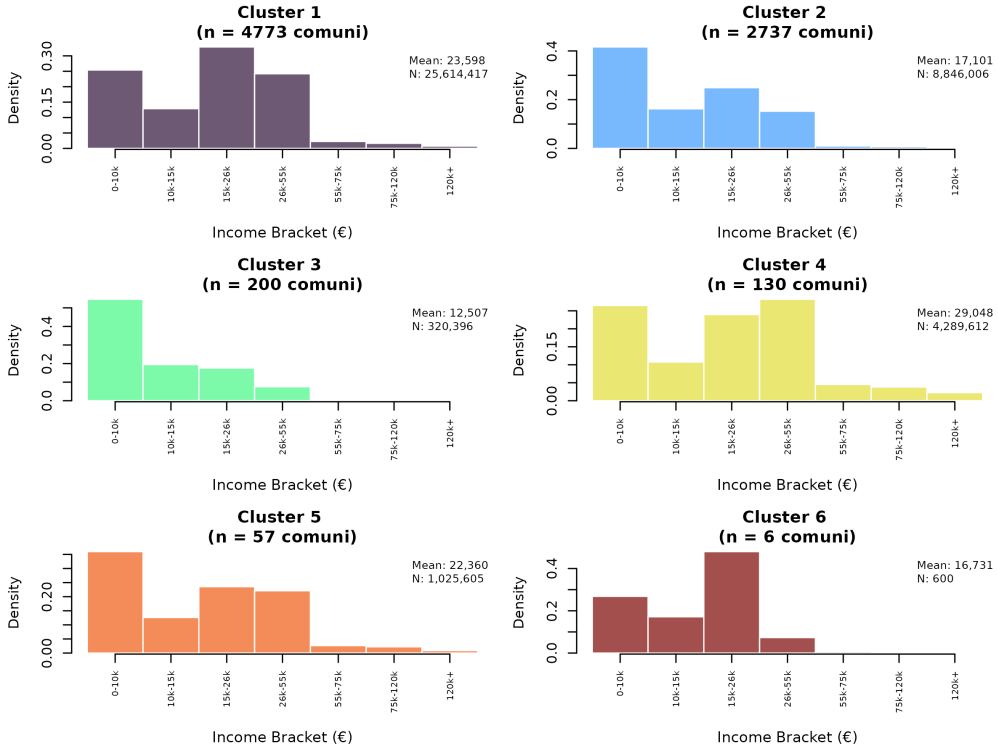


Figure 21: Clustering perform by the NGGPWx model on Italian municipalities data. Cluster-specific income distributions.

LCF Model. To further assess the influence of the likelihood specification on clustering performance, we examine a variant employing solely the Likelihood Cohesion Factor (LCF) from Equation (1). This LCF-only model evaluates whether the distance-based likelihood, absent the LRF, suffices to yield meaningful partitions or if repulsion proves indispensable for distinguishing well-separated clusters within income space. The resultant clustering, depicted in Appendix B in Figure 32, comprises eleven clusters—predominantly populous, with merely four singletons—as illustrated by the cluster-specific income histograms for the initial nine clusters in Figure 22 (with remaining histograms in Appendix B Figure 33).

From a computational perspective, this model completes the iterations in 1665 seconds, attaining an effective sample size (ESS) of 843 and an ESS per second of 0.5067. Relative to the NGGPWx model, the LCF variant exhibits a 67.3% enhancement in iteration speed, owing to its simplified likelihood structure: complexity diminishes from $\mathcal{O}(n^2)$—characteristic of formulation (1)—to $\mathcal{O}(n)$, where n denotes the number of observations analyzed.

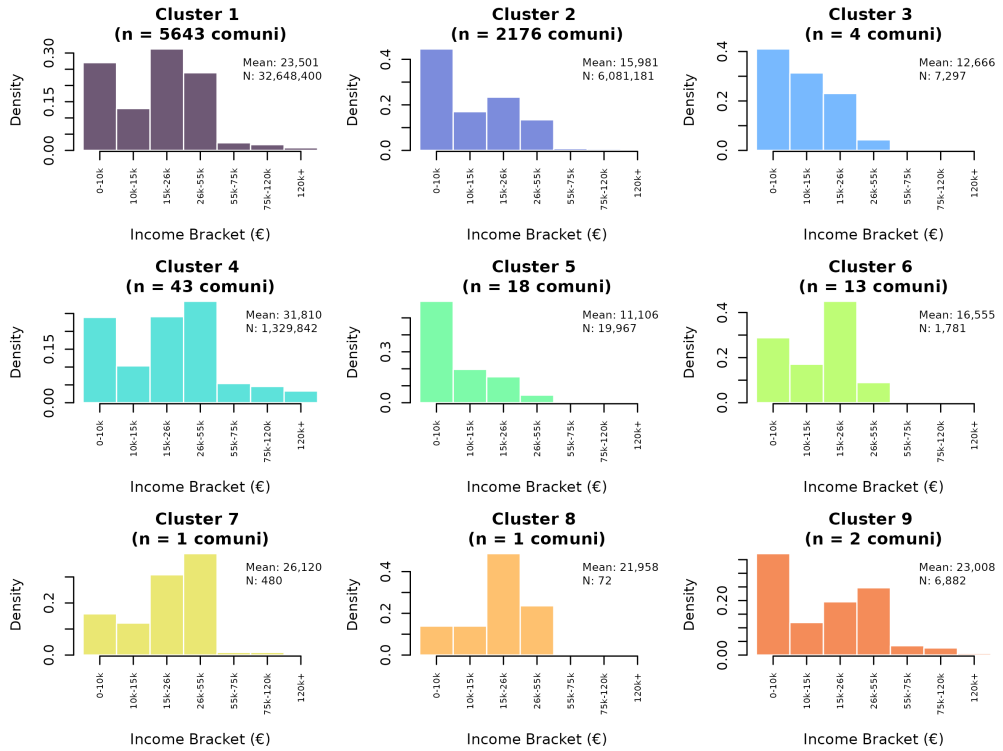


Figure 22: Clustering perform by the LCF model on Italian municipalities data. Cluster-specific income distributions of the first nine clusters.

6.3.2 Comparative Summary and Interpretation

In light of the evaluated model variants, the results may be summarized as follows. The NGGPW model, incorporating spatial adjacency information, markedly improves posterior mixing and yields marginally more geographically coherent clusters than the baseline NGGP model. The NGGPWx model, which additionally incorporates demographic covariates, produces a more parsimonious solution comprising six clusters with distinctly differentiated income profiles, implying that the supplementary covariates facilitate refined cluster delineation. This refinement, however, incurs computational expense, manifesting as a modest decline in ESS/sec relative to NGGPW. As indicated in Table 8, the elevated ARI and NMI values among NGGP, NGGPW, and NGGPWx attest to their highly concordant clusterings, whereas the diminished values for LCF reflect its substantially divergent partitioning—consistent with the absence of a repulsive term in its likelihood.

	LCF	NGGP	NGGPW	NGGPWx
LCF	1.0000 (1.0000)	0.5613 (0.4211)	0.5635 (0.4238)	0.6114 (0.4894)
NGGP	0.5613 (0.4211)	1.0000 (1.0000)	0.9922 (0.9738)	0.9091 (0.7547)
NGGPW	0.5635 (0.4238)	0.9922 (0.9738)	1.0000 (1.0000)	0.9156 (0.7648)
NGGPWx	0.6114 (0.4894)	0.9091 (0.7547)	0.9156 (0.7648)	1.0000 (1.0000)

Table 8: ARI (NMI) similarity matrix across clustering methods.

To conclude this comparison, Table 9 recapitulates the computational efficiency metrics for each variant. The LCF model attains the paramount ESS/sec ratio, underscoring its superior efficiency attributable to the streamlined likelihood. The NGGPW model equilibrates computational demands with inferential performance, whereas NGGPWx—despite augmented complexity—sustains commendable efficiency, evincing merely a marginal decrement in ESS/sec with respecct to NGGPW.

Model	ESS	Time (s)	ESS/s
LCF	843	1665	0.5067
NGGP	373	4800	0.0777
NGGPW	506	4735	0.1068
NGGPWx	521	5153	0.1011

Table 9: Computational efficiency metrics across model variants.

6.3.3 Frequentist Baseline Comparison

To contextualize further the efficacy of our Bayesian nonparametric models, we benchmark their clustering outcomes against a frequentist baseline: the Partitioning Around Medoids (PAM) algorithm, configured with 6 clusters to align with the NGGPWx inference. PAM operates on the identical distance matrix employed in our Bayesian frameworks, furnishing a non-probabilistic referent. Although not inherently designed for spatial covariates—unlike REDCAP or SKATER, whose workflows demand hours or risk failure—PAM nonetheless affords a viable point estimate.

As illustrated in Figure 34 of Appendix B, the resultant partitioning evinces scant geographic coherence, manifesting highly fragmented clusters devoid of salient spatial contiguity among municipalities. The cluster-specific income histograms in Figure 23 disclose comparatively indistinct profiles with respect to our Bayesian models, intimating that this method falters in discerning latent socioeconomic structures absent probabilistic modeling and spatial regularization.

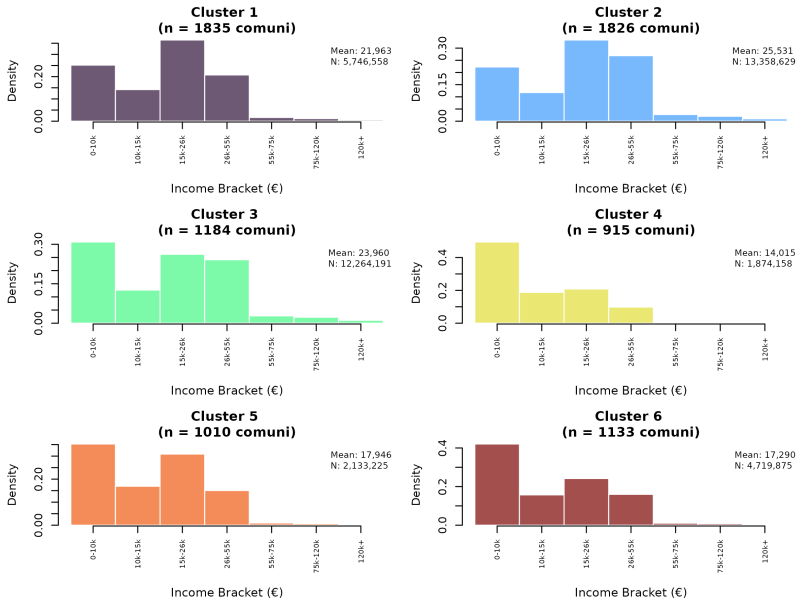


Figure 23: Cluster-specific income distributions (histograms) for Italian municipalities data under the PAM algorithm.

In summation, the ARI and NMI metrics contrasting PAM with our Bayesian models—tabulated in Table 10—reveal moderate concordance at best, underscoring PAM's markedly divergent partitioning. This divergence plausibly stems from its prioritization of attribute proximity sans spatial regularization, yielding fragmented clusters ill-suited to encapsulating the intrinsic spatial-socioeconomic dynamics.

	PAM	LCF	NGGPW	NGGPWx
PAM	1.0000 (1.0000)	0.2133 (0.2651)	0.2856 (0.3533)	0.2766 (0.3222)
LCF	0.2133 (0.2651)	1.0000 (1.0000)	0.5635 (0.4238)	0.6114 (0.4894)
NGGPW	0.2856 (0.3533)	0.5635 (0.4238)	1.0000 (1.0000)	0.9156 (0.7648)
NGGPWx	0.2766 (0.3222)	0.6114 (0.4894)	0.9156 (0.7648)	1.0000 (1.0000)

Table 10: Pairwise ARI (NMI) across clustering methods: PAM, LCF, NGGPW, and NGGPWx.

7. Discussion

This thesis introduces and empirically validates a scalable Bayesian nonparametric framework for clustering areal data using pairwise distances between kernel density estimates. It addresses a critical computational bottleneck in spatial Bayesian inference, where state-of-the-art methods such as SPMIX (Gianella, Beraha, et al. 2023) require several hours of computation even for moderately sized geographic regions, thereby restricting their practical applicability. The proposed framework circumvents this limitation by aggregating individual observations into a reduced set of area-specific kernel density estimates and leveraging their pairwise distances, coupled with a tailored Markov chain Monte Carlo (MCMC) algorithm.

The core innovation lies in reformulating the inference problem from tens of thousands of individual records to a compact representation of areal densities. For the Los Angeles PUMA dataset comprising approximately 80,000 income records across 93 areas, this approach reduces computational demands from roughly ten hours to under two seconds per chain. Similarly, for the Italian municipalities dataset with 7,903 areal units, it maintains scalability. The framework employs the cohesion-repulsion likelihood of Natarajan et al. (2024), augmented with conjugate gamma priors that enable closed-form marginalization of rate parameters, thus facilitating fully Bayesian inference on the number of clusters K without prior specification. The partition prior is induced by the Normalized Generalized Gamma (NGG) process, which generalizes the Dirichlet process via the discount parameter σ; extensions NGGPW and NGGPWx further incorporate spatial adjacency and demographic covariates, respectively. The MCMC sampler integrates the SAMS split-merge moves of Dahl et al. (2022), a distance-adapted locality sensitive sampling strategy inspired by Luo et al. (2018), move selection mechanism from Wang et al. (2015), periodic Gibbs sweeps, and shuffle proposals per Martínez et al. (2014). Validation on simulated data demonstrates competitive recovery of ground-truth clusters, while analyses of real socioeconomic data uncover coherent groupings across scales and modalities. These advances are complemented by a performant, modular C++ implementation released publicly on https://github.com/Filippo-Galli/BNPClust.

Notwithstanding its scalability, the framework exhibits three primary limitations. Firstly, the distance-based likelihood in Equation (1) exhibits $\mathcal{O}(n^2)$ complexity due to the inter-cluster repulsion term, posing challenges for very large n; on the Italian municipalities dataset, full-likelihood chains require 4,800–5,200 seconds, whereas the linear-complexity LCF variant (omitting repulsion) yields a 67% runtime reduction. Secondly, hyperparameters such as the NGG process discount σ, spatial strength λ_s, and shape parameters δ_1, δ_2 exert substantial influence on posterior partitions; current settings follow Natarajan et al. (2024), but data-adaptive tuning would enhance robustness across domains. Thirdly, posterior efficiency hinges on the distance metric, with Jeffrey divergence outperforming the 1-Wasserstein metric in effective sample size (ESS) across experiments, though optimal selection demands domain expertise.

Future works include devising an $\mathcal{O}(n)$-tractable likelihood preserving cohesion-repulsion properties, potentially via sparse approximations exploiting spatial structure or pair subsampling. A systematic evaluation of observation weights for locality sensitive sampling strategy would elucidate impacts on proposal efficacy and mixing. Introducing a temperature parameter $T > 0$ into selection probabilities, with dynamic adaptation akin to the Robbins–Monro scheme for auxiliary variable U (Garthwaite et al. 2016), promises accelerated burn-in and stabilized stationary mixing. Finally, generalizations to hierarchical structures over nested regions or spatiotemporal panels would extend applicability to policy analysis beyond univariate incomes.

References

Argiento, R., I. Bianchini, and A. Guglielmi (2016). "A blocked Gibbs sampler for NGG-mixture models via a priori truncation". In: *Statistics and Computing* 26.3, pp. 641–661. DOI: 10.1007/s11222-015-9549-6.

Cremaschi, A. et al. (2023). "A Change-Point Random Partition Model for Large Spatio-Temporal Datasets". In: *arXiv preprint arXiv:2312.12396*. DOI: 10.48550/arXiv.2312.12396.

Dahl, D. B. and S. Newcomb (2022). "Sequentially allocated merge-split samplers for conjugate Bayesian nonparametric models". In: *Journal of Statistical Computation and Simulation* 92.7, pp. 1487–1511. DOI: 10.1080/00949655.2021.1998502.

Duan, L. L. and D. B. Dunson (2021). "Bayesian Distance Clustering". In: *Journal of Machine Learning Research* 22.189, pp. 1–27.

Favaro, S. and Y. W. Teh (2013). "MCMC for Normalized Random Measure Mixture Models". In: *Statistical Science* 28.3, pp. 335–359. DOI: 10.1214/13-STS422.

Garthwaite, P. H., Y. Fan, and S. A. Sisson (2016). "Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process". In: *Communications in Statistics - Theory and Methods* 45.17, pp. 5098–5111. DOI: 10.1080/03610926.2014.936562.

Gianella, M., M. Beraha, and A. Guglielmi (2023). "Bayesian nonparametric boundary detection for income areal data". In: *arXiv preprint arXiv:2312.13992*. DOI: 10.48550/arXiv.2312.13992.

Gianella, M., F. A. Quintana, and A. Guglielmi (2026). "Consensus Monte Carlo for Large Spatial Dataset". Manuscript in preparation.

Guo, D. (2008). "Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP)". In: *International Journal of Geographical Information Science* 22.7, pp. 801–823. DOI: 10.1080/13658810701674970.

Jain, S. and R. M. Neal (2004). "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model". In: *Journal of Computational and Graphical Statistics* 13.1, pp. 158–182. DOI: 10.1198/1061860043001.

James, L. F., A. Lijoi, and I. Prünster (2009). "Posterior Analysis for Normalized Random Measures with Independent Increments". In: *Scandinavian Journal of Statistics* 36.1, pp. 76–97. DOI: 10.1111/j.1467-9469.2008.00609.x.

Jeffreys, H. (1946). "An Invariant Form for the Prior Probability in Estimation Problems". In: *Proceedings of the Royal Society of London. Series A* 186.1007, pp. 453–461. DOI: 10.1098/rspa.1946.0056.

Lau, J. W. and P. J. Green (2007). "Bayesian Model-Based Clustering Procedures". In: *Journal of Computational and Graphical Statistics* 16.3, pp. 526–550. DOI: 10.1198/106186007X230045.

Lijoi, A., R. H. Mena, and I. Prünster (2007). "Controlling the Reinforcement in Bayesian Non-Parametric Mixture Models". In: *Journal of the Royal Statistical Society Series B* 69.4, pp. 715–740. DOI: 10.1111/j.1467-9868.2007.00609.x.

Luo, C. and A. Shrivastava (2018). "Scaling-up Split-Merge MCMC with Locality Sensitive Sampling (LSS)". In: *arXiv preprint arXiv:1802.07444*. Proceedings of the 35th International Conference on Machine Learning (ICML).

Martínez, A. F. and R. H. Mena (2014). "On a Nonparametric Change Point Detection Model in Markovian Regimes". In: *Bayesian Analysis* 9.4, pp. 823–858. DOI: 10.1214/14-BA911.

Müller, P., F. Quintana, and G. L. Rosner (2011). "A Product Partition Model With Regression on Covariates". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 150–169. DOI: 10.1198/jcgs.2011.09066.

Natarajan, A. et al. (2024). "Cohesion and Repulsion in Bayesian Distance Clustering". In: *Journal of the American Statistical Association* 119.546, pp. 1374–1384. DOI: 10.1080/01621459.2023.2191821.

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of Computational and Graphical Statistics* 9.2, pp. 249–265.

Panaretos, Victor M. and Yoav Zemel (2019). "Statistical Aspects of Wasserstein Distances". In: *Annual Review of Statistics and Its Application* 6.1, pp. 405–431. DOI: 10.1146/annurev-statistics-030718-104938. URL: https://doi.org/10.1146/annurev-statistics-030718-104938.

Rigon, T., B. Hertigu, and D. B. Dunson (2023). "A Generalized Bayes Framework for Probabilistic Clustering". In: *Journal of the American Statistical Association* 118.544, pp. 2394–2406. DOI: 10.1080/01621459.2023.2200012.

Villani, C. (2003). *Topics in Optimal Transport*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society.

Wang, W. and S. Russell (2015). "A Smart-Dumb/Dumb-Smart Algorithm for Efficient Split-Merge MCMC". In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 846–855.

A. Appendix A

This section presents visualizations of prior mean incomes for the U.S. ACS dataset referenced in Section 2.2 and the Italian municipalities dataset referenced in Section 2.3.

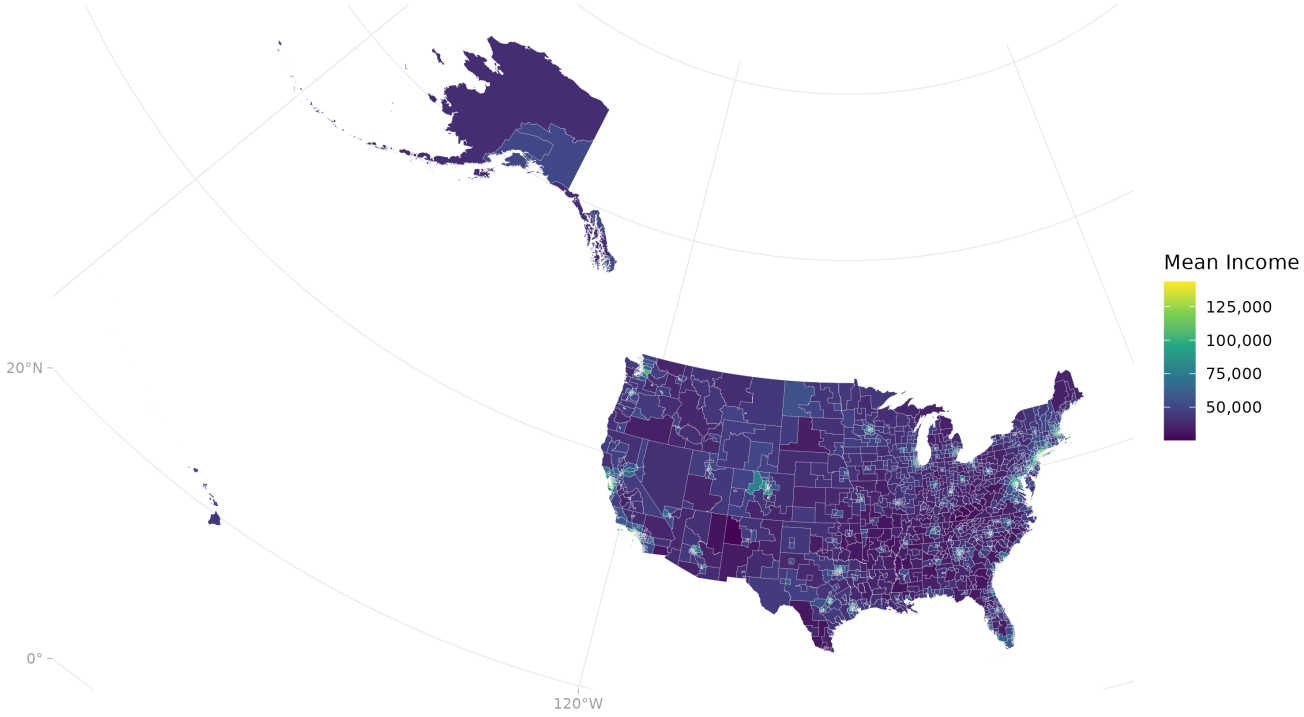


Figure 24: Average income for each areal unit in the U.S. ACS dataset described in Section 2.2.

B. Appendix B

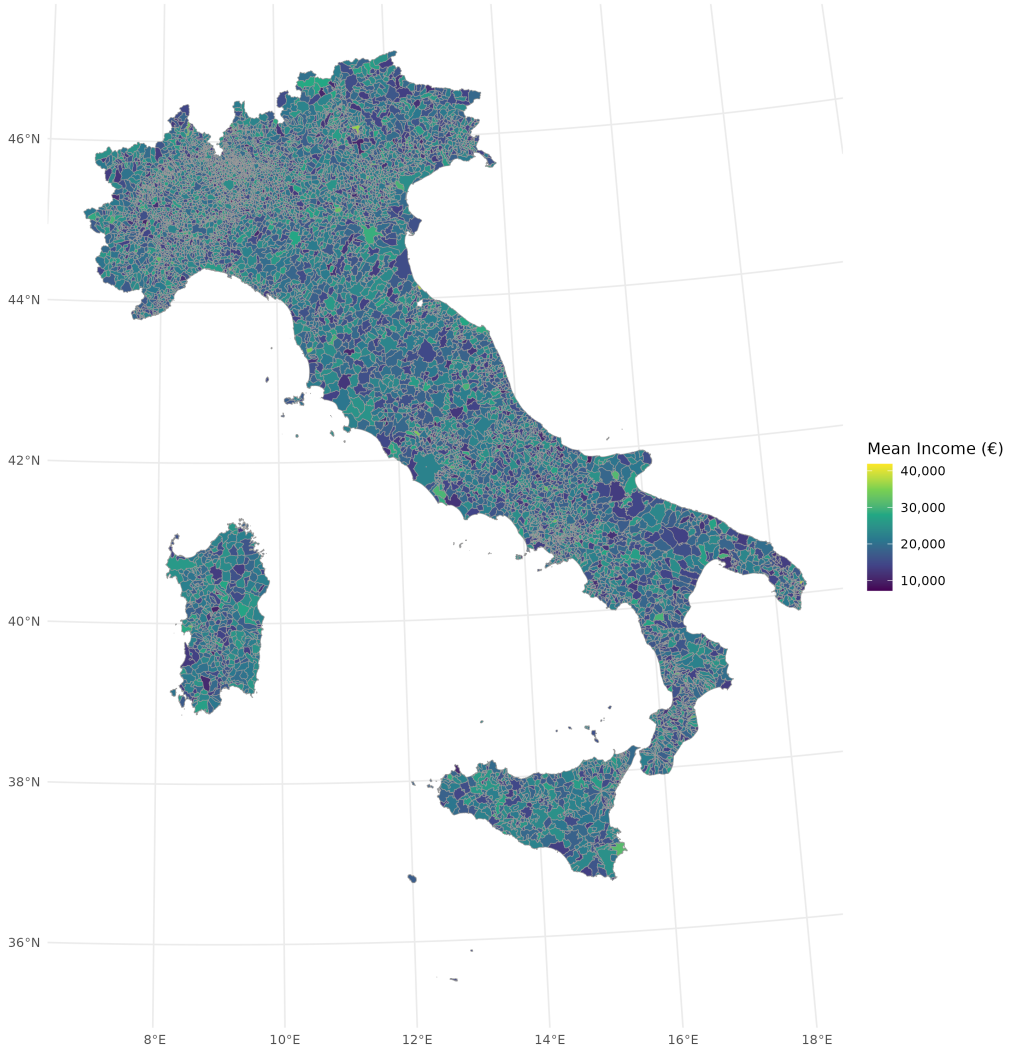


Figure 25: Average income for each areal unit in the Italian municipalities dataset described in Section 2.3.

B.1. Baseline NGGP Prior

Figure 26 displays the geographic map of posterior cluster assignments. Cluster-specific income distributions (clusters 10–23) are shown in Figure 27–Figure 28.

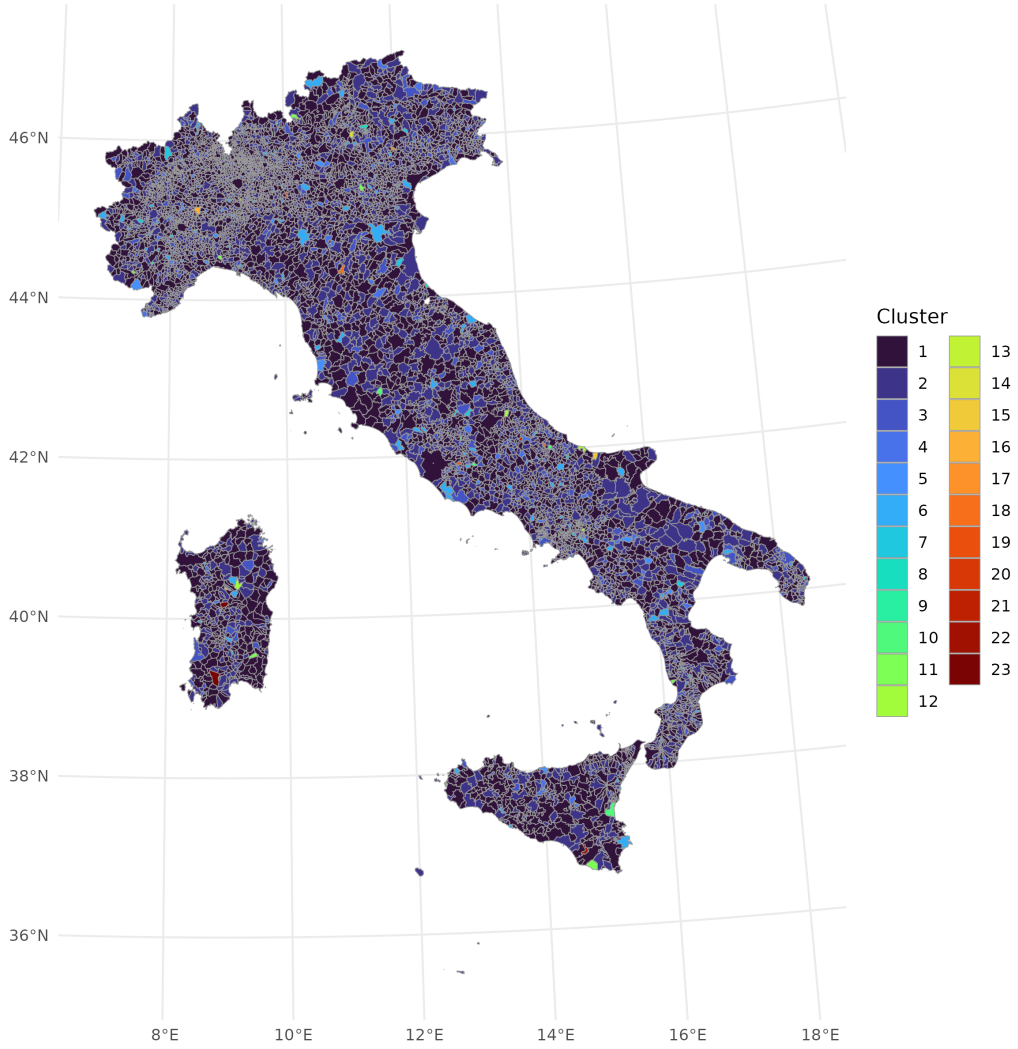


Figure 26: Geographic map of cluster assignments for the Italian municipalities dataset under the baseline NGGP prior.

B.2. NGGPW Spatial Prior

Figure 29 shows the geographic map of posterior cluster assignments. Cluster-specific income distributions (clusters 10–23) appear in Figure 30–Figure 31.

B.3. LCF Variant

Figure 32 depicts the geographic map of posterior cluster assignments. Representative cluster-specific income distributions (clusters 10–11) are provided in Figure 33.

B.4. PAM Benchmark

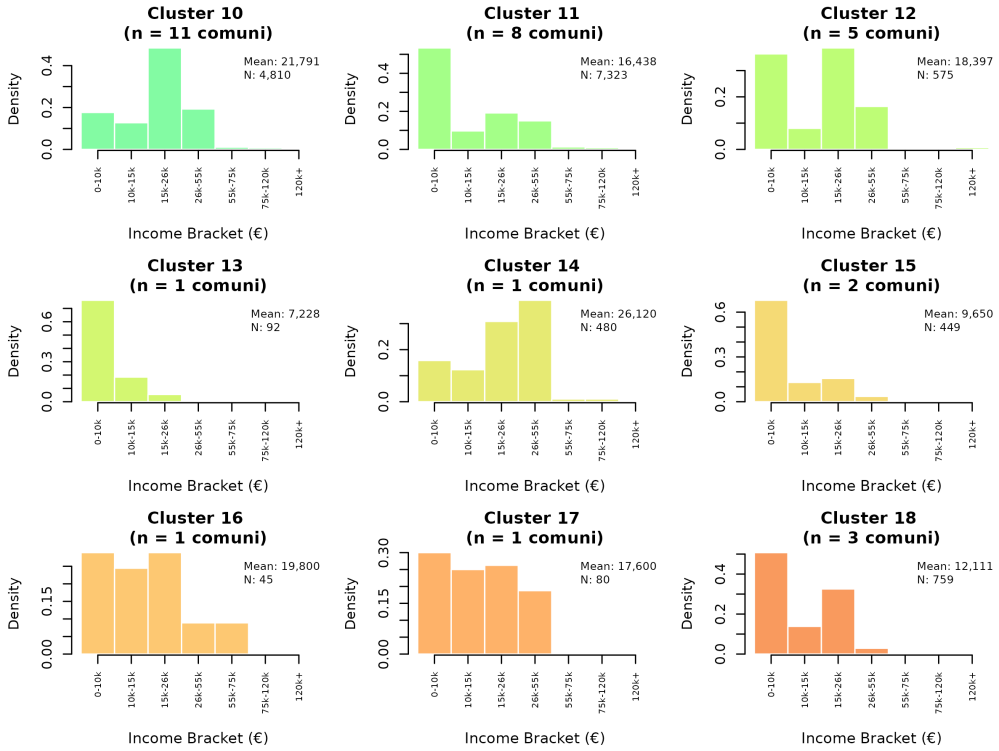


Figure 27: Cluster-specific income distributions (clusters 10–18) under the baseline NGGP prior.

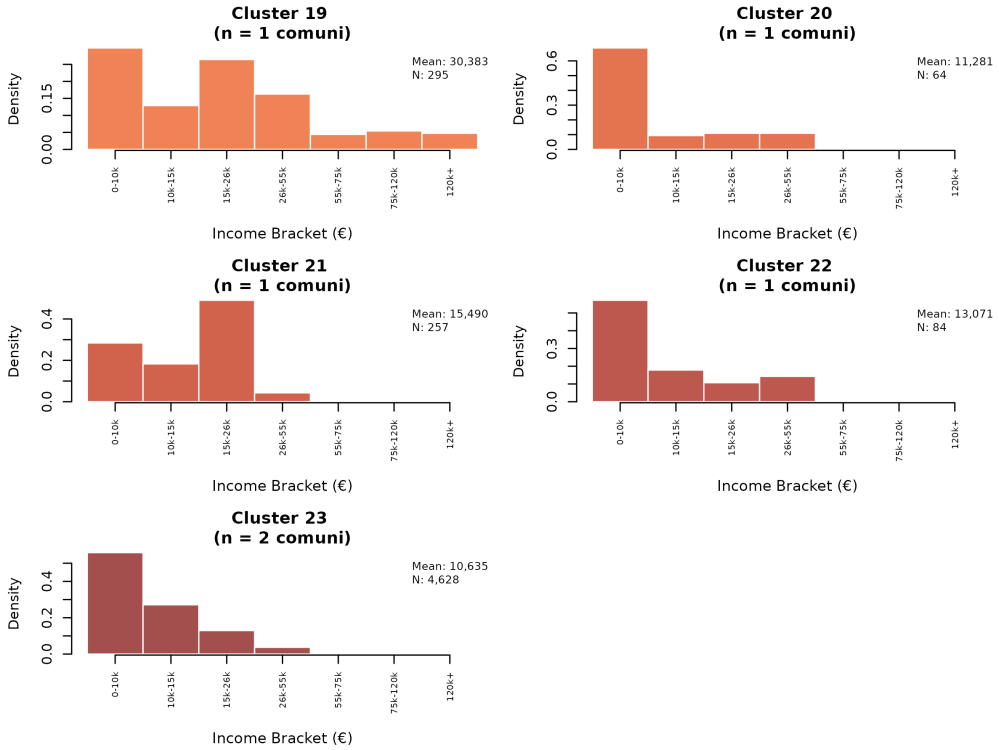


Figure 28: Cluster-specific income distributions (clusters 19–23) under the baseline NGGP prior.

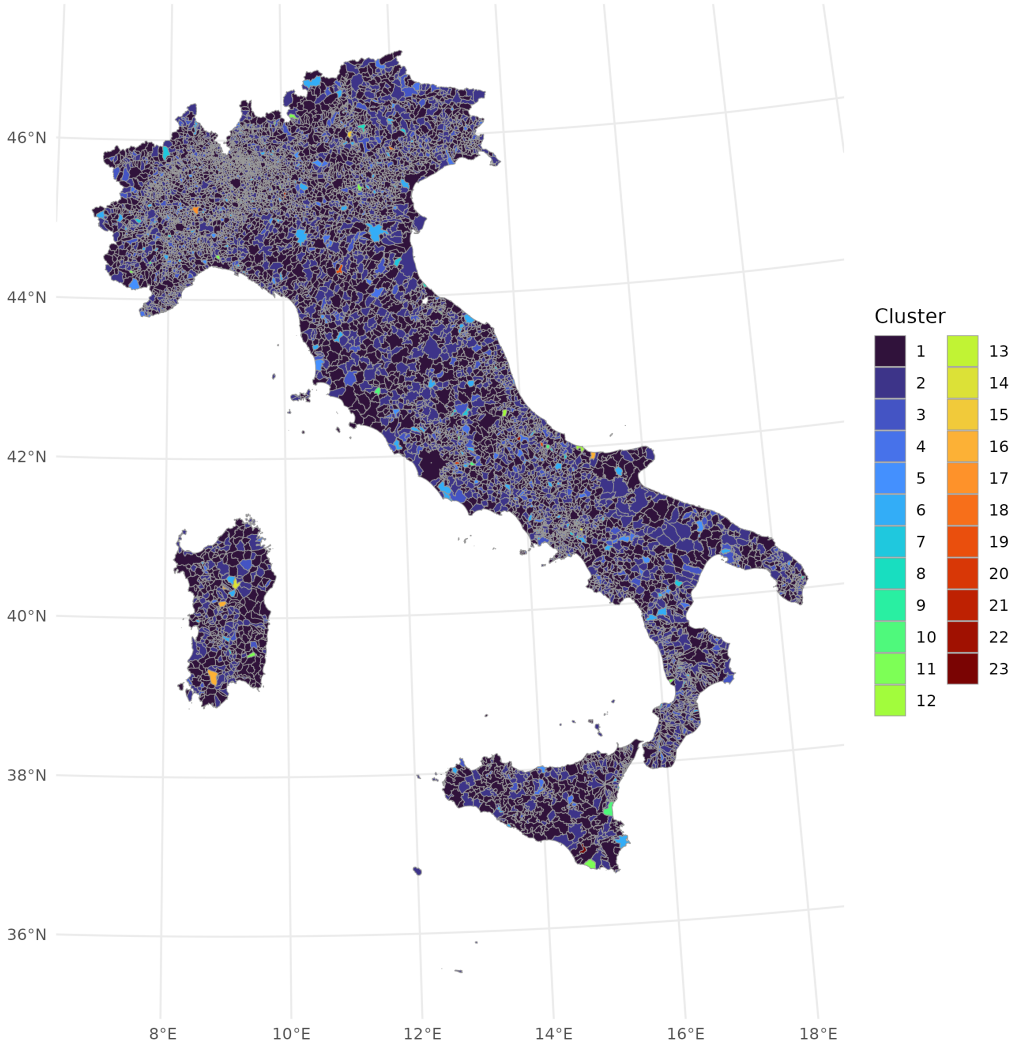


Figure 29: Geographic map of cluster assignments under the NGGPW spatial prior.

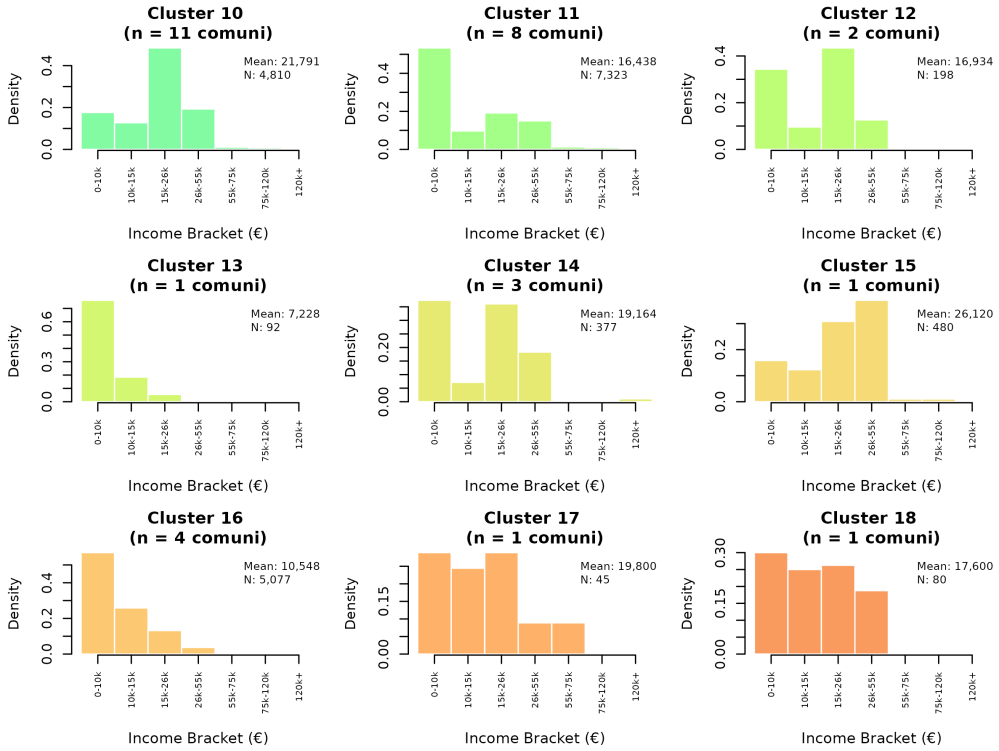


Figure 30: Cluster-specific income distributions (clusters 10–18) under the NGGPW spatial prior.

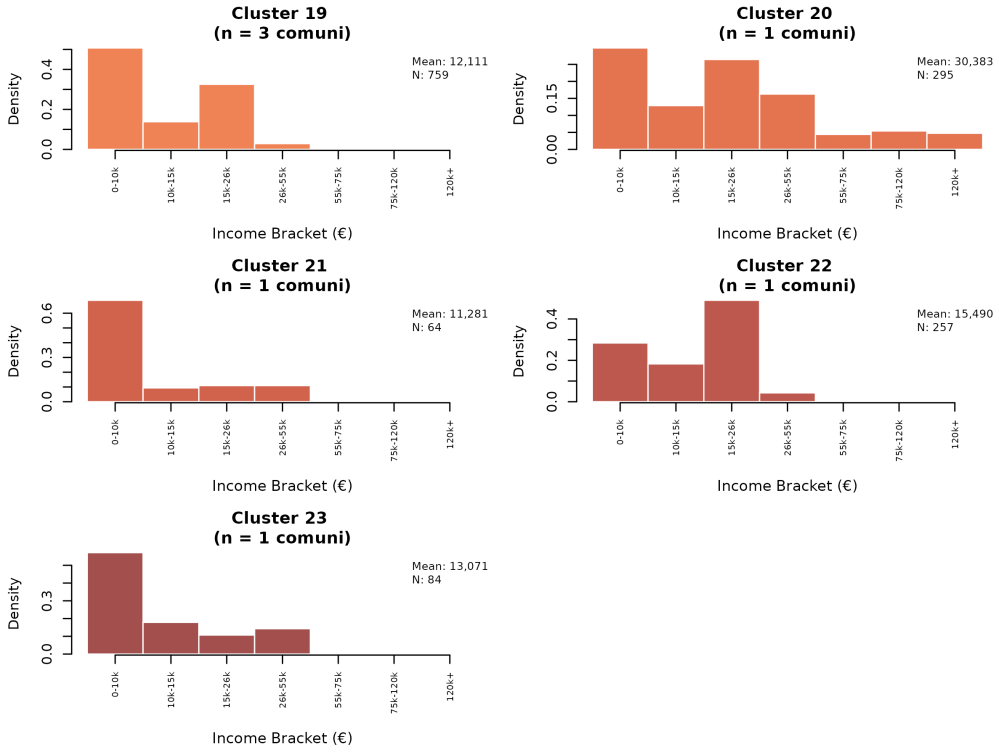


Figure 31: Cluster-specific income distributions (clusters 19–23) under the NGGPW spatial prior.

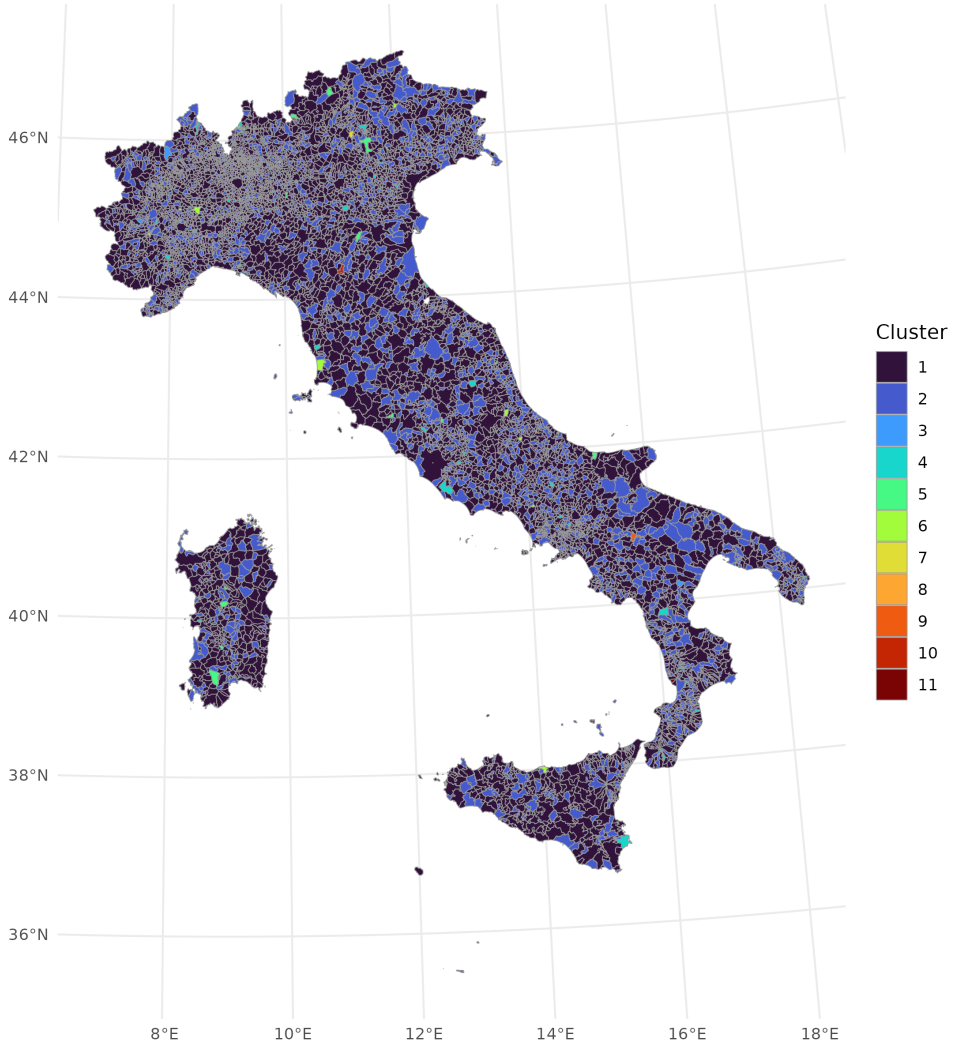


Figure 32: Geographic map of cluster assignments under the LCF variant.

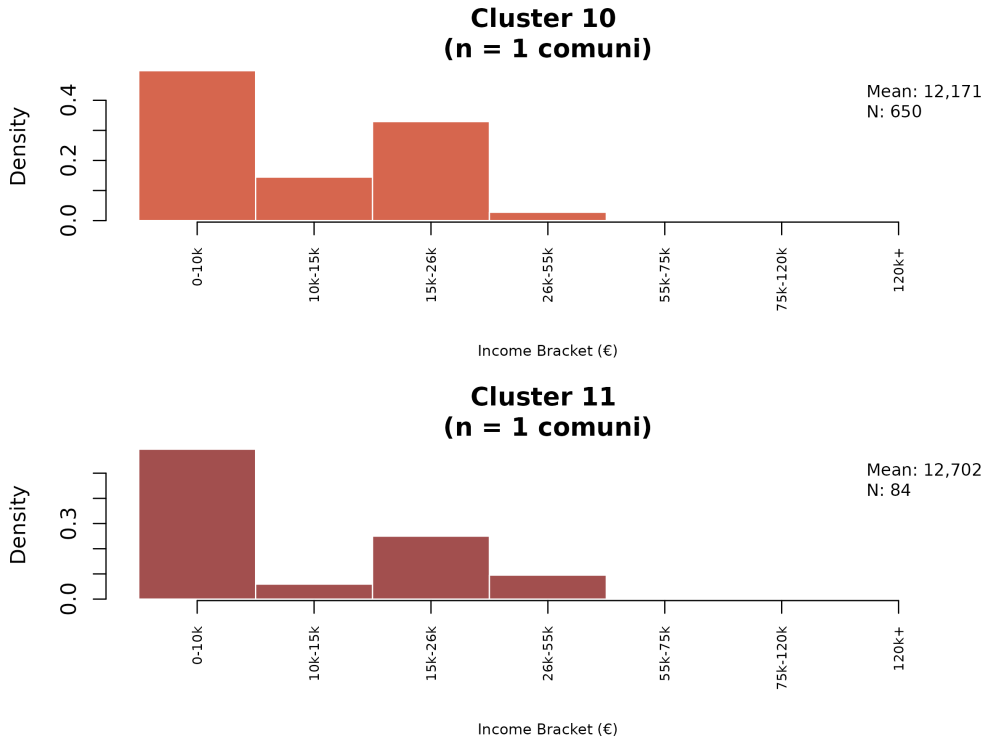


Figure 33: Cluster-specific income distributions (clusters 10–11) under the LCF variant.

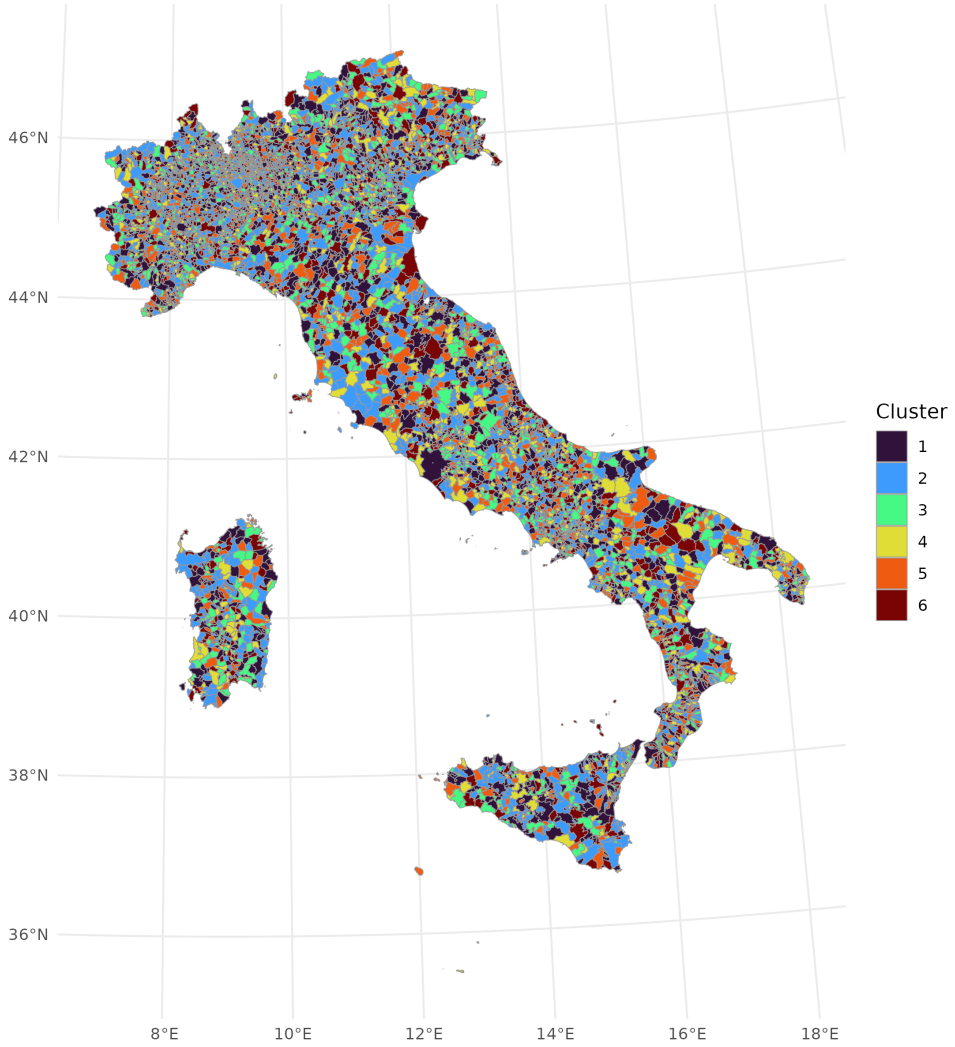


Figure 34: Geographic map of cluster assignments obtained via the PAM algorithm ($k = 6$) on the Italian municipalities dataset.

Abstract in lingua italiana

Questa tesi sviluppa e valida empiricamente un framework bayesiano non parametrico scalabile per il clustering di dati areali, basato sulle distanze a coppie tra stime di densità kernel. Motivati dall'intrattabilità computazionale dei modelli bayesiani spaziali allo stato dell'arte — che richiedono ore di calcolo anche per scale geografiche moderate — proponiamo una strategia che trasforma il problema inferenziale: da decine di migliaia di singole osservazioni individuali a un numero ridotto di rappresentazioni di densità specifiche per area e alle loro distanze a coppie. Basandoci sulla verosimiglianza coesivo-repulsiva basata sulle distanze di Natarajan et al. (2024), combiniamo una distribuzione a priori indotta da un processo Normalized Generalized Gamma (NGG) — per un'inferenza sul numero di cluster completamente guidata dai dati (data-driven) — con estensioni che incorporano la struttura di adiacenza spaziale e covariate demografiche ausiliarie. Un efficiente schema di campionamento MCMC sintetizza il campionatore Sequential Allocated Merge-Split, un adattamento basato sulle distanze del Locality-Sensitive Sampling per la selezione delle osservazioni, una strategia Smart-Dumb-Dumb-Smart per la selezione del tipo di mossa, scansioni di Gibbs periodiche e mosse di shuffle. Il framework è validato su dati simulati e applicato a due dataset reali: i dati sul reddito del censimento della California aggregati in 93 Public Use Microdata Areas (PUMA), dove l'inferenza viene completata in meno di due secondi rispetto alle dieci ore richieste dagli approcci esistenti nel campo della boundary detection, e i dati sul reddito dei comuni italiani, che comprendono 7.903 unità areali. I risultati dimostrano un'accuratezza di clustering competitiva, raggruppamenti socioeconomici interpretabili e sostanziali vantaggi computazionali in tutte le configurazioni del modello. L'intero framework è implementato come una libreria C++ modulare e ad alte prestazioni con binding in R, disponibile pubblicamente all'indirizzo https://github.com/Filippo-Galli/BNPClust.

Parole chiave: clustering bayesiano nonparametrico, dati areali, verosimiglianza distanza-basata, Processo Gamma Generalizzato Normalizzato, stima della densità kernel, clustering spaziale, MCMC, Product Partition Model