Intelligent n-Means Spatial Sampling

Bardia Panahbehagh[1], Mehdi Mohebbi[2], and Amir Mohammad HosseiniNasab[3]

[1]Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran, panahbehagh@khu.ac.ir

[2]School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran, mehdi.mohebbi@ut.ac.ir

[3]Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, s.a.m.hosseininasab@student.vu.nl

October 29, 2025

Abstract

Well-spread samples are desirable in many disciplines because they improve estimation when target variables exhibit spatial structure. This paper introduces an integrated methodological framework for spreading samples over the population's spatial coordinates. First, we propose a new, translation-invariant spreadness index that quantifies spatial balance with a clear interpretation. Second, we develop a clustering method that balances clusters with respect to an auxiliary variable; when the auxiliary variable is the inclusion probability, the procedure yields clusters whose totals are one, so that a single draw per cluster is, in principle, representative and produces units optimally spread along the population coordinates—an attractive feature for finite population sampling. Third, building on the graphical sampling framework of Panahbehagh (2025), we design an efficient sampling scheme that further enhances spatial balance. At its core lies an intelligent, computationally efficient search layer that adapts to the population's spatial structure and inclusion probabilities, tailoring a design—to each specific population to maximize spread. Across diverse spatial patterns and both equal- and unequal-probability regimes, this intelligent coupling consistently outperformed all rival spread-oriented designs on dispersion metrics, while the spreadness index remained informative and the clustering step improved representativeness.

MSC: 62D05, 62F40

Keywords: spatially balanced sampling; finite-population sampling; inclusion probabilities; k-means clustering

1 Introduction

Spreading samples is a central concern in studies involving spatially referenced data, particularly in environmental, ecological, and agricultural applications where spatial autocorrelation is prevalent. A well-spread sample improves domain coverage and thereby enhances the precision of estimators for spatially structured variables. Moreover, spatial spread often coincides with

balance on auxiliary variables, which is crucial when target parameters depend nonlinearly on those covariates (Grafström, 2013).

A substantial body of work has proposed designs that explicitly target spatial balance. One influential line maps a multidimensional population to a one-dimensional, neighborhood-preserving index and then applies systematic selection, as in Generalized Random Tessellation Stratified sampling (GRT; Stevens Jr. and Olsen, 2004). A second line induces repulsion among nearby units to discourage joint selection: the Local Pivotal Method (LPM; Grafström, Lundström and Schelin, 2012) iteratively updates neighboring pairs so that exactly one is retained, while Spatially Correlated Poisson sampling (SCP; Grafström, 2012) introduces dependence structures that reduce the likelihood of selecting adjacent units. Weakly Associated Vectors (WAV; Jauslin, 2020) extend this principle by iteratively perturbing the inclusion–probability vector along directions weakly associated with spatial proximity, preserving first-order inclusion probabilities while progressively driving its components to the boundary $\{0, 1\}$. Partitioning and quasi-random constructions form another family, such as Halton Iterative Partitioning (HIP; Robertson et al., 2018), which achieves low-discrepancy coverage through recursive subdivisions of the spatial domain. Recently, optimization-based approaches such as Dynamic Assignment Sampling (DAS; Robertson, Price and Reale, 2024) have emerged, combining geometric spread with adaptive control of sample size.

Assessing "how well-spread" a sample is typically relies on indices such as Moran (Moran, 1950; Tillé, Dickson, Espa and Giuliani, 2018), which measures spatial autocorrelation, the Voronoi index (Stevens Jr. and Olsen, 2004), and the Balanced Voronoi index (Prentius and Grafström, 2024). While informative, these indices can be insensitive to certain spatial configurations, may lack directional sensitivity—e.g., distinguishing over-concentration from over-dispersion—and may not be invariant to uniform translations of the sample configuration.

Building upon this background, the present study introduces a new sampling method that serves as a competitive alternative to existing spatially balanced designs. Our framework provides a unified approach to measuring, constructing, and exploiting spatial balance under both equal- and unequal-inclusion probability settings (EP and UP). Specifically, we (i) introduce the translation-invariant density disparity index, which evaluates a sample relative to a population-specific optimal configuration and is sensitive to both under- and over-dispersion; (ii) develop a n-means balanced clustering scheme that forms probability-balanced, spatially compact groups—when the auxiliary variable is the inclusion probability, each cluster is in principle representative with a sample of size one, yielding units optimally spread along the population index; and (iii) design an efficient sampling algorithm—built on the graphical framework of Panahbehagh (2025)—that enhances spatial balance while supporting design-based inference. The method incorporates an intelligent search component that examines the spatial configuration and inclusion probabilities of the population to tailor a sampling design specifically suited to its structure, thereby improving spread and representativeness without added computational burden.

Together, the three components—the density disparity index, n-means balanced clustering, and the graphical sampling method—form an efficient and effective framework for both measuring and achieving spatial spreadness of sample selection; moreover, when coupled with a lightweight

intelligent search that dynamically reorders the design, the procedure performs population-specific optimization of spread by adapting to the observed coordinates and inclusion probabilities, thereby delivering locally tailored dispersion gains without altering first-order inclusion probabilities.

To this end, the paper is organized as follows. Section 2 introduces notation and design-based preliminaries for finite-population sampling. Section 3 reviews existing indices for assessing spatial balance. Section 4 presents n-means balanced clustering. Section 5 introduces our new measure of spatial balance, the density disparity index, and establishes its properties. In Section 6, we adapt the graphical sampling framework of Panahbehagh (2025) to the spatial setting. Section 7 details the proposed sampling design, and Section 8 reports the simulation results. Finally, Section 9 offers concluding remarks.

2 Notations and Basic Concepts

Let $U = \{1, \ldots, N\}$ denote a finite population of size N, and let $\mathcal{S}$ be the set of all possible samples (subsets) of U. The objective is to select a sample $S \in \mathcal{S}$ with potentially unequal first-order inclusion probabilities $\boldsymbol{\pi} = \{\pi_\ell, \ell \in U\}$ and expected size n_S, where

$$\sum_{\ell \in U} \pi_\ell = E(n_S).$$

Let $\boldsymbol{p} = \{p_s = \Pr(S = s);\ s \in \mathcal{S}\}$ denote a sampling design that assigns a probability to each possible sample such that

$$\sum_{s \in \mathcal{S}} p_s = 1, \quad \text{and} \quad \sum_{s \in \mathcal{S},\, s \ni \ell} p_s = \pi_\ell \quad \text{for all } \ell \in U. \tag{1}$$

In this article, our focus is on designs with fixed sample size n, for which $E(n_S) = n$ and $\operatorname{var}(n_S) = 0$.

Consider y_ℓ and x_ℓ $(\ell \in U)$ as the main and auxiliary variables, respectively. For any subset $A \subset U$, let $\boldsymbol{a}(A) = \{a_\ell(A), \ell \in U\}$ denote a vector of indicators where $a_\ell(A) = 1$ if $\ell \in A$ and 0 otherwise. The main objective of sampling is typically to estimate parameters such as the population total,

$$Y(A) = \sum_{\ell \in U} y_\ell a_\ell(A),$$

which can be defined similarly for the auxiliary variable x. Estimation of such parameters and their variances is based on the sampling design, the first-order inclusion probabilities, and the second-order inclusion probabilities. An unbiased estimator of Y is the Narain–Horvitz–Thompson (NHT) estimator (Narain, 1951; Horvitz and Thompson, 1952):

$$\hat{Y}(A) = \sum_{\ell \in U} \frac{y_\ell}{\pi_\ell} a_\ell(A \cap S),$$

with variance

$$\operatorname{var}(\hat{Y}(A)) = \sum_{\ell \in U} \sum_{\ell' \in U} \frac{y_\ell}{\pi_\ell} \frac{y_{\ell'}}{\pi_{\ell'}} (\pi_{\ell\ell'} - \pi_\ell \pi_{\ell'}) a_\ell(A) a_{\ell'}(A), \tag{2}$$

and an unbiased estimator given by

$$\widehat{\text{var}}(\hat{Y}(A)) = \sum_{\ell \in U} \sum_{\ell' \in U} \frac{y_\ell}{\pi_\ell} \frac{y_{\ell'}}{\pi_{\ell'}} \frac{\pi_{\ell\ell'} - \pi_\ell \pi_{\ell'}}{\pi_{\ell\ell'}} a_\ell(A \cap S) a_{\ell'}(A \cap S), \tag{3}$$

or a more stable estimator for fixed-size designs:

$$\widehat{\text{var}}(\hat{Y}(A) = \frac{1}{2} \sum_{\ell \in U} \sum_{\ell' \in U} \left(\frac{y_\ell}{\pi_\ell} - \frac{y_{\ell'}}{\pi_{\ell'}} \right)^2 \frac{\pi_{\ell\ell'} - \pi_\ell \pi_{\ell'}}{\pi_{\ell\ell'}} a_\ell(A \cap S) a_{\ell'}(A \cap S). \tag{4}$$

Each population unit $\ell \in U$ is associated with a spatial coordinate denoted by $\boldsymbol{c}_\ell$. The spatial configuration of all population units in a D-dimensional space is represented by the matrix

$$\boldsymbol{C} = \{\boldsymbol{c}_\ell, \ell \in U\}, \quad \text{where} \quad \boldsymbol{c}_\ell = (\boldsymbol{c}_{\ell d},\, d = 1, 2, \ldots, D),$$

and $\boldsymbol{c}_{\ell d}$ denotes the coordinate of unit ℓ along the dth dimension. Here, D represents the dimensionality of the spatial domain under consideration.

3 Spatial Balance Indices

In this section, we discuss three indices that have been introduced to measure the spatial spread of samples: the Moran Index, the Voronoi Index, and the Balanced Voronoi Index.

3.1 Voronoi Index (VI)

Stevens Jr. and Olsen (2004) introduced the Voronoi Spatial Balance Measure to assess the spatial distribution of sample units within a population. Let $\{U_\ell\}_{\ell \in S}$ denote the Voronoi partition of the population U around the sample units S, such that U_ℓ is the set of population units closer to sample unit $\ell \in S$ than to any other. The VI measures how well the sample S is distributed across the population U and is defined as

$$\text{VI}(S) = n \sum_{\ell \in S} \left(\hat{G}_\pi(U_\ell) - G_\pi(U_\ell) \right)^2,$$

where

$$G_x(U^*) = F_x(U^*), \quad \hat{G}_x(U^*) = \frac{1}{X(U)} \sum_{\ell \in U} \frac{x_\ell}{\pi_\ell} a_\ell(U^*), \quad \text{and} \quad F_x(U^*) = \frac{X(U^*)}{X(U)}.$$

This index reflects the deviation of the observed inclusion probabilities within each Voronoi cell from their expected values under a uniform distribution.

3.2 Balanced Voronoi Index (BI)

Prentius and Grafström (2024) proposed an improved Voronoi–based measure that explicitly accounts for the balance of auxiliary variables within each Voronoi region. This enhanced index is computed as a weighted sum of squared discrepancies between the design-expected and design-estimated auxiliary totals within the respective regions, thereby capturing both

spatial spread and covariate balance. To balance the VI on covariates $\boldsymbol{x}$, define the scaled cellwise discrepancy $\boldsymbol{r}_{\boldsymbol{x}}(U_\ell^*)$ as the difference between the HT estimate and its design expectation (optionally scaled by a linear operator $X(U)$ to ensure commensurability of components). The index is then

$$\mathrm{BI}(S) := \sqrt{\frac{1}{n}\sum_{\ell\in S} \boldsymbol{r}_{\boldsymbol{x}}(U_\ell^*)^\top \mathbf{Q}^{-1} \boldsymbol{r}_{\boldsymbol{x}}(U_\ell^*)},$$

where discrepancies are aggregated via a positive definite weighting matrix $\mathbf{Q}$. Typical choices include $\mathbf{Q} = \mathrm{Cov}(\boldsymbol{r}_{\boldsymbol{x}}(U_\ell^*))$ (design- or model-assisted) or a diagonal variance matrix, with $\mathbf{Q} = \mathbf{I}$ recovering an unweighted Euclidean aggregation.

The use of $\mathbf{Q}$ instead of the covariance matrix is proposed to allow the inclusion of balancing with respect to the size of the neighbourhood. A specific choice for $\mathbf{Q}$ in this context is $\mathbf{Q} = \mathbf{x}^\top\mathbf{x}$, where $\mathbf{X}$ is the Gram matrix of $\mathbf{x} = [1, x_1, \ldots, x_p]$. The resulting measure, related to the **Mahalanobis distance**, measures the squared discrepancies in the neighbourhoods around the sample units, with respect to $\mathbf{Q}$. Also, n the univariate case,

$$r_x(U^*) := X(U)\big[\widehat{G}_x(U^*) - G_x(U^*)\big],$$

and a lower value of $\mathrm{BI}(S)$ indicates better balance of auxiliary variables across Voronoi cells.

3.3 Moran Index (MI)

Moran (1950) proposed a measure of spatial correlation, particularly suitable for data arranged on a grid. Tillé, Dickson, Espa and Giuliani (2018) introduced a normalized formulation using a spatial weight matrix W. Various approaches exist for defining W, as detailed by Jauslin (2020) and Tillé, Dickson, Espa and Giuliani (2018).

MI quantifies spatial autocorrelation within a sample S, indicating whether similar values cluster or disperse spatially. It is defined as

$$\mathrm{MI}(S) = \frac{[\boldsymbol{a}(S) - \bar{\boldsymbol{a}}(S)]^\top W [\boldsymbol{a}(S) - \bar{\boldsymbol{a}}(S)]}{(1^\top W 1)\,[\boldsymbol{a}(S) - \bar{\boldsymbol{a}}(S)]^\top [\boldsymbol{a}(S) - \bar{\boldsymbol{a}}(S)]},$$

where $W = [w_{\ell\ell'}]_{N\times N}$ is the spatial weight matrix with $w_{\ell\ell'} = 1/e_{\ell\ell'}$ for nearby units and $w_{\ell\ell'} = 0$ otherwise, and $e_{\ell\ell'}$ denotes the distance between units ℓ and ℓ'. Also,

$$\bar{\boldsymbol{a}}(S) = (\bar{a}(S), \ldots, \bar{a}(S)), \qquad \bar{a}(S) = \frac{1}{N}\sum_{\ell\in U} a_\ell(S).$$

A positive $\mathrm{MI}(S)$ indicates positive spatial autocorrelation (similar values cluster), while a negative value indicates negative spatial autocorrelation (dissimilar values are adjacent).

3.4 Discussion

While the above indices—Voronoi, Balanced Voronoi, and Moran's—provide valuable insight into the spatial structure of a sample, each focuses on a limited aspect of spatial configuration, such as clustering, dispersion, or autocorrelation. They may be sensitive to edge effects, scale dependencies, or fail to reflect the uniformity of coverage across the study region. As shown in

Section 5, these limitations can lead to suboptimal assessments of spatial representativeness in certain scenarios.

To address these shortcomings, we introduce a novel index that approaches the problem of spreadness from a new perspective. By leveraging the cluster structure of the population prior to sampling, our proposed measure provides a more comprehensive evaluation of how evenly a sample is distributed across space, integrating both local and global dispersion patterns to yield a more robust assessment of spatial representativeness. To define the new index and to lay the groundwork for a sampling design that achieves well-spread samples, we first introduce a clustering procedure for the population units; the next section develops this construction in detail.

4 n-Means UP-balanced clustering

A central challenge in spatial and design-based sampling from finite populations is to construct clusters that achieve both geographic dispersion and precise control over inclusion probabilities. Classical k-means clustering, rooted in Lloyd's least-squares quantization, optimizes within-cluster compactness without regard to design constraints (Lloyd, 1982). Constrained and balanced variants of k-means address empty clusters and enforce (near-)equal cluster sizes (Bradley, Bennett and Demiriz, 2000; de Maeyer, Sieranoja and Fränti, 2023), yet they do not guarantee control of cluster-level sums of prescribed auxiliary totals (e.g., inclusion probabilities), which is essential for unbiased design-based inference with unequal probabilities. In parallel, the spatial sampling literature has developed probability designs that explicitly promote spatial spread while respecting inclusion probabilities—most notably GRTS (Stevens Jr. and Olsen, 2004) and pivotal-method families that yield spatial balance and, in their doubly balanced forms, near-exact restitution of auxiliary totals (Grafström, Lundström and Schelin, 2012; Grafström and Tillé, 2013). However, these are selection mechanisms rather than partitioning schemes: they select samples directly but do not produce cluster partitions in which each group meets a target auxiliary total. This gap motivates *auxiliary-total-balanced clustering*, which partitions U into spatially compact groups whose auxiliary totals match prescribed targets, thereby aligning clustering with design-based objectives and enabling one-per-cluster selection that is simultaneously well spread and probability-consistent.

In this section, we introduce a practical algorithm, termed *n-Means UP-balanced clustering*, to partition a finite population U with spatial coordinates $\boldsymbol{C}$ and inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$, where $\sum_{\ell=1}^{N} \pi_\ell = n$, into n clusters $\{U_1, \ldots, U_n\}$ such that $\sum_{\ell \in U_i} \pi_\ell = 1$ for $i = 1, \ldots, n$. This construction enables "one-per-cluster" sampling designs that are simultaneously spatially well-spread and consistent with the specified inclusion probabilities.

The core idea is to convert the problem of balancing UP into a tractable size-balanced clustering problem via data expansion. For each unit ℓ with UP π_ℓ, define an integer

$$\texttt{count}_\ell = \max(1,\ \text{round}(\pi_\ell/\delta)),$$

where $\delta > 0$ is a small *split-size* parameter. We then construct an expanded dataset by replicating coordinate c_ℓ exactly $\texttt{count}_\ell$ times and record the mapping from each pseudocopy to its source

unit ℓ.

Next, we apply size-constrained n-means clustering (e.g., via `KMeansConstrained`[1]) to the expanded dataset, yielding n clusters of nearly equal size and cluster labels for all pseudocopies.

Aggregating the pseudocopy labels back to the original units yields, for each unit ℓ, the fraction of its pseudocopies assigned to each cluster. Define the *soft membership* matrix $M \in [0,1]^{N\times n}$ by

$$M[\ell, i] := \frac{\text{number of pseudocopies of unit } \ell \text{ assigned to cluster } i}{\texttt{count}_\ell}.$$

Then $\sum_i M[\ell, i] = 1$ and the cluster-attributed inclusion probabilities satisfy $\sum_i \pi_\ell M[\ell, i] = \pi_\ell$. A *hard membership* may be obtained as $i^*(\ell) = \arg\max_i M[\ell, i]$ (breaking ties by proximity to centroids), which induces a disjoint partition of the units.

While soft memberships preserve probabilistic mass exactly at the unit level, they may create multiple boundary units per cluster, which is undesirable for subsequent use. Hard memberships avoid fractional boundaries but can produce clusters whose total inclusion probabilities deviate from 1. Our aim is therefore to retain the probabilistic fidelity of the soft assignments while enforcing, downstream, clusters that each carry total probability exactly equal to 1 and have at most a single boundary.

To facilitate this, we impose a coherent global order on the preliminary clusters. Let $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n$ denote the centroids computed from the hard memberships. We obtain an ordered sequence by approximating the shortest open path through these centroids (nearest-neighbor initialization followed by a light 2-opt refinement suffices in practice). This yields an arrangement $U_{(1)}, \ldots, U_{(n)}$ in which consecutive clusters are geometrically proximate, encouraging natural joins at interfaces.

Conditioned on this cluster order, the within-cluster arrangement of units is aligned with neighboring clusters. For each ordered cluster $U_{(i)}$, we compute, for its members, their 1-nearest-neighbor distances to $U_{(i-1)}$ and $U_{(i+1)}$ (when defined), and sort $U_{(i)}$ by a simple contrast (e.g., distance to predecessor minus distance to successor). The member closest to $U_{(i-1)}$ is pinned as the "left" endpoint, and the member closest to $U_{(i+1)}$ as the "right" endpoint, so that concatenating consecutive blocks incurs minimal geometric discordance.

Concatenating the ordered blocks produces a single total order of all units. Traversing this order, we form cumulative sums of the original inclusion probabilities π_ℓ and place quota thresholds at integer levels $(1, 2, \ldots, n)$. A threshold that falls strictly between two units closes one cluster and opens the next. If a threshold falls within a unit of mass π_ℓ, that unit is provisionally split fractionally between the adjacent clusters according to the portion of π_ℓ needed to satisfy the left quota. This yields crisp, interpretable borders while preserving target masses. Consequently, each cluster U_i satisfies $\sum_{\ell\in U_i} \pi_\ell = 1$, producing an n-way partition that is UP-balanced and spatially well-organized for one-per-cluster selection.

The details are summarized in Algorithm 1.

Conceptually, this generalizes to any auxiliary variable—such as EP—without loss of structure or interpretation, since the same balancing principle applies regardless of the weighting scheme. It ensures design consistency by forming clusters whose compositions mirror the prescribed

[1] https://pypi.org/project/k-means-constrained/

Algorithm 1 n-Means UP-balanced clustering

Require: Coordinates $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N)$, UP $\boldsymbol{\pi}$ with $\sum_{\ell=1}^{N} \pi_\ell = n$, number of clusters n, expansion parameter δ

Ensure: Centroids $\{\boldsymbol{o}_k\}_{k=1}^{n}$, membership matrix $M \in \mathbb{R}^{N \times n}$ (soft)

Expand Data to Encode UP

for $\ell = 1, \ldots, N$ **do**

    $\texttt{count}_\ell \leftarrow \max(1, \mathrm{round}(\pi_\ell/\delta))$

end for

Construct the expanded coordinate sequence $\boldsymbol{C}_{\exp} = (\boldsymbol{c}_{e(1)}, \ldots, \boldsymbol{c}_{e(N_{\exp})})$ by repeating each $\boldsymbol{c}_\ell$ exactly $\texttt{count}_\ell$ times

Define $I_{\exp} : \{1, \ldots, N_{\exp}\} \to \{1, \ldots, N\}$ by $I_{\exp}(u) = \ell$ iff $\boldsymbol{c}_{e(u)} = \boldsymbol{c}_\ell$ ▷ maps each expanded index to its source unit

Constrained n-Means Clustering

Apply equal-size constrained n-means to $\boldsymbol{C}_{\exp}$ to obtain labels $L_{\exp}(u) \in \{1, \ldots, n\}$ for $u = 1, \ldots, N_{\exp}$

Soft Membership (per original unit)

Initialize $M \leftarrow \mathbf{0}_{N \times n}$

for $\ell = 1, \ldots, N$ **do**

    $J_\ell \leftarrow \{\, u \in \{1, \ldots, N_{\exp}\} : I_{\exp}(u) = \ell \,\}$

    $M[\ell, k] \leftarrow |J_\ell|^{-1} \sum_{u \in J_\ell} \mathbb{I}(L_{\exp}(u) = k) \quad$ for $k = 1, \ldots, n$

end for

Preliminary Hard Labels and Centroids

$z_\ell^{\mathrm{raw}} \leftarrow \arg\max_k M[\ell, k] \quad$ (ties by nearest mean in $\{\boldsymbol{c}_\ell\}$)

$\boldsymbol{o}_k \leftarrow \mathrm{mean}\{\, \boldsymbol{c}_\ell : z_\ell^{\mathrm{raw}} = k \,\} \quad$ for $k = 1, \ldots, n$

Order Clusters Along a Path

Obtain an ordered sequence $(U_{(1)}, \ldots, U_{(n)})$ by approximating the shortest open path through $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n\}$

Within-Cluster Alignment and Total Order Φ

for $i = 1, \ldots, n$ **do**

    Sort $U_{(i)}$ by a 1-NN contrast to $U_{(i-1)}$ and $U_{(i+1)}$ (when defined), pinning endpoints nearest to neighbors

end for

Concatenate the ordered blocks to form the total order $\Phi : \{1, \ldots, N\} \to \{1, \ldots, N\}$; $\Phi(i)$ is the i-th ordered unit

Quota Split by Cumulative UP

Form $sum(t) = \sum_{u \le t} \pi_{\Phi(u)}$; place thresholds at $\{1, 2, \ldots, n\}$

If a threshold falls strictly between $\Phi(t)$ and $\Phi(t{+}1)$, cut there; if it falls within $\Phi(t)$, split $\pi_{\Phi(t)}$ fractionally between adjacent clusters

Snap and Promote Borders

Snap near-degenerate fractional shares to $\{0, 1\}$ under tight tolerances; if one side receives 1, promote $\Phi(t)$ fully to that side (no fractional border)

Final Centroids by Hard Clustering

Convert free/border allocations to per-cluster fractions; set $z_\ell \leftarrow \arg\max_k$ (fraction of unit ℓ in cluster k)

$\boldsymbol{o}_k \leftarrow \mathrm{mean}\{\, \boldsymbol{c}_\ell : z_\ell = k \,\}$ for $k = 1, \ldots, n$

return $\{\boldsymbol{o}_k\}_{k=1}^{n}$, M

inclusion probabilities, thereby enabling feasible, calibrated one-per-cluster selection. Using size-constrained n-means that preserves spatial compactness, the clusters remain geographically coherent and deliver strong spatial balance. Algorithmically, the method is practical: it recasts inclusion-probability balancing as a tractable cardinality-constrained clustering problem solvable with standard routines, and the subsequent construction of the total order is likewise fast and scalable. The resulting procedure yields soft allocations in which consecutive ordered clusters have at most one border unit and each cluster attains total probability exactly 1. In sum, n-Means UP-balanced clustering provides a systematic, extensible basis for constructing spatial strata or clusters in complex surveys, maintaining spatial coherence and enabling design-based, well-spread one-per-cluster sampling.

Example

Figures 1 and 2 illustrate the performance of the proposed n-Means UP-balanced clustering compared with alternative methods under different population structures and probability regimes.

Figure 1 displays three canonical spatial populations of size 100: gridded (left), random (middle), and clustered (right). Each layout is shown under equal probabilities (EP, top row) and unequal probabilities (UP, bottom row). In the UP setting, inclusion probabilities increase monotonically from left to right across the domain, visualized by larger point sizes. This configuration was chosen to highlight the comparative behavior of different methods.

Building on these populations, Figure 2 reports cluster assignments produced by four clustering strategies. The first two rows correspond to the proposed n-means EP- and UP-balanced clustering respectively (in general, we refer to both as n-means balanced clustering); the third and fourth rows display results for size-balanced clustering and conventional k-means. Colored polygons delineate cluster partitions, and annotated values indicate the total inclusion probability (cluster weight) in each case.

The results highlight a clear advantage of the proposed method. The n-means balanced clustering attains both broad spatial coverage and tight probability balance, with cluster totals summing to one across scenarios. By contrast, size-balanced and conventional k-means exhibit visible imbalances, particularly under heterogeneous spatial layouts or strongly unequal probabilities. This example underscores n-means balanced clustering as a practical tool for finite-population sampling, yielding well-spread clusters while respecting prescribed inclusion probabilities.

In addition to its clustering interpretation, the n-means balanced clustering can be used directly as a sampling device: each polygon in Figure 2 defines a region from which exactly one unit may be selected, and the balanced construction ensures that this unit is representative of the population within the polygon. In this way, the method partitions the population into spatially compact and probability-balanced clusters, so that subsequent one-per-cluster sampling yields a well-spread design that reflects both spatial structure and the specified inclusion probabilities.

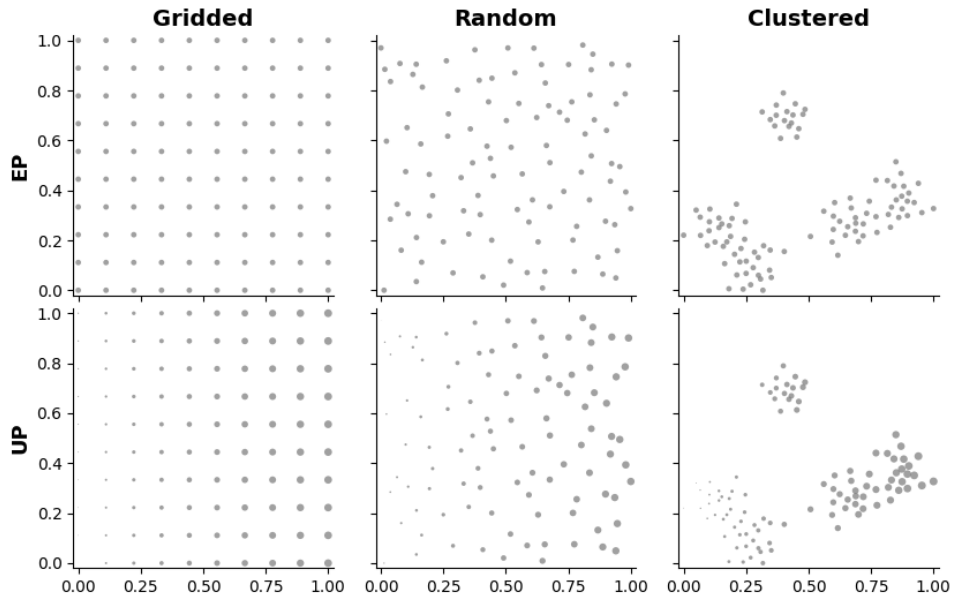

Figure 1: Simulated populations of size 100 under three spatial layouts: gridded (left), random (middle), and clustered (right). The top row shows equal probabilities (EP); the bottom row shows unequal probabilities (UP), where inclusion probabilities increase monotonically from left to right (larger point size indicates higher probability).

5 A New Index for Spread of Sample

In this section, we introduce a novel index for assessing the spatial spread of a sample. The core idea is that partitioning the population into clusters, each with a total inclusion probability of one, provides a meaningful representation of the population's spatial structure for evaluating spread, and the centroids of these clusters constitute spread-optimal samples for that population. Building on this idea, we assess spread via the population's spatial density distribution: after matching each sample unit to a cluster centroid, we translate each cluster so that its centroid coincides with its assigned unit and then compare the original and translated density surfaces; the closer these distributions, the better the sample's spread, with near-invariance indicating a well-spread sample.

This index exhibits several key properties that distinguish it from existing measures of spatial spread. First, it provides clear, population-specific insight into what optimal spread samples look like, as these correspond directly to the centroids obtained from n-Means UP-balanced clustering of the population. In contrast, for indices such as MI or Voronoi-based, the form of an optimal spread sample is neither explicit nor practically identifiable. Second, the index reveals whether a lack of spread arises from sample units being too close to one another or too far apart; such diagnostic information is absent in existing indices. Third, the index is *spatially translation invariant*: the spread value of a sample remains unchanged if all its units are uniformly translated by the same vector, ensuring that absolute position does not affect the index. For example, in Figure 3, the first panel shows a sample that coincides with the benchmark centroids and is therefore judged optimally spread by the index; uniform translations of this arrangement in the subsequent panels receive the same spread value and are likewise

Figure 2: Cluster assignments for simulated populations of 100 units arranged on gridded, random, and clustered layouts. Methods compared are n-means balanced clustering (EP and UP), size-balanced clustering, and conventional k-means. In the UP case, inclusion probabilities increase from left to right. Colored polygons delineate clusters; annotated values indicate cluster inclusion-probability totals.

deemed optimal. Existing indices, by contrast, can change under such translations despite all interpoint relationships being preserved. These properties make the index a more robust and interpretable assessment of sample spread.

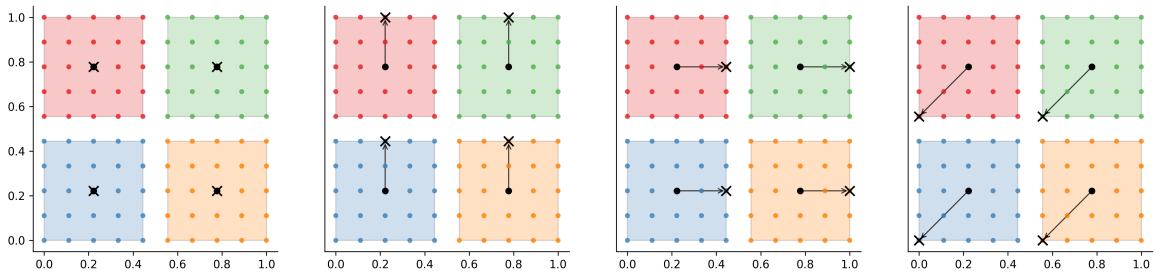

Figure 3: Spatial translation invariance of the index. In the first panel, the sample (black $\times$) coincides with the benchmark centroids O (black dots) within n-means UP-balanced clusters (colored parts) and is therefore judged optimally spread. The subsequent panels show uniform translations of this same arrangement; by translation invariance, all translated samples receive the same spread value and are likewise deemed optimal.

Since the central mechanism of this index for assessing the spread of a sample is the quantification of the disparity between the estimated density surfaces of the population before and after the translation procedure, we refer to it as the Density Disparity Index (DI). The details of each step are presented in the remainder of this section.

Given a sample $S \in \mathcal{S}$, we begin by clustering the population using n-Means UP-balanced clustering, taking the units of S as the initial centroids. This partitions U into n clusters, each with a total inclusion probability of one. The centroids of these clusters, denoted O, serve as the *benchmark* for evaluating the spread optimality of S, with the idea that the closer the arrangement of S is to that of O, the more it represents an optimal spread. The arrangement of O is determined jointly by the spatial structure of the population and the arrangement of S, making the benchmark both population-specific and sample-specific. This procedure allows O to adapt to each given sample. For instance, Figure 4 illustrates this with two samples for the same gridded population that have the same level of spread despite their distinct arrangements. The vertically arranged sample (left panel) induces horizontal clusters, while the horizontally arranged sample (right panel) induces vertical clusters. In each case, the sample units S are positioned near the centroids O of their respective clusters. If a single, fixed clustering were imposed—for example, the horizontal partitioning from the left panel—the second sample would be inaccurately assessed as poorly spread due to its distance from those fixed centroids. Therefore, this approach ensures that different arrangements are evaluated equitably, enabling a more appropriate comparison of their spread optimality.

Next, each unit of S is assigned to one of the centroids in O and its corresponding cluster. The arrangement of O can differ markedly from that of S, particularly when S exhibits poor spatial spread. To determine an optimal one-to-one correspondence between sample units and centroids based on spatial proximity, we use the Hungarian assignment algorithm (Kuhn, 1955). Following this assignment, the coordinates of the sample units are denoted by $\{\boldsymbol{s}_i : i \in S\}$, and the coordinates of their matched centroids are denoted by $\{\boldsymbol{o}_i : i \in S\}$.

With the centroid–sample assignment fixed, we introduce a clusterwise translation that

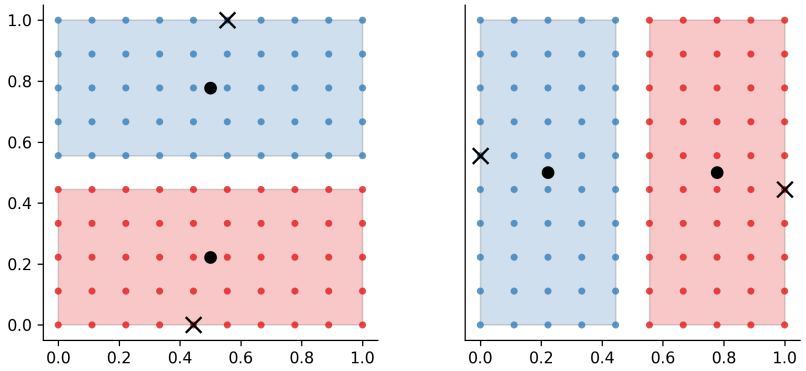

Figure 4: Sample-specific benchmark generation via n-Means UP-balanced clustering for two distinct samples of size $n = 2$ drawn from the same gridded population. The sample units of S (marked by $\times$) initialize the clustering. Left: A vertically arranged sample induces a horizontal partition, and the resulting benchmark centroids O (black dots) reflect this configuration. Right: A horizontally arranged sample induces a vertical partition, yielding a different set of benchmark centroids O.

recenters each population cluster at its matched sample unit in order to quantify the discrepancy between the sample arrangement and the centroid configuration. Let $U_1, \ldots, U_n$ denote the clusters with centroids $\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n$, and let $\boldsymbol{s}_i$ be the coordinate of the sample unit matched to $\boldsymbol{o}_i$. For any unit $\ell \in U_i$ with coordinate $\boldsymbol{c}_\ell$, define

$$T(\boldsymbol{c}_\ell) = \boldsymbol{c}_\ell + (\boldsymbol{s}_i - \boldsymbol{o}_i),$$

which uniformly shifts all units in U_i so that the cluster centroid moves from $\boldsymbol{o}_i$ to $\boldsymbol{s}_i$. Because T is a rigid translation within each cluster, within-cluster pairwise distances are preserved exactly; changes arise only from the relative placement of clusters. If the sample's spatial relationships among units closely resemble those of the centroid arrangement, then applying T to every unit has only a small effect on the relative placement of clusters. Conversely, any deviation between the sample arrangement and the centroid arrangement produces a non-trivial perturbation of the population's spatial structure.

To quantify the population's spatial structure before and after translation, we will estimate two spatial density functions via MKDE[2] (Wand and Jones, 1994) on the coordinates of all population units. For any location $\boldsymbol{c} \in \mathbb{R}^2$, the MKDE estimator is

$$f(\boldsymbol{c}) \;=\; \frac{1}{N} \sum_{i \in S} \sum_{\ell \in U_i} K_{\mathbf{H}}(\boldsymbol{c} - \boldsymbol{c}_\ell),$$

where

- $K_{\mathbf{H}}(\cdot)$ is the multivariate kernel function defined by

$$K_{\mathbf{H}}(\boldsymbol{c}) = |\mathbf{H}|^{-1/2}\, K(\mathbf{H}^{-1/2}\, \boldsymbol{c}),$$

  with $K(\cdot)$ a symmetric multivariate density.

[2]Multivariate Kernel Density Estimation.

- $\mathbf{H} \in \mathbb{R}^{2\times 2}$ is the symmetric positive-definite bandwidth matrix, governing the amount and orientation of smoothing and selected according to Scott's rule.

In this work we adopt the uniform kernel

$$K(\boldsymbol{c}) = \frac{1}{2^2}\,\mathbf{1}\{\|\boldsymbol{c}\|_\infty \le 1\},$$

so that

$$K_{\mathbf{H}}(\boldsymbol{c}) = |\mathbf{H}|^{-1/2}\,\frac{1}{2^2}\,\mathbf{1}\{\|\mathbf{H}^{-1/2}\boldsymbol{c}\|_\infty \le 1\}.$$

By applying this estimator to the original population we obtain the original spatial density surface, and by applying it to the population after cluster-wise translations $T(\cdot)$ we obtain the translated spatial density surface.

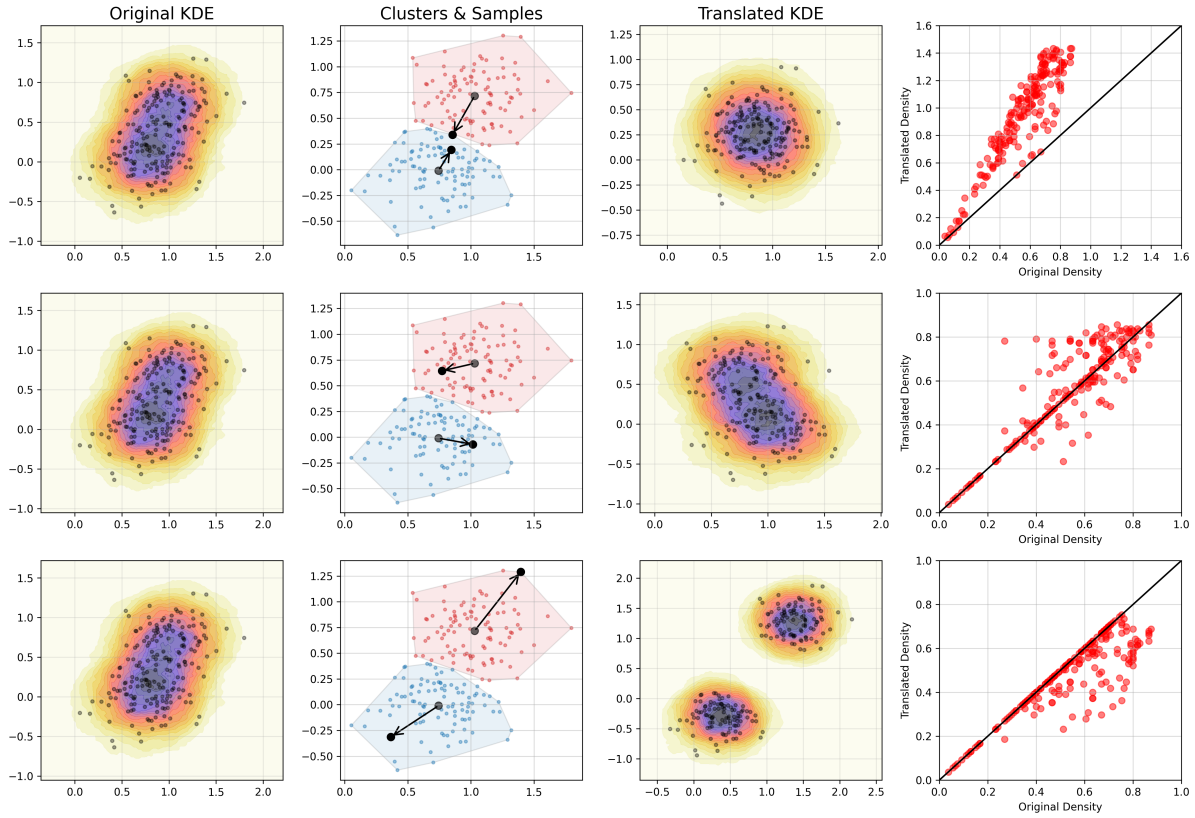

Figure 5: Illustration of cluster-wise translation and density agreement for three sample configurations in a clustered population. Columns (left to right): (i) original spatial density over the population; (ii) partition into 2-Means UP-balanced clustering (blue/red) with sample units (bold black dots), cluster centroids (pale black dots), and arrows indicating the translation from each centroid to its assigned sample; (iii) density of the translated population after applying the cluster-wise translations; (iv) density–density comparison plotting original (horizontal) versus translated (vertical) values computed with *MKDE*, where the 45° identity line denotes equality. Rows correspond to three alternative sample arrangements (near–near, cross, and far–far), illustrating how sample arrangement influences distortion in the translated density.

Figure 5 illustrates how cluster-wise translation affects agreement between the original and translated spatial densities. In each row, panels show (i) the original density, (ii) the n-Means UP-balanced clustering with sample units, centroids, and translation arrows, (iii) the translated

density after applying the cluster-specific shifts, and (iv) a density–density scatter plot of translated (vertical) versus original (horizontal) values with the 45° identity line. The top row depicts a configuration where the sample units are too close to each other; translation causes the two clusters to overlap and locally inflate density, yielding scatter points predominantly above the identity line. The middle row shows sample units that are reasonably close to their respective centroids, but the translation arrows point in opposite directions; this introduces mild variation, so some points lie above and some below the identity line, while the majority remain near it. The bottom row presents sample units that are far apart: translation separates the clusters and locally deflates density, producing scatter points that fall below the identity line.

In populations with many clusters, translation-induced effects become more intricate; some clusters converge while others diverge, placing points above or below the 45° identity line in the density–density scatter plot (see Figure 6). The translated spatial structure thus emerges from inter-cluster interactions, and the distribution of points relative to the identity line succinctly conveys the global response—predominantly above indicates a net increase in local density, predominantly below indicates a net decrease, and close alignment indicates minimal structural change. Leveraging scuh plot as a diagnostic, we summarize the overall departure from optimal spread by the net deviation of points from the identity line: a net deviation of zero corresponds to a perfectly spread sample—or to an arrangement in which local contractions and expansions (some units closer, others farther) offset to yield a balanced outcome; its magnitude quantifies the loss of spread quality; and its sign reveals the mode of failure—positive when units are on average too close (inflated translated density) and negative when they are too far apart (deflated translated density). This signed characterization captures both the extent and the nature of deviation from optimal spread, offering information unavailable from conventional indexes.

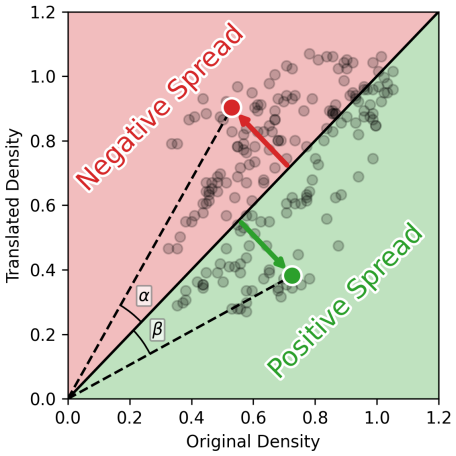

Figure 6: Angle-based contributions in the density–density scatter plot. Points show translated versus original densities with the identity line (solid) separating the *negative spread* region (above; units too close) from the *positive spread* region (below; units too far). For any point, the ray from the origin forms an angle with the identity line (illustrated as α for the red point above and β for the green point below). The signed contribution to the index is proportional to the scaled value of $-\sin$ of this angle (negative above the line, positive below), while $1-\cos$ of the angle quantifies dissimilarity from the identity line used in the dispersion coefficient.

To formalize the notion of *Net Deviation*, we examine each point of the density–density

scatter plot at the level of individual population units. For a unit $\ell \in U$ with location $\boldsymbol{c}_\ell$, define

$$\mathbf{f}_\ell \;=\; \begin{pmatrix} f(\boldsymbol{c}_\ell) \\ f(T(\boldsymbol{c}_\ell)) \end{pmatrix}, \qquad \mathbf{1} \;=\; \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Let α_ℓ be the oriented angle from $\mathbf{1}$ to $\mathbf{f}_\ell$. A direct calculation yields

$$\sin\alpha_\ell = \frac{\|\mathbf{1}\times\mathbf{f}_\ell\|}{\|\mathbf{1}\|\,\|\mathbf{f}_\ell\|} = \frac{f(T(\boldsymbol{c}_\ell)) - f(\boldsymbol{c}_\ell)}{\sqrt{2}\,\sqrt{f(T(\boldsymbol{c}_\ell))^2 + f(\boldsymbol{c}_\ell)^2}},$$

$$\cos\alpha_\ell = \frac{\mathbf{1}\cdot\mathbf{f}_\ell}{\|\mathbf{1}\|\,\|\mathbf{f}_\ell\|} = \frac{f(T(\boldsymbol{c}_\ell)) + f(\boldsymbol{c}_\ell)}{\sqrt{2}\,\sqrt{f(T(\boldsymbol{c}_\ell))^2 + f(\boldsymbol{c}_\ell)^2}}.$$

Here, $\sin\alpha_\ell$ is a signed angular deviation of the point $\mathbf{f}_\ell$ from the identity line. To align signs with our diagnostic convention (negative for local inflation after translation and positive for local deflation), we use $-\sin\alpha_\ell$, hence values < 0 indicate local inflation (points above the identity line), values > 0 indicate local deflation (points below the line), and 0 indicates local agreement. In order to quantify the angular dispersion (dissimilarity) of points from the identity line, we will use $1-\cos\alpha_\ell$; the factor $1-\cos\alpha_\ell$ will serve as a natural weight when aggregating pointwise deviations into a global signed measure.

Since α_ℓ typically takes small values, we apply the following scaling to increase sensitivity and capture finer differences:

$$\mathrm{scale}(\beta;\gamma) \;=\; \begin{cases} 1, & \frac{\beta}{\gamma} \geq 1, \\ -1, & \frac{\beta}{\gamma} \leq -1, \\ \frac{\beta}{\gamma}, & \text{otherwise}, \end{cases}$$

where $\gamma > 0$. Under this transformation, values with $|\beta| \geq \gamma$ are saturated to ± 1, while those with $|\beta| < \gamma$ are mapped linearly from $(-\gamma, \gamma)$ to $(-1, 1)$.

Figure 6 provides an angle-based visualization of these definitions. Each point corresponds to a population unit with coordinates $(f(\boldsymbol{c}_\ell),\, f(T(\boldsymbol{c}_\ell)))$; the solid 45° line is the identity. Points above the line indicate *negative spread* (local inflation after translation), while points below indicate *positive spread* (local deflation). For any point, the ray from the origin forms an oriented angle α_ℓ with the identity line. The signed contribution entering the net component is given by the scaled value of $-\sin\alpha_\ell$ (negative above, positive below), and the dispersion contribution is measured by $1-\cos\alpha_\ell$, which quantifies angular dissimilarity from the identity line.

The net deviation D_{net} is defined as the average of the scaled negative sin over all population units:

$$D_{\mathrm{net}} \;=\; \frac{1}{N}\sum_{\ell\in U} \mathrm{scale}\big(-\sin\alpha_\ell;\, \sin(\tfrac{\pi}{8})\big)\,.$$

Equivalently, D_{net} decomposes into the contributions of positive and negative deviations,

$$D_+ \;=\; \sum_{\ell:\,-\sin\alpha_\ell - \geq 0} \mathrm{scale}\big(-\sin\alpha_\ell;\, \tfrac{\pi}{8}\big)\,, \qquad D_- \;=\; \sum_{\ell:\,-\sin\alpha_\ell < 0} \mathrm{scale}\big(-\sin\alpha_\ell;\, \tfrac{\pi}{8}\big)\,,$$

$$D_{\text{net}} \;=\; \frac{D_+ + D_-}{N},$$

where D_+ captures *positive spread*—the component arising when sample units are, on average, too far apart (points below the identity line)—and D_- captures *negative spread*—the component arising when sample units are, on average, too close (points above the line).

Thus $D_{\text{net}} \approx 0$ in two situations: (i) points lie predominantly on the identity line, indicating near-equality of original and translated densities (sample units effectively coincide with the UP-Balanced centroids, i.e., optimal spread); or (ii) nonzero positive and negative deviations are present but offset each other, reflecting an arrangement with some units closer and others farther such that their effects balance.

The second case has a useful interpretation: it identifies a form of well-spread arrangement arising from *balanced imperfections* (some units too close, others too far). However, to avoid awarding near-perfect values when $|D_+|$ and $|D_-|$ are both large but offset, we introduce a *Dispersion Coefficient* that measures the overall angular dissimilarity from the identity line, irrespective of sign, as

$$\eta \;=\; \frac{1}{N} \sum_{\ell \in U} \text{scale}(1 - \cos \alpha_\ell;\, 1 - \cos(\tfrac{\pi}{8}))\,,$$

which is small when the translated and original densities closely align pointwise and increases as the scatter departs from the identity line.

For a given sample $S \in \mathcal{S}$ and population U, combining the net deviation with the Dispersion Coefficient while preserving the $[-1, 1]$ range yields the DI:

$$\text{DI}(S) \;=\; D_{\text{net}} \;+\; (\text{sign}(D_{\text{net}}) - D_{\text{net}})\, \eta,$$

which interpolates between D_{net} (when $\eta = 0$) and $\text{sign}(D_{\text{net}})$ (when $\eta = 1$).

Algorithm 2 summarizes the full computation of the DI: starting from a n-Means UP-balanced clustering partition initialized at the sample, it assigns centroids via the Hungarian algorithm, applies cluster-wise translations, estimates original and translated densities, computes pointwise angles, and aggregates scaled $-\sin \alpha_\ell$ and $1 - \cos \alpha_\ell$ into the signed component D_{net} and the dispersion coefficient η, respectively, to yield $\text{DI} \in [-1, 1]$.

Example

To clarify the limitations of existing spatial balance indices and the advantages of the proposed DI, we consider two simplified population settings shown in Figure 7. In panel (a), the population forms three well-separated clusters placed at large mutual distances. In panel (b), the population forms four compact clusters arranged symmetrically in close proximity.

The numerical results in Table 1 detail the limitations of classical indices and the strength of DI.

For the three-cluster setting, panel (a), both the VI and the MI fail to discriminate among samples: all are evaluated identically as "perfect" (VI $= 0$ and MI $= -1$), despite clear structural differences. The BI is more informative but is unsigned; it cannot indicate whether a sample is overly concentrated or overly dispersed and therefore does not distinguish extreme cases such as

Algorithm 2 Density Disparity Index (DI)

Require: Population $U = \{1, \dots, N\}$ with coordinates $\{\boldsymbol{c}_\ell \in \mathbb{R}^2\}_{\ell \in U}$; inclusion probabilities $\boldsymbol{\pi} = (\pi_\ell)_{\ell \in U}$ with $\sum_\ell \pi_\ell = n$; sample $S = \{i_1, \dots, i_n\} \subset U$; kernel density estimator MKDE$(\cdot)$; scaling function scale$(\beta; \gamma)$.

Ensure: Density Disparity Index DI $\in [-1, 1]$.

Clustering and benchmark centroids

Run n-**Means UP-balanced clustering** on U with initialization by the sample units S.

Let $\{U_i\}_{i=1}^n$ be the clusters with $\sum_{\ell \in U_i} \pi_\ell = 1$.

Denote the set of centroids by $O = \{\boldsymbol{o}_1, \dots, \boldsymbol{o}_n\}$ (benchmark).

Assignment and translation

Compute a one-to-one assignment $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ between sample units $\{\boldsymbol{s}_i\}$ and centroids $\{\boldsymbol{o}_i\}$ by the Hungarian algorithm.

Define the cluster-wise translation $T(\boldsymbol{c}_\ell) = \boldsymbol{c}_\ell + \boldsymbol{s}_i - \boldsymbol{o}_i$ for all $\ell \in U_i$.

Density estimation

Estimate the spatial density on original and translated locations:

$$f(\boldsymbol{c}_\ell) \leftarrow \text{MKDE}(\boldsymbol{c}_\ell; \{\boldsymbol{c}_m\}_{m \in U}), \qquad f(T(\boldsymbol{c}_\ell)) \leftarrow \text{MKDE}(T(\boldsymbol{c}_\ell); \{T(\boldsymbol{c}_m)\}_{m \in U}).$$

Pointwise angles

For each $\ell \in U$, compute

$$\sin \alpha_\ell = \frac{f(T(\boldsymbol{c}_\ell)) - f(\boldsymbol{c}_\ell)}{\sqrt{2}\sqrt{f(T(\boldsymbol{c}_\ell))^2 + f(\boldsymbol{c}_\ell)^2}}, \qquad \cos \alpha_\ell = \frac{f(T(\boldsymbol{c}_\ell)) + f(\boldsymbol{c}_\ell)}{\sqrt{2}\sqrt{f(T(\boldsymbol{c}_\ell))^2 + f(\boldsymbol{c}_\ell)^2}}.$$

Aggregation

Net deviation: $D_{\text{net}} \leftarrow \dfrac{1}{N} \displaystyle\sum_{\ell \in U} \text{scale}(-\sin \alpha_\ell;\ \tfrac{\pi}{8})$.

Dispersion Coefficient: $\eta \leftarrow \dfrac{1}{N} \displaystyle\sum_{\ell \in U} \text{scale}(1 - \cos \alpha_\ell;\ 1 - \cos(\tfrac{\pi}{8}))$.

Index

Density Disparity Index:

$$\text{DI} \leftarrow D_{\text{net}} + \big(\text{sign}(D_{\text{net}}) - D_{\text{net}}\big)\, \eta.$$

return DI.

`aaa`, `aac`, and `ccc`. By contrast, DI produces a meaningful ordering: `aaa` is least well spread (-0.84), `aac` is intermediate (-0.37), and `ccc` is most spread (0.74). The range (from -0.84 to 0.74) shows that DI captures gradations of spread rather than collapsing configurations into a single class as VI and MI do here.

For the four-cluster setting, panel (b), the first three samples (`e1 e2 e3 e4`, `a1 a2 a3 a4`, `d1 d2 d3 d4`) are simple translations of the centroid configuration. DI correctly treats all as well spread (0.00), while other indices are ambiguous. The fourth sample (`f1 h2 d3 b4`) is essentially a rotation, which DI rates near balance (0.11). The fifth and sixth samples (`b1 f2 h3 d4` and `h1 d2 b3 f4`) move one unit far from, or very close to, another cluster; DI reflects this asymmetry in both sign and magnitude (0.78 versus -0.71). The seventh sample (`a1 d1 b1 e1`) exhibits extreme proximity, yielding the lowest DI (-0.86) and correctly flagging the worst imbalance. Again, BI and VI fail to discriminate many of these configurations. MI detects clustering in the extreme case (`a1 d1 b1 e1`, MI $= 0.61$) but is less informative in intermediate cases: `b1 f2 h3 d4` (overspread) and `h1 d2 b3 f4` (overconcentrated) both receive middling values (about -0.50 to -0.69), indicating that MI lacks the directional sensitivity that DI provides.

Taken together, these examples show that DI not only separates dense from spread samples but also provides a signed, directionally sensitive measure indicating whether a sample is too compact or too dispersed. By comparison, BI is unsigned, VI is insensitive to global configuration, and MI can be misleading in clustered layouts. DI therefore offers a richer and more interpretable diagnostic of spatial spread, responsive to both global separation and local concentration.

Table 1: Index values for example samples in Figure 7. Columns report Density Disparity Index (DI), Balanced-Voronoi (BI), Voronoi (VI), and Moran (MI). Left block corresponds to subplot (a) with three clusters; right block to subplot (b) with four clusters.

(a) 3 clusters							**(b) 4 clusters**							
Sample			DI	BI	VI	MI	Sample				DI	BI	VI	MI
a	a	a	-0.84	0.26	0.00	-1.00	e1	e2	e3	e4	0.00	0.00	0.00	-1.00
a	a	b	-0.69	0.21	0.00	-1.00	a1	a2	a3	a4	0.00	0.43	0.11	-0.59
a	a	c	-0.37	0.26	0.00	-1.00	d1	d2	d3	d4	0.00	0.28	0.02	-0.83
a	b	b	-0.41	0.15	0.00	-1.00	f1	h2	d3	b4	0.11	0.19	0.00	-0.75
a	b	c	-0.12	0.21	0.00	-1.00	b1	f2	h3	d4	0.78	0.26	0.00	-0.50
a	c	c	0.21	0.26	0.00	-1.00	h1	d2	b3	f4	-0.71	0.26	0.00	-0.69
b	b	b	0.00	0.00	0.00	-1.00	a1	d1	b1	e1	-0.86	1.01	0.73	0.61
b	b	c	0.30	0.15	0.00	-1.00	a1	c1	g1	i1	-0.70	0.85	0.58	0.38
b	c	c	0.56	0.21	0.00	-1.00	c1	c2	g3	g4	0.56	0.36	0.00	-0.40
c	c	c	0.74	0.26	0.00	-1.00	i1	a2	a3	i4	-0.51	0.36	0.00	-0.05

6 Graphical Sampling for Spatial Populations

Panahbehagh (2025) introduced a graphical framework for finite-population sampling (GFS) that affords intuitive control over sampling designs. The original development treated the one-dimensional case, where units are arranged along a single axis without explicit spatial

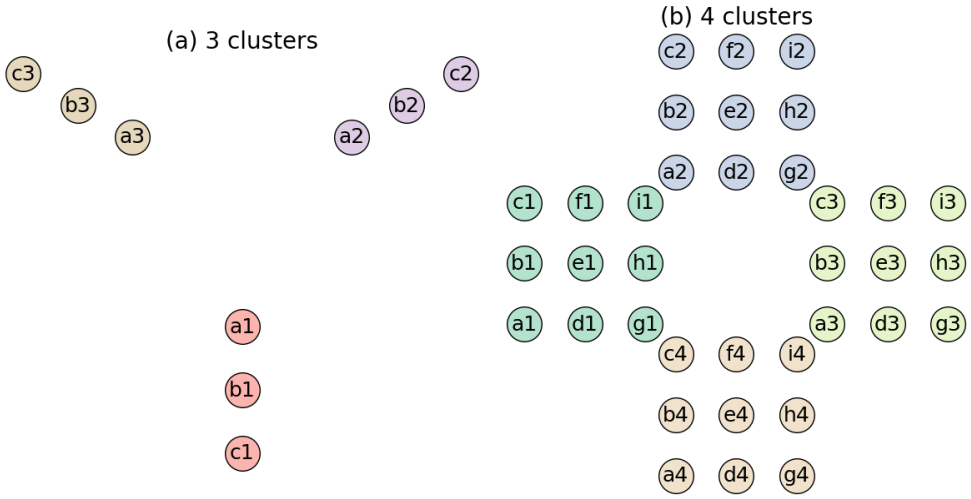

Figure 7: Illustrative population structures used to compare classical spread indices with the proposed Density Disparity Index (DI). Panel (a): three rotated clusters placed far apart. Panel (b): four uniformly arranged clusters in close proximity. These contrasting cases highlight situations where MI, the VI, or the BI may provide indistinguishable or misleading results, whereas DI adapts to the underlying population geometry and more faithfully reflects differences in spatial spread.

coordinates, yet it accommodates a broad class of unequal-probability designs ranging from Madow's systematic sampling (Madow, 1949) to maximum-entropy sampling (Tillé, 2006). In this paper, we extend the algorithm to populations with D-dimensional coordinates, with particular emphasis on the spatial case $D = 2$. Throughout, we assume a fixed-size design with $\sum_{\ell=1}^{N} \pi_\ell = n \in \mathbb{N}$. The resulting methodology is detailed in Algorithm 3.

Algorithm 3 Fixed-Size Spatial GFS for a Population with D-Dimensional Coordinates

1: Construct a $(D{+}1)$-dimensional system: represent the spatial coordinates of each population unit on the first D axes, and use the $(D{+}1)$-th axis to display its inclusion probability in $(0, 1]$.
2: Order the population units (arbitrarily or at random) to obtain a sequence $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_N$, where each $\boldsymbol{c}_\ell \in \mathbb{R}^D$ denotes a point in the coordinate space.
3: Initialize the first bar at $\boldsymbol{c}_1$ to cover $[0, \pi_1]$ on the $(d{+}1)$-th axis; set $b_1 \leftarrow \pi_1$.
4: **for** $\ell = 2$ to N **do**
5: **if** $b_{\ell-1} + \pi_\ell \leq 1$ **then**
6: Place the ℓ-th bar at $\boldsymbol{c}_\ell$ covering $(b_{\ell-1}, , b_{\ell-1}{+}\pi_\ell]$; set $b_\ell \leftarrow b_{\ell-1}{+}\pi_\ell$.
7: **else**
8: Place the ℓ-th bar at $\boldsymbol{c}_\ell$ covering the wrapped interval $(b_{\ell-1}, 1] \cup [0, , b_{\ell-1}{+}\pi_\ell{-}1]$;
9: set $b_\ell \leftarrow b_{\ell-1}{+}\pi_\ell{-}1$.
10: **end if**
11: **end for**
12: Draw $r \sim U(0, 1)$. The final sample consists of all units whose assigned bar contains r along the $(d{+}1)$-th axis.

As two illustrative examples for $D = 1$ and $D = 2$, consider Figure 8 with the following inclusion probabilities:

$$\pi_1 = 0.7, \pi_2 = 0.3, \pi_3 = 0.4, \pi_4 = 0.8, \pi_5 = 0.3, ; \pi_6 = 0.65, \pi_7 = 0.45, \pi_8 = 0.2, \pi_9 = 0.2. \tag{5}$$

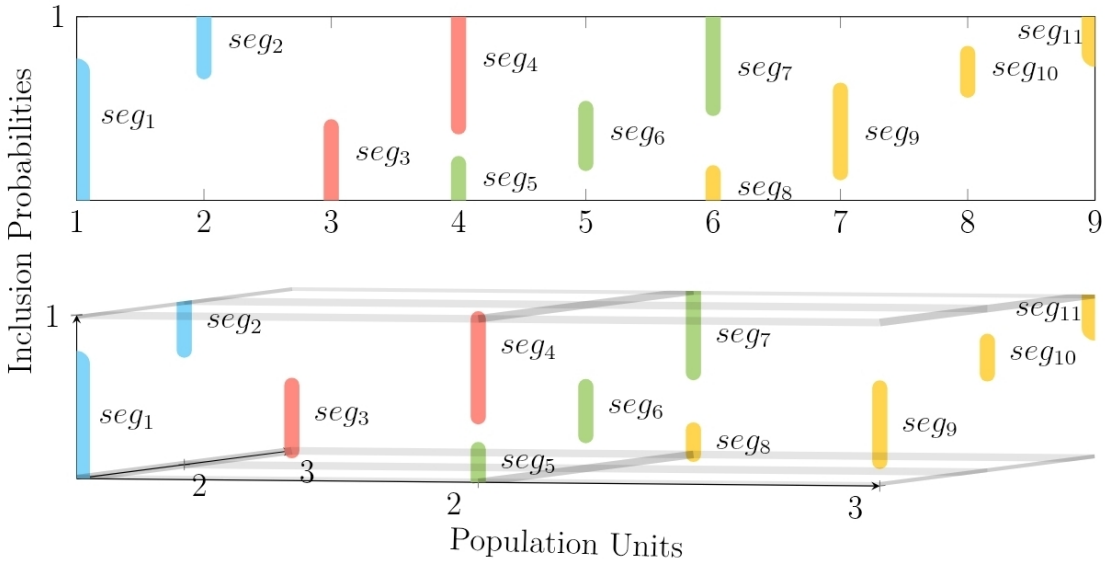

Figure 8: Visualization of the GFS design for one-dimensional (top) and two-dimensional (bottom) populations, as described in Example 5. Vertical bars, potentially segmented by $seg_1, seg_2, \ldots, seg_{11}$, represent each unit's inclusion probability (UP) in lexicographic order. Colors indicate groups of units whose total UP equals 1, ensuring that a single uniform draw selects exactly one unit from each color group, thereby producing a fixed-size design.

For the one-dimensional case, the population coordinate set is

$$\boldsymbol{C} = \{\boldsymbol{c}_1 = 1,\ \boldsymbol{c}_2 = 2,\ \boldsymbol{c}_3 = 3,\ \boldsymbol{c}_4 = 4,\ \ldots,\ \boldsymbol{c}_9 = 9\},$$

$$\boldsymbol{C} = \{\boldsymbol{c}_1 = (1,1),\ \boldsymbol{c}_2 = (1,2),\ \boldsymbol{c}_3 = (1,3),\ \boldsymbol{c}_4 = (2,1),\ \ldots,\ \boldsymbol{c}_9 = (3,3)\}.$$

In constructing the bars, we group those that collectively span the probability axis from 0 to 1 into a single stack. Within each stack, every bar requires at most two segments (indicated by *seg* in Figure 8), splitting only when the cumulative probability reaches 1 and wraps to 0. By contrast, Panahbehagh (2025) proposed an optional refinement that increases design entropy by subdividing bars into smaller pieces and rearranging them while preserving fixed sample size. Although this refinement could be incorporated here to further diversify the design, we do not pursue it in the present exposition.

The sampling step begins by drawing a random horizontal line across the ordered bars—this line represents a uniform random threshold within $[0, 1]$. Units whose bars intersect this line are selected into the sample. Conceptually, units with bars aligned at similar heights have a synchronized chance of being selected together. This geometric mechanism naturally preserves first-order inclusion probabilities while introducing spatial coordination through the bar arrangement.

Arranging the bars to implement a sampling design affords precise control over the selection mechanism. By strategically choosing their placement and order, one can directly modulate the spatial spread and distribution of selected units, tailoring the design to specific objectives. This enables a form of "inverse engineering," wherein desired sample properties guide the bar configu-

ration: prescribing the intended allocation of selections across the population yields arrangements that achieve these targets while preserving the inherent randomness of the procedure.

The framework is highly flexible, supporting designs that meet statistical requirements and practical constraints simultaneously—for example, enforcing uniform spatial coverage or adjusting selection probabilities for designated units. Through careful configuration, it is possible to balance methodological rigor with operational needs, maintaining the robustness of probabilistic selection throughout.

7 n-Means Spatial Sampling

In this section, we introduce *n-Means Spatial Sampling*, a design for drawing well-spread samples from a finite population with prescribed inclusion probabilities. The design uses nested clustering to capture both global and local spatial structure. Its core rule is twofold; units from the same cluster are not selected together, while units occupying the same relative position across different clusters are selected jointly. Leveraging the GFS framework, we translate this rule into a sampling design by constructing a total order of the population units induced by the nested clustering.

The procedure first uses the n-Means UP-balanced clustering to shape spatially compact clusters $\{U_1, \ldots, U_n\}$ such that $\sum_{\ell \in U_i} \pi_\ell = 1$ for every $i \in \{1, \ldots, n\}$. Within each cluster U_i, units are further partitioned into m zones $\{U_{i,1}, \ldots, U_{i,m}\}$ satisfying $\sum_{\ell \in U_{i,j}} \pi_\ell = 1/m$ for each $j \in \{1, \ldots, m\}$, obtained by a second application of n-Means UP-balanced clustering targeted to the within-cluster total $1/m$. The n-Means UP-balanced clustering provide a global spatial structure for the population; units within each cluster U_i are geographically similar, so co-selecting them would degrade spread and is therefore avoided. However, cluster-level control alone may be insufficient, because units from two neighboring clusters can lie close to a shared boundary; selecting such pairs together would again reduce spread. To regulate this local geometry, we rely on the within-cluster zones. For examples of within-cluster zones, see the shaded backgrounds labeled 1–4 in the first two rows of Figure 2. Within the GFS framework, the guiding principle is to order the UP bars so that zones occupying the same relative position across clusters—thus inducing a well-spread zoning—begin at the same height and are processed in parallel during selection. Under this alignment, selecting units from equally labeled zones naturally disperses the sample across the population, while boundary-near units from adjacent clusters typically fall into different zones and are not co-selected. This yields samples that are well spread at both global and local levels.

To identify which zones occupy the same relative positions across clusters, we assign an order to the zones within each cluster via a ranking map $\psi_1 : \{1, \ldots, m\} \to \{1, \ldots, m\}$. The same ordering rule must be applied in every cluster to make ranks comparable. Figure 9 illustrates several such rules, including *Horizontally Lexicographic*, *Vertically Lexicographic*, *Random*, *Radial Distance from Origin*, *Projection along Diagonal*, *Radial Distance from Centroid*, *Centroidal Polar*, *Max-Coordinate*, and *Hilbert Space-Filling Curve*. After applying the common ranking ψ_1 in every cluster, zones receiving the same rank j are taken to define the same relative positioning zones across clusters. We list the zones of cluster i in this order as $U_{i,(1)}, U_{i,(2)}, \ldots, U_{i,(m)}$.

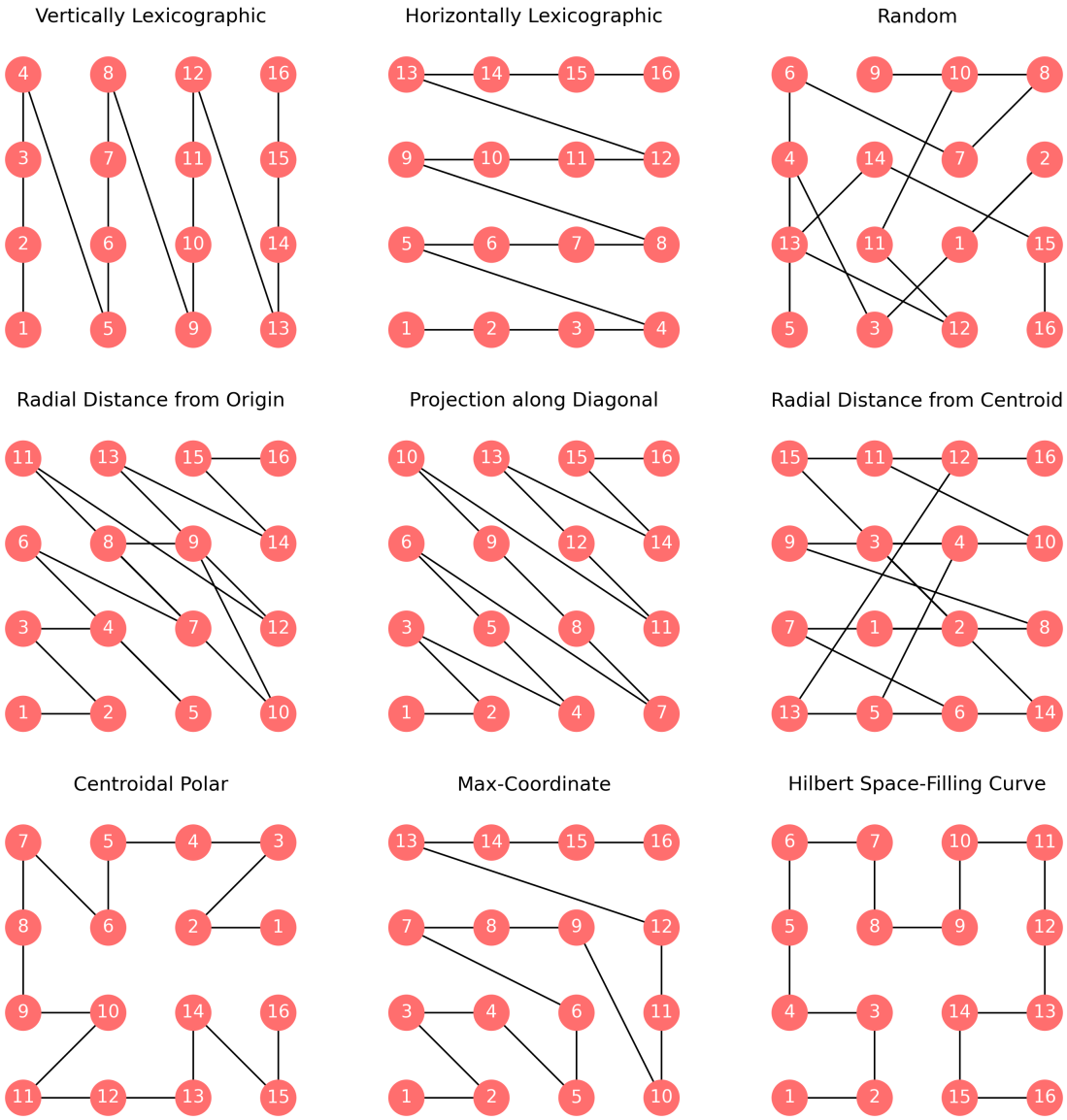

Figure 9: Illustrative ordering rules applied to spatial representatives—(i) zone centroids within a cluster or (ii) unit coordinates within a zone. In each panel, points are connected in ascending rank, visualizing the induced total order.

An analogous ordering is imposed on units within each zone. For zone $U_{i,(j)}$ and coordinates $\mathbf{c}_\ell \in \mathbb{R}^2$ for $\ell \in U_{i,(j)}$, define a ranking map $\psi_2 : U_{i,(j)} \to \{1, \ldots, |U_{i,(j)}|\}$ by ordering the units according to a common within-zone rule (e.g., any of the ordering schemes in Figure 9) and letting $\psi_2(\ell)$ be the position of unit ℓ in that order. Using the same rule in every zone ensures comparable ranks. With this convention, units that share the same zone (by ψ_1) and the same within-zone rank (by ψ_2) are treated as having the same relative positioning and are scheduled to co-occur in samples, while units with different ranks are separated, enhancing local spread.

We now assemble the total order by traversing clusters along the fixed path supplied by n-Means UP-balanced clustering, and, within each cluster, concatenating zones and within-zone ranks with border adjustments. Denote the clusters in path order by $U_{(1)}, \ldots, U_{(n)}$. For each consecutive pair on this path, there is at most one border unit whose inclusion probability is split across the two clusters. In cluster $U_{(i)}$, we list (1) the border unit with $U_{(i-1)}$ first, if it exists; (2) the units of zones $U_{(i),(1)}, \ldots, U_{(i),(m)}$ in increasing within-zone rank; and (3) the border unit with $U_{(i+1)}$ last, if it exists. If the same unit appears in consecutive positions in this construction (as may occur for border units), we collapse these into a single occurrence, ensuring that each population unit appears exactly once in the final order. Therefore, concatenating these within-cluster sequences in path order yields a unique total order Ψ, which encodes all the intended global and local structure in a way such that the GFS framework could interpret it correctly.

The total order Ψ is tailored to the GFS construction. Because each cluster has total probability 1, every cluster forms exactly one complete stack, so precisely one unit from each cluster is selected in any sample—ensuring that units within the same cluster do not co-occur. Border units behave consistently with this scheme; a border unit is the one that splits across adjacent stacks, with its remainder initiating the next stack. Zones likewise align naturally. Since each zone carries probability $1/m$ and zones of a cluster are placed contiguously in Ψ, the cluster's stack on $[0, 1]$ is partitioned into m consecutive subintervals, each assigned to one zone. Consequently, zones with the same rank across clusters occupy the same subinterval of $[0, 1]$; all samples arising from a given subinterval therefore select units from the same relative positioning zones in every cluster, achieving the intended cross-cluster co-occurrence pattern.

Providing the total order Ψ to the GFS framework completes the procedure and yields the n-Means Spatial Sampling design with the desired properties. Importantly, n-Means Spatial Sampling defines not a single design but a *family* parameterized by the number of zones m per cluster and by the ranking functionals used to order zones and units. This structure offers flexibility within a systematic plan; fixing n clusters guarantees a baseline level of spread, while alternative within-cluster orderings induce distinct designs. At one extreme, choosing $m = 1$ (one zone per cluster) and a random within-zone ordering produces a high-entropy design within this family. At the other, adopting a larger m together with strict, geometry-aware ranking rules reduces entropy and strengthens spatial spread. This tunable trade-off between entropy and spread is valuable in practice, allowing the analyst to calibrate the design to application-specific priorities and constraints while preserving the prescribed first-order inclusion probabilities.

Algorithm 4 summarizes the construction of n-Means Spatial Sampling. It inputs spatial coordinates and inclusion probabilities, builds n-Means UP-balanced clustering and within-cluster

zones, derives a total order Ψ consistent with the GFS framework, and finally draws a fixed-size, well-spread sample of size n.

7.1 Intelligent n-Means Spatial Sampling

The n-Means Spatial Sampling leverages GFS to represent the population with n-Means UP-balanced clusters, zones, and simple ordering rules. This representation yields a *family of valid designs*: by choosing the number of zones and the ordering of zones and units, we obtain many feasible options that all respect inclusion probabilities and sample size. The resulting flexibility makes it natural to *search* for a design that best serves our goal—spreading the sample over the population. In this section, we make that search *intelligent*: we use a greedy best-first procedure that proposes small, structure-preserving edits to a seed design, scores each candidate with a spreadness index, keeps the most promising versions, and explores nearby alternatives. Because edits act only on orderings (and, when allowed, the zone count), feasibility is preserved, evaluation is fast, and the approach adapts to the specific population at hand.

The algorithm consists of the following key stages:

- **Seeding:** Construct one or more plausible starting designs and score each with a chosen spreadness index. In the GFS framework, a design is fully specified by (i) the partition of the population into n-Means UP-balanced clusters, (ii) the number of zones per cluster, and (iii) the ordering rules applied to zones and to units within zones. Because these elements are native to GFS, every seed automatically satisfies inclusion probabilities and sample size constraints. All seeds are inserted into a max-priority queue keyed by their spread score.

- **Selection:** Extract the top-ranked design from the queue to serve as the current parent. This concentrates computation on the most promising region of the design space while avoiding redundant expansion of the same candidate.

- **Local variation:** Generate a small batch of child designs by applying limited, structure-preserving edits to the parent. Leveraging GFS's graphical representation (UP bars, zones, and ordering primitives), edits are simple and safe: re-rank zones within selected clusters (e.g., align or permute equally labeled zones across clusters), adjust within-zone unit orders using predefined rules (e.g., radial, sweep, serpentine, space-filling). These operators modify only the ordering logic, so feasibility and design-based validity are maintained.

- **Evaluation:** Score each child with the same spreadness index (e.g., DI, MI, VI, BI) and insert all children into the priority queue so that higher-scoring candidates rise to the top. The GFS data structures allow re-use of cluster/zone information, keeping evaluation costs low.

- **Best-tracking:** If a child exceeds the best score observed so far, update the incumbent design and record the sequence of edits that produced it. This provenance makes improvements transparent and reproducible.

- **Iteration or stop:** Repeat selection–variation–evaluation until an iteration budget is reached or a target spread level is achieved, then return the best design encountered. Because only local ordering changes are applied within GFS, the method is *anytime*: it yields a valid, progressively improved design at every iteration and remains compatible with design-based inference throughout.

Further implementation details appear in Algorithm 5.

In the simulation study, we show that this approach reliably upgrades reasonable seeds to designs with substantially higher spread, both globally and within clusters.

Algorithm 4 n-Means Spatial Sampling

Require: Population $U = \{1, \dots, N\}$ with coordinates $\{\boldsymbol{c}_\ell \in \mathbb{R}^2\}_{\ell \in U}$; inclusion probabilities $\boldsymbol{\pi} = (\pi_\ell)_{\ell \in U}$ with $\sum_\ell \pi_\ell = n$.

Require: Number of zones per cluster $m \geq 1$; zone-ranking function ψ_1; within-zone ranking function ψ_2.

Ensure: A fixed-size, well-spread sample $S \in \mathcal{S}$ with $|S| = n$.

UP-Balanced clustering and path

Run n-**Means UP-balanced clustering** to obtain clusters $\{U_i\}_{i=1}^n$ with $\sum_{\ell \in U_i} \pi_\ell = 1$ and a fixed path order $U_{(1)}, \dots, U_{(n)}$.

For each consecutive pair $(U_{(i)}, U_{(i+1)})$, identify at most one border unit $b_{(i)}$ whose probability is split across the two clusters (if present).

Within-cluster zoning and rankings

For each i, run n-**Means UP-balanced clustering** on U_i *targeting* m *zones* with per-zone total $1/m$, yielding $\{U_{(i),1}, \dots, U_{(i),m}\}$.

Using ψ_1 ranks zones as $U_{(i),(1)}, \dots, U_{(i),(m)}$.

For each zone, using ψ_2 to rank $\ell \in U_{(i),(j)}$.

Total order compatible with GFS

for $i = 1$ to n **do**
    Initialize Stack$_{(i)} \leftarrow [\,]$.
    if border with $U_{(i-1)}$ exists **then** prepend b_{i-1} to Stack$_{(i)}$.
        for $j = 1$ to m **do**
            Append units of $U_{(i),(j)}$ to Stack$_{(i)}$ in increasing ψ_2.
        end for
        if border with $U_{(i+1)}$ exists **then** append b_i to the end of Stack$_{(i)}$.
        end if
    end if
end for

Concatenate Stack$_{(1)}, \dots,$ Stack$_{(n)}$ to form a preliminary sequence; if the same unit appears in consecutive positions (due to border handling), keep a single occurrence. Denote the resulting total order by $\Psi = (\Psi_1, \dots, \Psi_N)$.

GFS sampling from the order

Build n stacks on $[0, 1]$ by sequentially placing bars of lengths $\pi_{\Psi_1}, \dots, \pi_{\Psi_N}$; when a bar would cross 1, split it so the leading piece closes the current stack and place the remainder at 0 to start the next stack.

Draw a single $r \sim U(0, 1)$; from each stack, select the unique unit whose bar intersects the vertical line at abscissa r.

return The n selected units S.

Algorithm 5 An Intelligent n-Means Spatial Sampling

Require: Population $U = \{1, \ldots, N\}$ with coordinates $\{\boldsymbol{c}_\ell\}_{\ell \in U}$; inclusion probabilities $\boldsymbol{\pi} = (\pi_\ell)_{\ell \in U}$.

Require: Set of initial designs $\mathcal{D}esign_0 = \{Design^{(1)}, \ldots, Design^{(q)}\}$ (each with fixed number of zones m and predefined zone and within-zone ordering rules).

Require: spreadness index $\textsc{Spread}(\cdot)$ for scoring designs; iteration limit $L \in \mathbb{N}$; optional target spread $\tau \in \mathbb{R}$.

Ensure: A design $Design^\star$ with high spread under $\textsc{Spread}(\cdot)$.

1: Initialize a max-priority queue Q keyed by $\textsc{Spread}(Design)$.
2: **for** each $Design \in \mathcal{D}esign_0$ **do**
3: Insert $Design$ into Q with key $\textsc{Spread}(Design)$.
4: **end for**
5: $Design^\star \leftarrow \arg\max_{Design \in \mathcal{D}esign_0} \textsc{Spread}(Design)$; best $\leftarrow \textsc{Spread}(Design^\star)$.
6: $\ell \leftarrow 0$. ▷ iteration counter
7: **while** $\ell < L$ **and** $Q \neq \emptyset$ **and** (τ unset **or** best $< \tau$) **do**
8: $Design_{\text{parent}} \leftarrow \textsc{PopMax}(Q)$. ▷ top of the priority queue
9: Generate a set $\mathcal{C}$ of child designs by copying $Design_{\text{parent}}$ and applying one or more random modifications to:
▷ the ordering of zones in one or several clusters; and/or
▷ the ordering of units in one or several zones.
10: **for** each $Design_{\text{child}} \in \mathcal{C}$ **do**
11: Compute $spread \leftarrow \textsc{Spread}(Design_{\text{child}})$.
12: Insert $Design_{\text{child}}$ into Q with key $spread$.
13: **if** $spread >$ best **then**
14: $Design^\star \leftarrow Design_{\text{child}}$; best $\leftarrow spread$.
15: **end if**
16: **end for**
17: $\ell \leftarrow \ell + 1$.
18: **end while**
19: **return** $Design^\star$.

Notes:

- Line 11: The spread score is computed by a chosen index from those introduced earlier (e.g., DI/MI/VI/BI).

8 Simulations

In this section, we compare the proposed design and spreadness index with established spatially balanced sampling methods and existing indices, using a diverse collection of synthetic and real-world populations. To ensure transparency and reproducibility, all algorithms, population generators, and simulation scripts are implemented and publicly available in the `graphical-sampling` package (Panahbehagh et al., 2025), accessible at https://github.com/mehdimhb/graphical-sampling. The simulation study is organized around three main components: the populations to which the designs are applied, the set of sampling designs included for comparison, and the indices used to evaluate the degree of spatial spread, as detailed below.

- **Populations:**
  - **synthetic Populations** (Figure 1):
    Three synthetic population types of size $N = 100$ were considered: gridded, random, and clustered configurations. Each population was evaluated under both equal-probability (EP) and unequal-probability (UP) scenarios. In the UP case, inclusion probabilities increase gradually from left to right, represented by larger point sizes for higher-probability locations. This setting provides a controlled framework to examine how the n-Means method adapts to both uniform and spatially varying inclusion-probability structures.
  - **Swiss Amphibians Population**[3] (Figure 10):
    The dataset, provided by the *Centre de coordination pour la protection des amphibiens et des reptiles de Suisse* (karch), contains spatial coordinates for 959 observation sites across Switzerland along with site-level attributes. Geographic coordinates were used as spatial information, and the variable *Area* (surface extent of each site) served as the auxiliary variable to define UP inclusion probabilities. Extremely large *Area* values were truncated at 150 to prevent over-weighting.
    When applying one of the most computationally demanding spread designs, the WAV method, a runtime warning[4] indicated that simulations would be prohibitively slow for $N = 959$. To retain the spatial and auxiliary-variable structure while ensuring feasibility, we constructed a reduced population using conventional k-means clustering with $k = 100$. The observation nearest to each cluster centroid was retained, yielding a representative population of size $N = 100$.
  - **Meuse Population** (Figure 11):
    The Meuse dataset, available in the `sp` package of R (Pebesma and Bivand, 2005), contains spatial coordinates and soil characteristics measured in the floodplain of the River Meuse near Stein, the Netherlands. The data include topsoil heavy-metal concentrations and several soil and landscape attributes, based on composite samples covering areas of approximately 15×15 meters. In our analysis, spatial coordinates were used as location information, and copper concentration (Cu) was employed as the auxiliary variable to define UP inclusion probabilities.

[3] https://www.karch.ch

[4] `WARNING: Your population size is greater than 500; this could be quite time-consuming...`

  - **Simulated Regular and Aggregated Populations** (Figure 12):
    To further evaluate performance of our desing, we used simulated populations introduced by Robertson, Price and Reale (2024). Two configurations were examined: (i) *Regular populations*, generated as balanced acceptance samples (Robertson et al., 2013) with $N = 1000$ points over the unit square; and (ii) *Aggregated populations*, consisting of $N = 1027$ points generated via a conditional Neyman–Scott process with clustered spatial structure. The regular configuration emphasizes uniform spatial spread, whereas the aggregated configuration represents strongly clustered distributions, providing complementary benchmarks for assessing design robustness.

- **Sampling Designs:** The following sampling designs were included in the simulations. Some were implemented directly using the authors' available code, while others were faithfully reproduced based on the descriptions and results reported in their respective papers.

  - **WAV:**
    The Weakly Associated Vector method constructs spatially balanced samples by iteratively perturbing the inclusion-probability vector along directions weakly associated with spatial proximity, progressively fixing units while preserving first-order inclusion probabilities (Jauslin, 2020).

  - **LP1:**
    The Local Pivotal Method (variant 1) applies pairwise updates to neighbouring units so that exactly one is retained in each step. This induces spatial repulsion and yields samples with strong local balance and reduced clustering (Grafström, Lundström and Schelin, 2012; Grafström and Lisic, 2016).

  - **SCP:**
    Spatially Correlated Poisson sampling introduces dependence among the selection indicators of nearby units, reducing the likelihood of jointly including close neighbours and improving the spatial spread of the resulting sample (Grafström, 2012).

  - **GRT:**
    Generalized Random Tessellation Stratified sampling maps multi-dimensional populations to a one-dimensional ordered frame through recursive tessellation, from which systematic selection produces well-spread, spatially balanced samples (Stevens Jr. and Olsen, 2004).

  - **HIP:**
    Halton Iterative Partitioning exploits the low-discrepancy properties of Halton sequences to recursively partition the spatial domain, generating quasi-random, uniformly distributed samples suitable for finite two-dimensional populations (Robertson et al., 2018).

  - **DAS:**
    Dynamic Assignment Sampling formulates the selection process as a sequence of linear assignment problems that iteratively optimize geometric spread while maintaining

prescribed inclusion probabilities. It supports dynamic sample-size adjustment and general auxiliary structures (Robertson, Price and Reale, 2024).

- **NMS** (Algorithm 4):
  n-Means Spatial Sampling partitions the population into n-means UP-balanced clusters and, within each cluster, into zones; a GFS step is then applied to achieve both global and local spatial spread. Each cluster has total inclusion probability 1 and is subdivided into $m = 4$ zones, each with total inclusion probability 0.25 . In our simulations, we used the *Centroidal Polar* ranking, as defined in Section 7, both for ordering zones and for ordering units within zones. Comparative experiments indicated that the choice of ranking can materially affect NMS performance; among the alternatives we tested, *Centroidal Polar* was generally the most efficient. A comprehensive analysis of ranking schemes is of independent interest; in this study, we standardize on *Centroidal Polar* to focus on the main design features.
- **GMS** (Algorithm 5):
  An intelligent greedy search built on NMS refines the GFS ordering while preserving fixed sample size and first-order inclusion probabilities. Starting from one or more NMS-based initial designs (fixed , predefined orderings), the procedure iteratively proposes local reorderings of zones within clusters and of units within zones, evaluates candidates using a spreadness index, and adopts the best-scoring configuration at each step. In our simulations, MI served as the guiding index. The search terminates when a target spread is achieved or a prespecified computational budget is reached.
- **SRS:**
  Simple Random Sampling Without Replacement serves as the baseline equal-probability design. Although it does not target spatial balance, it provides a natural benchmark for evaluating the gains achieved by spread-oriented methods (Tillé, 2006).
- **MAX:**
  Maximum-Entropy sampling selects samples that maximize entropy subject to given inclusion probabilities. While not explicitly spatial, it offers a design-based benchmark for assessing how spatially oriented algorithms enhance representativeness (Tillé, 2006).

Cube-based and sequential balancing variants proposed by Jauslin, Panahbehagh and Tillé (2022) and Panahbehagh, Jauslin and Tillé (2024) were not included. Although these approaches have potential to promote spatial balance through auxiliary-variable control, they have not yet been extended to operate directly on spatial coordinates or to incorporate geometric distances into their balancing mechanism.

- **Indices:**
  - **Moran Index (MI):** Assesses spatial autocorrelation via a spatial-weights matrix; negative values indicate dispersion, positive values indicate clustering, and values near zero indicate spatial randomness (Moran, 1950; Tillé, Dickson, Espa and Giuliani, 2018).

– **Voronoi Index (VI):** Measures deviation between observed and expected inclusion probabilities across Voronoi cells; lower values indicate greater spatial balance, while higher values indicate clustering or uneven coverage (Stevens Jr. and Olsen, 2004).

– **Balanced Voronoi Index (BI):** Extends VI by incorporating auxiliary-variable balance within cells; lower values indicate both spatial and auxiliary balance, whereas higher values signal imbalance across one or both dimensions (Prentius and Grafström, 2024).

– **Density Disparity Index (DI):** Quantifies spread through a translation-invariant, signed comparison under cluster-wise translations (balanced -means); values near zero indicate well-spread samples, negative values reflect over-dispersion, and positive values reflect under-dispersion.

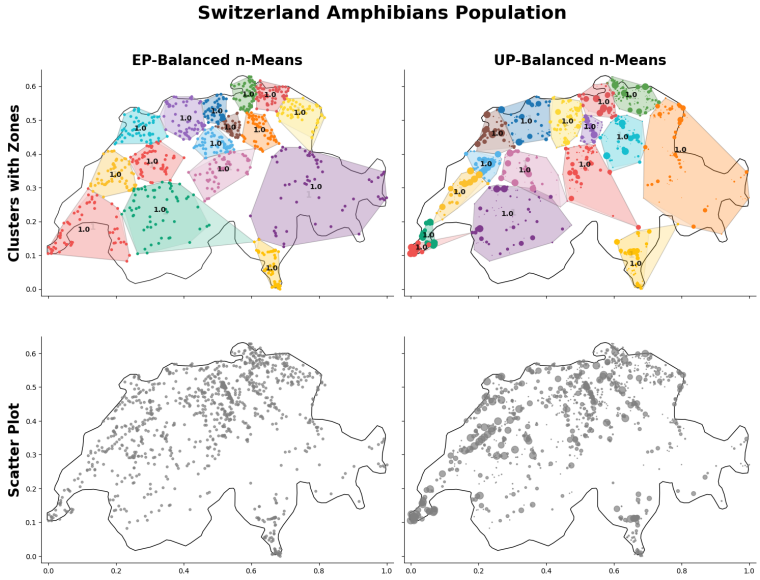

Figure 10: Amphibian observation sites in Switzerland under equal (EP, left) and unequal (UP, right) probabilities based on site area. Bottom row: scatter plots of sites; top row: n-Means UP-balanced clustering, one sample per cluster.

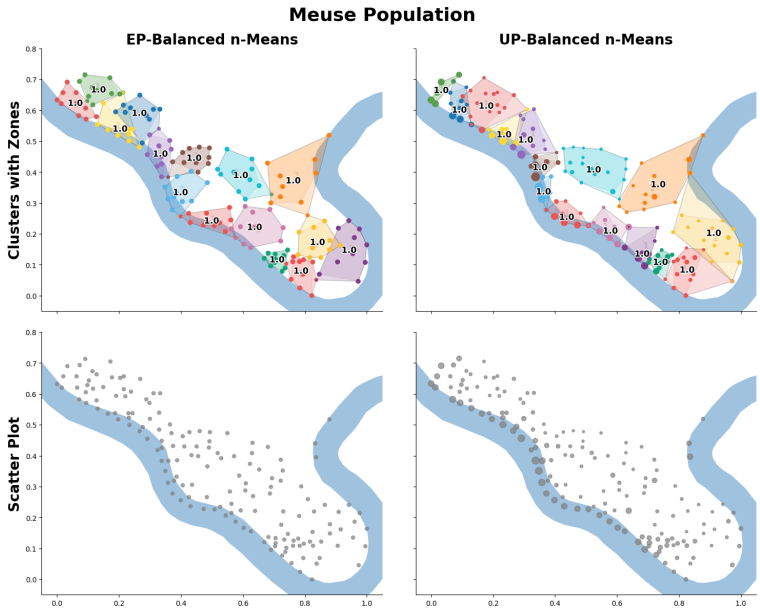

Figure 11: Meuse population showing sampling sites along the river under equal (EP, left) and copper-based unequal (UP, right) probabilities. Bottom row: scatter plots; top row: n-Means UP-balanced clustering, one sample per cluster.

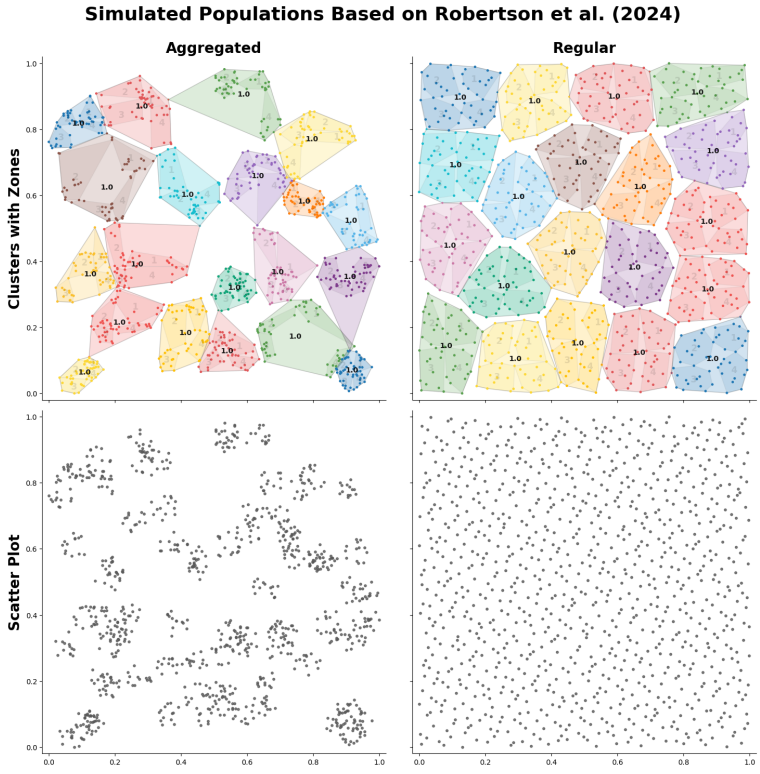

Figure 12: Simulated populations from Robertson et al. Robertson, Price and Reale (2024). Bottom row: scatter plots of aggregated (left) and regular (right) configurations; top row: n-Means UP-balanced clustering, one sample per cluster.

Using the selected populations, designs, and indices, we conducted extensive simulation experiments. Practical constraints limited direct replication of some competitors: WAV became

computationally prohibitive for larger N, and the available DAS implementation is in `Matlab`, which we could not feasibly integrate into our pipeline. Accordingly, we adopted the following protocol. When author-supplied generators were available, we regenerated their populations and applied our designs and indices under identical settings. When faithful reimplementation was not practical, we reported the competing method's results as published in the respective sources on the same (or author-provided) populations and metrics, explicitly indicating their provenance. This approach preserves comparability while maintaining transparency regarding which results are reproduced and which are drawn from prior work.

It is important to emphasize that the focus of this paper is solely on assessing the efficiency of sampling designs in terms of their ability to spread the sample. To avoid diverting attention from this objective, we did not consider efficiency with respect to parameter estimation (such as population totals) or issues related to variance estimation under these designs.

Preliminary comparisons indicated that WAV and LP1 consistently produced more efficient spreading than most other existing methods. Based on this, the simulations proceeded as follows. First, for the three synthetic populations and the Swiss amphibian population, we compared our proposed designs—NMS and GMS—with WAV, LP1, SRS, and MAX. For the Meuse population, we relied on the results of Jauslin (2020), and in the case of $n = 15$, we extended the comparison of NMS and GMS to include WAV, LP1, SCP, GRT, MAX, and SRS. Finally, to evaluate our methods against DAS, we incorporated the simulation results of Robertson, Price and Reale (2024), where we compared GMS with GRT, HIP, LP1, and DAS.

8.1 Results for Clustered, Gridded, Random, and Swiss Populations

The results of the Monte Carlo simulations are presented in Figures 13, 14, and 15, corresponding to MI, DI, and BI, respectively. As the BI represents a refined and extended version of the VI, incorporating auxiliary-variable balance, we report only BI to represent the family of Voronoi-based indices and avoid redundancy in presentation.

Across the three indices considered—MI, DI, and BI—the relative performance of the designs was broadly consistent with expectations, while also revealing several noteworthy findings. WAV and LP1 achieved strong spatial dispersion across all indices, with WAV consistently—and in most cases significantly—outperforming LP1. On BI and DI, NMS surpassed WAV, whereas on MI, WAV outperformed NMS in most settings. Motivated by this contrast, GMS was implemented with MI as the optimization criterion; as intended, GMS surpassed WAV on MI across all populations and sample sizes, with gains particularly pronounced for the Swiss data, while remaining competitive on BI and DI (generally comparable to WAV and LP1).

SRS and MAX, by contrast, demonstrated clear limitations in spreading samples, as anticipated. Their weaknesses were particularly evident under DI, which revealed a much broader range of outcomes (-0.60 to 0.30) compared with MI (approximately -0.15 to 0.15). This extended range highlights the sensitivity of the DI in detecting differences among designs and its ability to capture the long-tailed behavior of weakly spreading methods. Although SRS is not ideal under unequal probabilities, it was retained in the simulations to preserve comparability across plots.

For BI, occasional extreme values exceeding one were observed, which complicated interpre-

tation and distorted the figures; accordingly, the plots were clipped at 1.22.

Overall, the DI demonstrated particular interpretability, with values around zero indicating well-spread samples, increasingly negative values reflecting stronger dispersion, and positive values indicating concentration. This directional property makes DI a valuable complement to existing indices for assessing spatial spread.

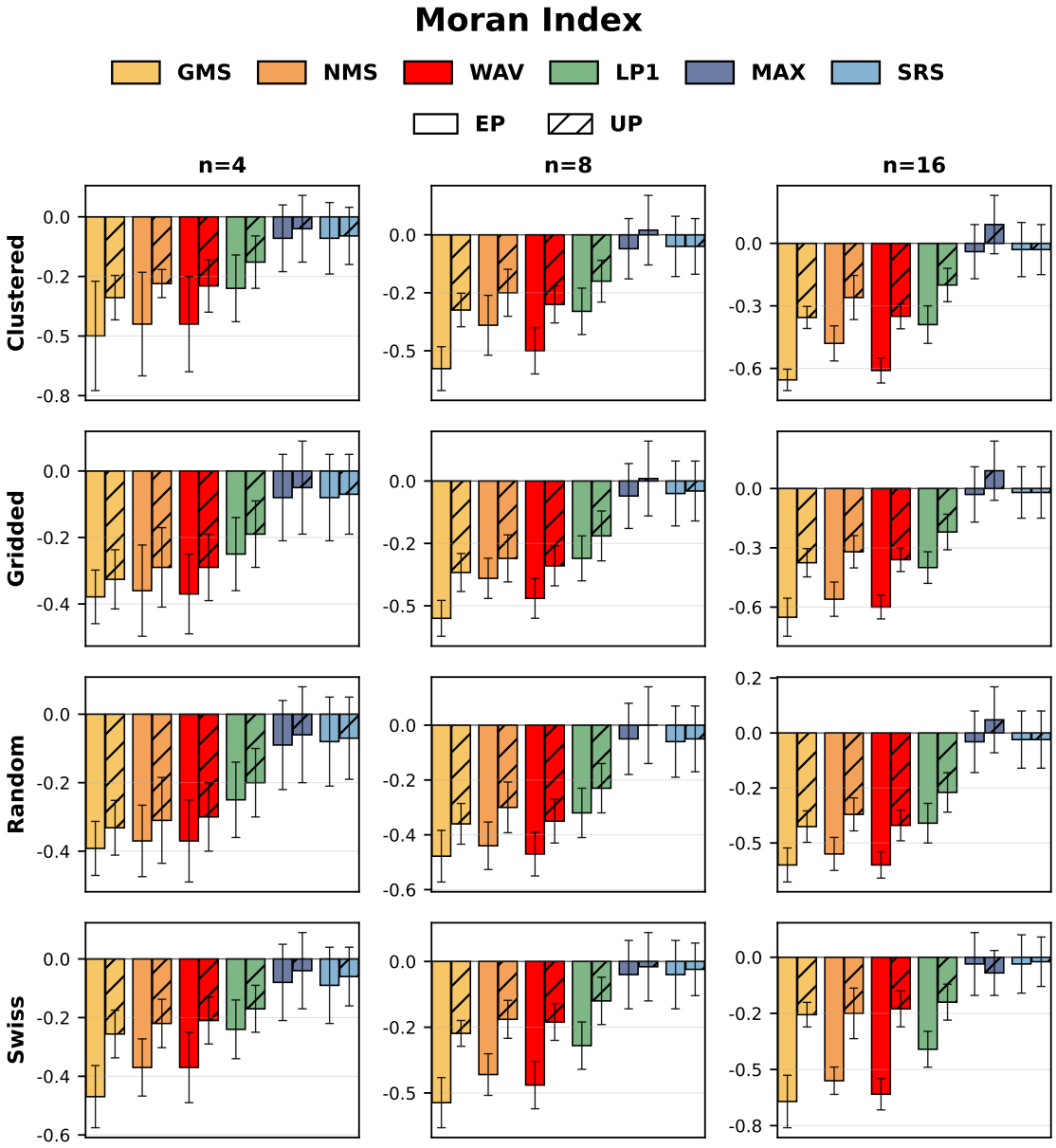

Figure 13: Results from Monte Carlo simulations under equal (EP) and unequal (UP) probability sampling schemes. Boxplots display the distribution of the Moran index obtained from different designs applied to three simulated populations (Clustered, Gridded, Random) and the observed Swiss population, with sample sizes $n = 4, 8, 16$.

8.2 Results for the Meuse Population

The results for the Meuse population with $n = 15$ clearly demonstrate the superior spreading performance of GMS relative to all competing methods. Under both EP and UP settings, and

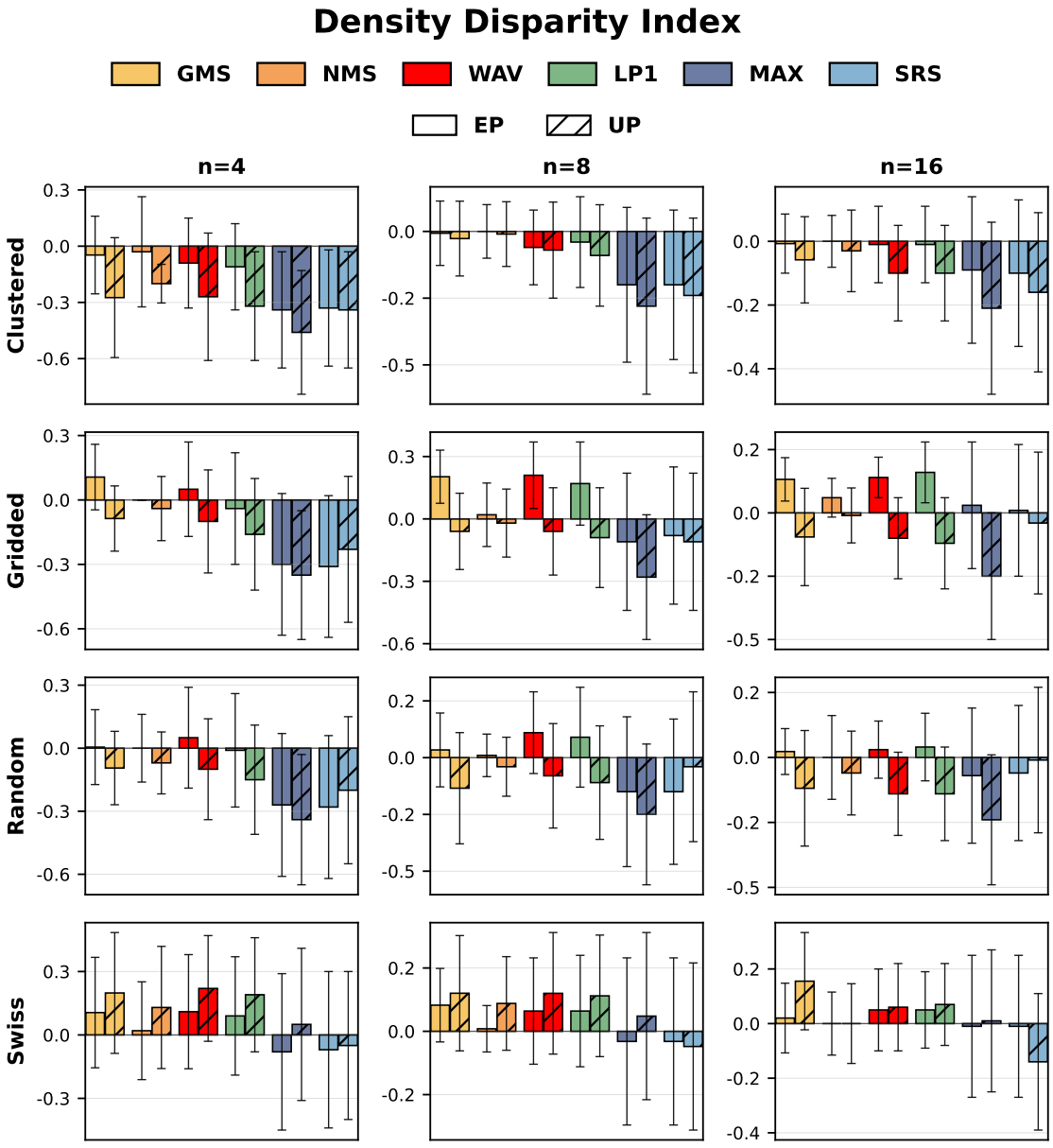

Figure 14: Results from Monte Carlo simulations under equal (EP) and unequal (UP) probability sampling schemes. Boxplots display the distribution of the Density Disparity Index obtained from different designs applied to three simulated populations (Clustered, Gridded, Random) and the observed Swiss population, with sample sizes $n = 4, 8, 16$.

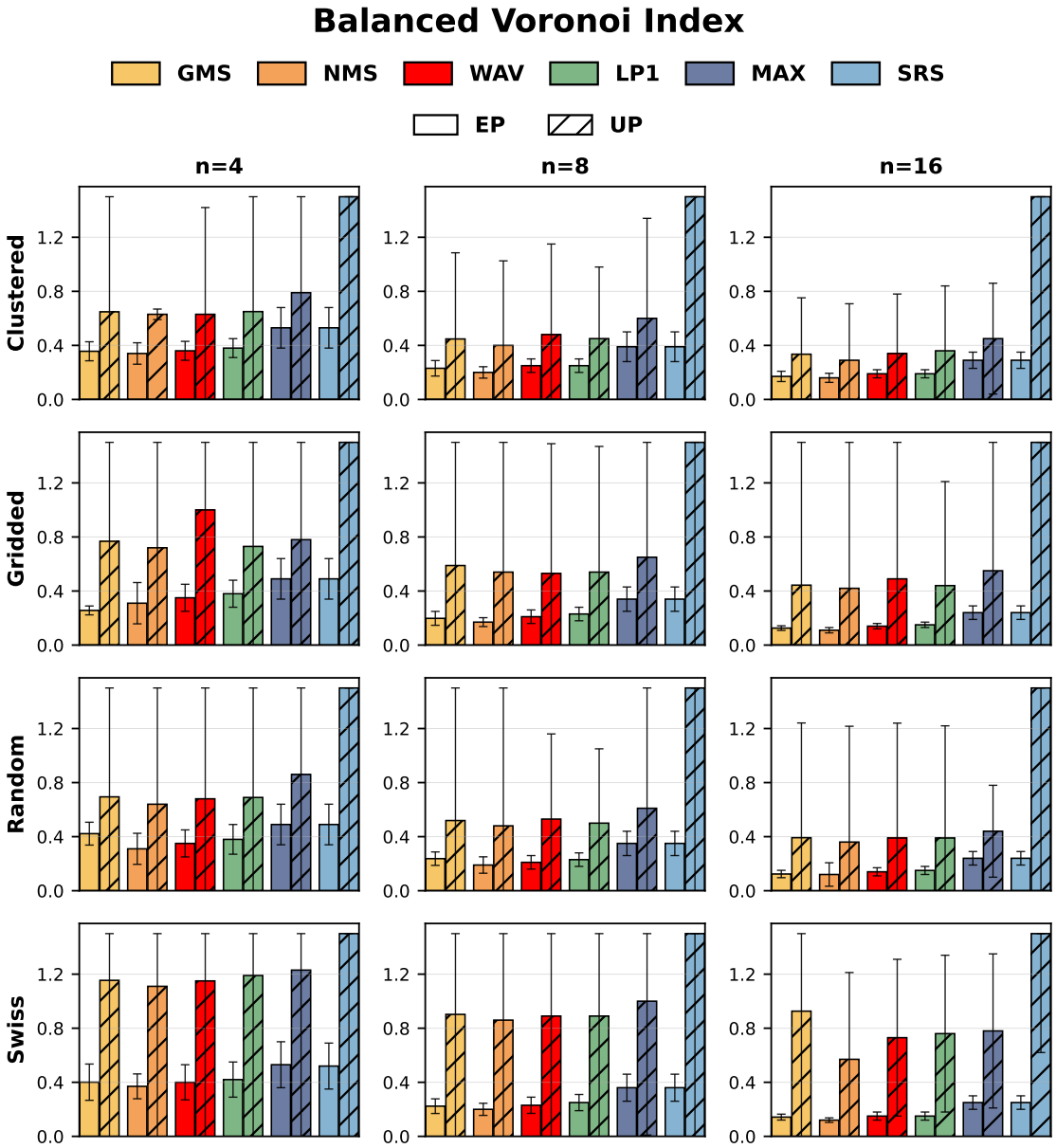

Figure 15: Results from Monte Carlo simulations under equal (EP) and unequal (UP) probability sampling schemes. Boxplots display the distribution of the Balanced Voronoi Index obtained from different designs applied to three simulated populations (Clustered, Gridded, Random) and the observed Swiss population, with sample sizes $n = 4, 8, 16$.

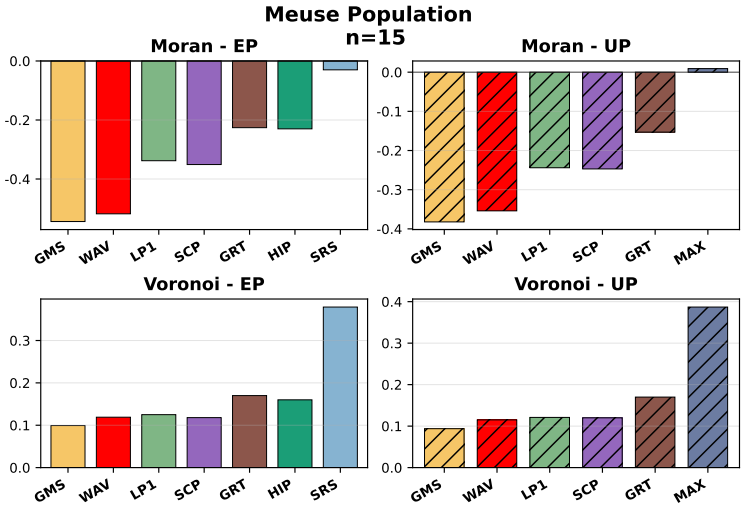

Figure 16: Comparison of spatial spreading indices for the Meuse population with $n = 15$. Results are shown under both EP and UP settings using Moran and the Voronoi Index. The proposed methods (GMS) are evaluated against existing spread-oriented designs (WAV, LP1, SCP, GRT, HIP) as well as baseline methods (SRS, MAX).

across both indices considered (Moran and the Voronoi index), GMS consistently achieves values closer to the optimal benchmark. This indicates that GMS produces samples that are more dispersed and spatially balanced than those obtained from established designs.

The advantage of GMS is particularly pronounced when evaluated under the UP setting. In both Moran and Voronoi-based measures, the separation between GMS and the next-best competitors (notably WAV and LP1) is considerably larger in UP compared to EP. This highlights the robustness of GMS in handling heterogeneity induced by unequal inclusion probabilities, a common challenge in practical survey scenarios.

SRS and MAX serve here as natural baselines rather than true competitors. While other methods such as SCP, GRT, HIP, and WAV demonstrate improvements relative to these baselines, none achieve dominance over GMS. This consistent superiority suggests that the geometric construction underlying GMS provides a more stable and generalizable framework for spreading, ensuring reliable performance across both EP and UP settings.

8.3 Results for Simulated Regular and Aggregated Populations

The results reported in Table 2 are based on the simulated regular and aggregated populations introduced by Robertson, Price and Reale (2024). For comparability, we reproduce the outcomes of GRT, HIP, LP1, and DAS from their study and supplement them with the performance of our proposed GMS design. Two indices are reported: the Voronoi index, where lower values indicate better spatial balance, and Moran, where more negative values indicate stronger dispersion.

Across both regular and aggregated populations, GMS consistently achieves Moran values that are substantially more negative than those of all other designs. This indicates that GMS produces samples with a markedly stronger degree of spatial dispersion. The superiority of GMS is maintained across all sample sizes, from $n = 20$ to $n = 200$, and becomes even more pronounced as the sample size increases, where the gap between GMS and competing methods

steadily widens.

In terms of the Voronoi index, GMS also dominates all competitors. For both population structures and across all sample sizes, GMS yields the lowest Voronoi values, indicating a level of spatial balance that is unmatched by GRT, HIP, LP1, or DAS. These differences are not marginal.

Taken together, the results show that GMS achieves simultaneous improvements on both indices: it strongly outperforms in Moran by delivering greater dispersion, and at the same time it provides superior Voronoi values, reflecting balanced coverage. This dual advantage underscores the effectiveness of the geometric construction underlying GMS, which ensures robust performance across different spatial structures and sample sizes.

Table 2: Comparison of spatial spreading indices (Voronoi and Moran) for regular and aggregated simulated populations from Robertson, Price and Reale (2024). Results for GRT, HIP, LP1, and DAS are reproduced from their study, while GMS is newly added. GMS consistently attains more negative Moran and lower Voronoi-based index values than all competitors across all sample sizes, thereby dominating on both criteria.

n	GRT		HIP		LP1		DAS		GMS	
	Voronoi	Moran	Voronoi	Moran	Voronoi	Moran	Voronoi	Moran	Voronoi	Moran
Regular Population										
20	0.38	−0.09	0.25	−0.13	0.24	−0.16	0.15	−0.24	0.08	−0.25
50	0.40	−0.12	0.25	−0.20	0.23	−0.24	0.16	−0.32	0.06	−0.36
100	0.44	−0.16	0.30	−0.24	0.24	−0.32	0.20	−0.38	0.06	−0.42
200	0.52	−0.19	0.41	−0.26	0.31	−0.43	0.32	−0.43	0.09	−0.47
Aggregated Population										
20	0.40	−0.11	0.37	−0.12	0.28	−0.18	0.24	−0.26	0.09	−0.27
50	0.44	−0.14	0.42	−0.17	0.29	−0.26	0.27	−0.34	0.12	−0.38
100	0.47	−0.18	0.46	−0.21	0.30	−0.35	0.30	−0.40	0.14	−0.45
200	0.52	−0.22	0.51	−0.25	0.33	−0.46	0.37	−0.45	0.17	−0.48

9 Summary and Conclusion

This paper introduced a graphical and intelligent framework for assessing and enhancing spatial spread in finite-population sampling. We proposed the Density Disparity Index, a kernel-based measure comparing the original population density with its cluster-wise translated counterpart, to quantify spatial distortion induced by a sample. DI complements MI and BI by distinguishing local redistribution effects from global autocorrelation or imbalance across cells. In addition, we developed n-Means UP-balanced clustering to form probability-balanced, spatially compact groups that support representative single-unit sampling, and used this structure as the basis for a GFS optimized for spatial dispersion under design-based validity. An intelligent search mechanism was further integrated, yielding the GMS design, which adaptively refines GFS orderings to maximize spread while preserving inclusion probabilities. Monte Carlo experiments on vast range of populations demonstrate that GMS dominates all competing designs, achieving more negative Moran's , close to zero DI, and lower or comparable BI values across sample sizes. The framework is practical—scaling efficiently through simple clustering and localized intelligent

search—modular, allowing flexible choice of kernels and balancing targets, and inferentially coherent, preserving design-based validity. Future work includes increasing the entropy of NMS by segmenting GFS bars into finer subcomponents, thereby enlarging the feasible neighborhood of reorderings while preserving fixed size and first-order inclusion probabilities. This higher-resolution search space would enable more intelligent optimization—e.g., genetic algorithms, artificial bee colony methods, and related metaheuristics—to explore diversified local and global permutations under design-based constraints.

References

Bradley, Paul S.; Bennett, Kristin P.; Demiriz, Ayhan. Constrained k-Means Clustering. Microsoft Research Technical Report MSR-TR-2000-65, Redmond, WA, 2000. https://www.microsoft.com/en-us/research/publication/constrained-k-means-clustering/.

de Maeyer, Rieke; Sieranoja, Sami; Fränti, Pasi. Balanced k-means revisited. Applied Computing and Intelligence, 3(2):145–179, 2023.

Grafström, Anton; Lundström, Niklas L. P.; Schelin, Lina. Spatially balanced sampling through the pivotal method. Biometrics, 68(2):514–520, 2012.

Grafström, Anton; Tillé, Yves. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics, 24(2):120–131, 2013.

Grafström, Anton; Lisic, Jonathan A. BalancedSampling: Balanced and spatially balanced sampling. R package, 2016. https://CRAN.R-project.org/package=BalancedSampling.

Grafström, Anton. Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference, 142(1):139–147, 2012.

Grafström, Anton. Why well spread probability samples are balanced? Open Journal of Statistics, 3(1):36–41, 2013.

Panahbehagh, Bardia; Mohebbi, Mehdi; HosseiniNasab, AmirMohammad; Hosseini Moghadam, Mehdi. graphical-sampling. PyPI software package, 2025. https://pypi.org/project/graphical-sampling/. Code: https://github.com/mehdimhb/graphical-sampling.

Horvitz, Daniel G.; Thompson, Donovan J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47:663–685, 1952.

Jauslin, Raphaël. Spatial spread sampling using weakly associated vectors. Journal of Agricultural, Biological and Environmental Statistics, 25(3):431–451, 2020.

Jauslin, Raphaël; Panahbehagh, Bardia; Tillé, Yves. Sequential spatially balanced sampling. Environmetrics, 33(8):e2776, 2022.

Kuhn, Harold W. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1–2):83–97, 1955.

Lloyd, Stuart P. Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2):129–137, 1982.

Madow, William G. On the theory of systematic sampling. Annals of Mathematical Statistics, 20(3):333–354, 1949.

Moran, Patrick A. P. Notes on continuous stochastic phenomena. Biometrika, 37(1/2):17–23, 1950.

Narain, Rupendra D. On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 3:169–174, 1951.

Pebesma, Edzer J.; Bivand, Roger S. Classes and methods for spatial data in R. R News, 5(2):9–13, 2005. https://journal.r-project.org/articles/RN-2005-014/.

Panahbehagh, Bardia; Jauslin, Raphaël; Tillé, Yves. A general stream sampling design. Computational Statistics, 39(6):2899–2924, 2024.

Panahbehagh, Bardia. Graphical Finite Population Sampling. arXiv:2308.07715, 2025. http://arxiv.org/abs/2308.07715.

Prentius, Wilmer; Grafström, Anton. How to find the best sampling design: A new measure of spatial balance. Environmetrics, 35(7):e2878, 2024.

Robertson, Blair L.; Brown, Jennifer A.; McDonald, Trent; Jaksons, Peter. BAS: Balanced acceptance sampling of natural resources. Biometrics, 69(3):776–784, 2013.

Robertson, Blair L.; McDonald, Trent; Price, Chris; Brown, Jennifer. Halton iterative partitioning: spatially balanced sampling via partitioning. Environmental and Ecological Statistics, 25(3):305–323, 2018.

Robertson, Blair L.; Price, Chris; Reale, Marco. Well-spread samples with dynamic sample sizes. Biometrics, 80(2):ujae026, 2024.

Stevens Jr., Don L.; Olsen, Anthony R. Spatially balanced sampling of natural resources. Journal of the American Statistical Association, 99(465):262–278, 2004.

Tillé, Yves. Sampling Algorithms. Springer, New York, 2006.

Tillé, Yves; Dickson, Maria Michela; Espa, Giuseppe; Giuliani, Diego. Measuring the spatial balance of a sample: A new measure based on the Moran's I index. Spatial Statistics, 23:182–192, 2018.

Wand, Matt P.; Jones, M. Chris. Kernel Smoothing. Chapman and Hall/CRC, London, 1994.