Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LOCATE: Context-Aware GeoAI Decision Support for Population-Oriented Air-Quality Sensor Placement in Data-Constrained Cities

Fidel Raja Wabinyai

raja@airqo.net

Makerere University/ AirQo

Deogratius Okedi

Makerere University/ AirQo

Priscah Adrine

Makerere University/ AirQo

George William Settala

Makerere University/ AirQo

Usman Abdul-Ganiy

Makerere University/ AirQo

Jennifer Kutesakwe

National Environment Management Authority, Uganda

Alex Ndyabakira

Kampala Capital City Authority

Sharifah Buzeki

Kampala Capital City Authority

Sarah Zalwango

Kampala Capital City Authority

Engineer Bainomugisha

Makerere University/ AirQo

Deo Okure

Makerere University/ AirQo

Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11439350/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

LOCATE: Context-Aware GeoAI Decision Support for Population-Oriented Air-Quality Sensor Placement in Data-Constrained Cities

Fidel Raja Wabinyai*[1], Deogratius Okedi[1], Priscah Adrine[1], George William Settala[1], Usman Abdul-Ganiy[1], Jennifer Kutesakwe[2], Alex Ndyabakira[3], Sharifah Buzeki[3], Sarah Zalwango[3], Engineer Bainomugisha[1], and Deo Okure[1]

[1]*AirQo Project, Department of Computer Science, Makerere University, Kampala, Uganda*
[2]*National Environment Management Authority, Kampala, Uganda*
[3]*Kampala Capital City Authority, Kampala, Uganda*

Abstract

Air-quality sensor placement that prioritizes population proximity while accounting for urban context can support environmental health surveillance, but many cities lack reliable pollution surfaces to guide placement. LOCATE is a data-light, context-aware GeoAI decision-support framework for pre-deployment planning that prioritizes proximity to modelled residential population while incorporating OpenStreetMap road and building information. We compared 6 feature-informed clustering methods, 2 feature-informed allocation methods, and 2 uninformed spatial baselines in Kampala, Jinja, Nairobi, and Lagos City across nine sensor budgets (N = 5, 10, 20, 30, 40, 50, 60, 70, and 80), with 20 seeds for stochastic methods. Internal clustering diagnostics were evaluated separately from final-layout planning metrics. Under the declared population-and-context composite objective, the context-augmented population-weighted p-median heuristic ranked first in all 4 cities, while KMeans produced the strongest mean internal diagnostics among fixed-count clustering methods across the retained seven-budget diagnostic subset. Coverage improved with budget but remained lower in the larger Lagos City and Nairobi domains. LOCATE therefore supports population-oriented and urban-context-aware planning before installation, when observations at proposed sites do not yet exist; its outputs are planning shortlists rather than evidence of pollution reconstruction, exposure accuracy, or field feasibility.

Keywords: air-quality monitoring; geospatial artificial intelligence; population-oriented sensor placement; urban context; location–allocation; environmental health surveillance; decision support; low-cost sensors; OpenStreetMap.

*Corresponding author: Fidel Raja Wabinyai, email: raja@airqo.net

1 Introduction

Air-quality monitoring is an important component of environmental health surveillance, supporting exposure assessment, public communication, source-control policy, and evaluation of environmental interventions. In many low- and middle-income cities, however, reference-grade networks remain sparse because instruments, station infrastructure, calibration, specialist personnel, and long-term maintenance are costly (deSouza et al., 2020; Bainomugisha et al., 2023). Satellite products and atmospheric models extend spatial information, but locally operated ground observations remain necessary for evaluating those products and interpreting neighbourhood-scale conditions. Lower-cost sensors provide a practical route to denser and more flexible monitoring, provided that calibration, quality assurance, and operational performance are adequately managed (Clements et al., 2022; Frederickson et al., 2022; Okorn and Iraci, 2024).

The value of a sensor network depends not only on the number of instruments but also on where they are placed and what the network is intended to represent. Urban-background monitoring generally favours spatial balance and separation from immediate sources, traffic-oriented monitoring favours road corridors and intersections, and population-oriented monitoring favours proximity to residential concentrations. In this study, population-oriented and context-aware placement means that proximity to modelled residential population is the primary objective, while road and building variables provide secondary information about the urban settings represented by proposed monitors. It does not estimate individual exposure or imply spatially uniform pollution. Source-attribution and intervention studies may require paired, repeated, or upwind-downwind locations. The same city and sensor budget defined here as the number of sensors available for placement can therefore support several defensible layouts, each corresponding to a different objective. Site placement should consequently be treated as an explicit design problem rather than as a final logistical decision (Kanaroglou et al., 2005; Clements et al., 2022).

This problem is especially consequential in African cities, where monitoring may exist but routine observations, detailed emissions inventories, and validated urban dispersion models can remain insufficient to construct a reliable citywide prior pollution surface. In this study, *data-constrained* describes limited inputs for placement optimization rather than the complete absence of deployed sensors. At the pre-deployment stage, observations from proposed sites do not yet exist, so planning must use the spatial demand and contextual information already available while reserving pollution and field validation for later stages. Relevant sources include road transport, industrial activity, open waste burning, and household fuel use, as documented for Kampala (Okure et al., 2025). Cities also differ substantially in administrative extent, boundary geometry, settlement density, road structure, and population distribution. A network size that provides broad population proximity in a compact secondary city may leave extensive gaps in a larger metropolitan domain. Sensor requirements should therefore be evaluated from city-specific evidence rather than transferred directly from another city or prescribed as a universal target.

Network-design approaches include location-allocation models, geostatistical optimization, emissions and dispersion models, wireless-network optimization, suitability analysis, and multi-objective formulations (Kanaroglou et al., 2005; Boubrima et al., 2019; Mukherjee et al., 2020; Fattoruso et al., 2020; Ajnoti et al., 2024). Their principal strength is alignment with a specified objective such as reconstruction error, information gain, source influence, communication cost, exposure, or equity. Their application can nevertheless be constrained when a city lacks a

prior pollution field, a validated atmospheric model, a detailed emissions inventory, or consistently complete local covariates. Under these conditions, a transparent coverage-oriented design provides a defensible starting point, although its conclusions must remain narrower than those of a network validated against pollution observations.

Unsupervised clustering and location-allocation methods can generate such initial layouts without historical pollutant measurements, but they formalize different design questions. Clustering methods represent the candidate feature space through compact, probabilistic, hierarchical, or density-defined groups. Location-allocation methods instead select feasible facilities according to explicit criteria such as demand-weighted proximity or separation across spatial and contextual conditions. These method families are related but should not be judged using a single form of evidence. Internal clustering diagnostics describe the structure of candidate groups before site mapping and repair, whereas deployment metrics evaluate residents' proximity to proposed monitors, population coverage, residual geographic gaps, contextual representation, and compliance with placement constraints. A valid comparison should therefore use common candidate locations, budgets, contextual inputs, and feasibility rules, while reporting stochastic variability where applicable.

Open geospatial data can make placement context-aware without implying that it is optimized for pollution. WorldPop provides a gridded representation of population distribution that can enter fitting as a demand weight and can also evaluate nearest-site distance and threshold coverage (Tatem, 2017; WorldPop and Center for International Earth Science Information Network, 2018). OpenStreetMap (OSM) provides road density, intersections, building counts, and footprint coverage that can distinguish urban contexts represented by proposed sites (Haklay and Weber, 2008; OpenStreetMap contributors, 2026). In LOCATE, these variables were used both to construct a common feature representation for feature-informed methods and to characterize final layouts. They remain proxies: neither source directly measures emissions or pollutant concentrations, and neither confirms access, security, power, communications, airflow, or host willingness.

The study was designed for the pre-deployment stage of public-sector network planning. LOCATE produces candidate-site shortlists for institutional and field review rather than automatic installation decisions.

LOCATE combines boundary-constrained candidate generation; a robustly scaled feature space containing projected coordinates, population density, roads, intersections, buildings, and building coverage; six clustering methods; two feature-informed allocation methods; two uninformed spatial baselines; common feasibility repair; population, geographic, and urban-context assessment; and repeated runs of stochastic methods. Internal cluster structure and final-layout planning performance were evaluated separately so that compact feature-space representation was not treated as equivalent to deployment coverage and contextual representation. Existing monitors can be retained as fixed sites before the remaining budget is allocated.

The main research question is: *How can spatial allocation, modelled population, and OSM-derived urban context be combined in a reproducible data-light framework to support population-oriented and urban-context-aware pre-deployment planning across cities and sensor budgets when a reliable prior pollution surface is unavailable or deliberately excluded?* The study addresses this question in Kampala and Jinja, Uganda; Nairobi, Kenya; and Lagos City, Nigeria, for budgets of $N = 5$, 10, 20, 30, 40, 50, 60, 70, and 80 sites. Existing monitor locations and measurements were withheld

from the main comparison so that the experiment represented prospective planning from spatial inputs available before proposed sites are instrumented. Specifically, the study examines: (i) how population-weighted and geographic coverage change with budget; (ii) which method performs best under a declared population-coverage-urban-context composite; (iii) which fixed-count clustering method produces the strongest internal diagnostics; and (iv) whether the conclusions are consistent across cities of different scale and form. The study does not estimate pollution reconstruction accuracy, hotspot-detection performance, source attribution, regulatory representativeness, realized exposure, or field feasibility.

2 Data and Methods

The analysis used five stages: input and constraint definition; spatial and contextual feature preparation; model fitting; feasibility repair and layout evaluation; and ranking and output generation (Figure 1). Operational pollutant measurements were excluded. Internal clustering diagnostics were calculated from pre-repair labels, and layout metrics were calculated from the final repaired coordinates.

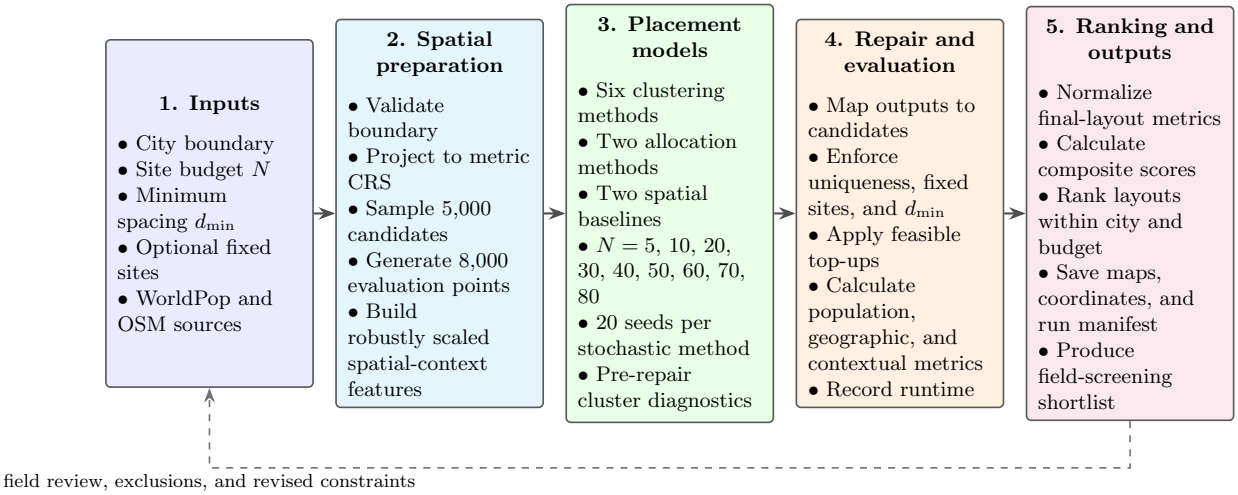


Figure 1: Analytical workflow for generating, evaluating, and ranking pre-deployment sensor layouts. Internal clustering diagnostics are calculated before feasibility repair, and field review can revise constraints before deployment.

2.1 Study area

The evaluation covers Kampala and Jinja in Uganda, Nairobi in Kenya, and Lagos City in Nigeria. These cities were purposively selected to test the framework across contrasting population sizes, urban forms, infrastructure systems, and environmental constraints rather than to represent all African urban areas statistically. Kampala and Jinja provide a within-country comparison between a densely developed national capital and a smaller secondary industrial city; Nairobi adds a large regional capital with heterogeneous land uses and multimodal transport; and Lagos City provides

a coastal metropolitan-region case with extensive waterways and fragmented developable land. Uganda's 2024 census recorded 1,797,722 residents in Kampala Capital City and 279,184 in Jinja City, Kenya's 2019 census recorded 4,397,073 residents in Nairobi City County, and UN-Habitat estimated 14.368 million residents in the Lagos urban agglomeration in 2020 (Uganda Bureau of Statistics, 2024; Kenya National Bureau of Statistics, 2019; United Nations Human Settlements Programme, 2020). Because these figures refer to different geographic definitions and reference years, and the Lagos figure is an estimate rather than a census count, they are used only to indicate relative urban scale and were not model inputs.

The cities also provide distinct infrastructure conditions for sensor placement. Kampala combines dense residential, commercial, and industrial development with a radial road network and continuing road, drainage, mass-transit, and commuter-rail investment, whereas Jinja combines established industrial and urban roads with newly incorporated, less densely developed areas (Kampala Capital City Authority, 2025; Jinja City Council and Ministry of Lands, Housing and Urban Development, 2022). Nairobi contains major highways, arterial roads, railways, airports, and dense formal and informal development, but its administrative domain also includes the approximately 117 km^2 Nairobi National Park (Nairobi City County Government, 2023; Kenya Wildlife Service, n.d.). Lagos City combines roads, ports, buses, and ferries across islands, lagoons, creeks, and wetlands, creating a substantially different infrastructure and accessibility pattern (Lagos State Resilience Office, 2020).

2.2 Evaluation design and sensor budgets

The sensor budget N was defined as the target number of monitoring locations, with one sensor assigned to each selected site. Nine budgets were evaluated: $N \in \{5, 10, 20, 30, 40, 50, 60, 70, 80\}$. A minimum pairwise separation of $d_{\min} = 0.5$ km was applied to every layout. The same budgets and spacing value were used for all four cities. Table 1 summarizes the experimental design.

Table 1: Component and specificication

Component	Specification
Study cities	Kampala, Jinja, Nairobi, and Lagos City
Site budgets	N = 5, 10, 20, 30, 40, 50, 60, 70, and 80
Master candidate locations	5,000
Evaluation locations	8,000
Minimum site separation	0.5 km
Placement approaches	6 clustering, 2 feature-informed allocation, and 2 uninformed baseline methods
OSM context radii	50 and 100 m
Population surface	WorldPop 2020, approximately 1 km

2.3 Population and urban context data

The 2020 WorldPop population-density surfaces for Uganda, Kenya, and Nigeria, provided at 30 arc-second resolution (approximately 1 km at the equator) and adjusted to United Nations Population Division totals, supplied candidate features, demand weights, and evaluation weights (Tatem, 2017; WorldPop and Center for International Earth Science Information Network, 2018). The country-specific surfaces were supplied as GeoTIFF (`.tif`) raster files: the Uganda raster was used for Kampala and Jinja, the Kenya raster for Nairobi, and the Nigeria raster for Lagos City. These modelled values were treated as comparative spatial weights rather than household-level counts. For a proposed sensor layout T, the exploratory population-context index using a 500-m (0.5-km) buffer radius was defined as

$$L_{500}(T) = \pi(0.5)^2 \sum_{t \in T} \bar{\rho}_{500}(t), \tag{1}$$

where $\bar{\rho}_{500}(t)$ is the mean population density within a 500-m (0.5-km) buffer radius of site t. Overlapping buffers were not dissolved before summation. OSM driveable-road and building-footprint attributes were attached to candidates using projected layers, nearest-feature joins, and 50- and 100-m circular buffers (OpenStreetMap contributors, 2026; Haklay and Weber, 2008). Major roads comprised motorway, trunk, primary, and secondary classes, and intersections were nodes with at least three incident streets. Table 2 lists the derived variables. The 100-m road density, intersection count, building count, and building coverage fraction entered the feature matrix; the remaining variables were retained as site attributes. Query failures and missing layers were recorded. Features with fewer than two valid values or no within-comparison variation were excluded from composite-score normalization.

Table 2: Road and building context variables derived from OpenStreetMap.

Layer	Nearest-feature and classification variables	Variables calculated within 50- and 100-m buffers
Driveable roads	Nearest-road class; nearest-road distance; road-within-50-m and road-within-100-m indicators; major-road-within-100-m indicator	Clipped total road length; road density in m km^{-2}; number of nodes with at least three incident streets
Buildings	Nearest-building type; nearest-building distance; building-within-50-m and building-within-100-m indicators	Number of intersecting building footprints; clipped footprint area in m^2; building coverage fraction of the circular buffer

2.4 Candidate generation and spatial representation

For each city, a master candidate set $C = \{c_1, \ldots, c_{5000}\}$ and a separate evaluation set $E = \{e_1, \ldots, e_{8000}\}$ were sampled uniformly by area within the projected boundary using independent seeds. The same C and E sets were reused for every method and sensor budget within that city: C supplied the common candidate locations, while E supplied the common support for coverage evaluation. Beyond holes already encoded in the boundary geometry, no separate permanent-water or protected-area exclusion mask was applied. Candidates and uniform evaluation points could therefore occur in water bodies or protected land retained within an analytical polygon. This

preserves a standardized full-domain comparison, but it means that the upper-tail geographic endpoint represents the full boundary rather than eligible residential land alone.

Uniform candidates represent equal geographic area rather than equal population or emissions potential. Each candidate was represented by projected easting and northing, log-transformed population density, log-transformed 100-m road density and intersection count, log-transformed 100-m building count, and 100-m building coverage fraction. Missing values were median-imputed within each city. Features were robustly scaled using the 10th–90th percentile range and clipped to $[-6, 6]$ to limit the influence of extreme values. Group weights were 1.00 for space, 1.25 for population, 1.00 for roads, and 1.00 for buildings; weights for space, roads, and buildings were divided equally between their two component variables. Longitude and latitude were retained for exchange and display, while spacing, buffering, and coverage calculations used projected metre coordinates.

2.5 Location-allocation, clustering, and baseline methods

The ten placement approaches were organized into four groups. Direct allocation comprised the context-augmented population-weighted p-median heuristic and feature-space maximin. Fixed-count clustering comprised KMeans, the approximate K-medoids procedure, Gaussian mixture modelling (GMM), and Ward-linkage agglomerative clustering. DBSCAN and HDBSCAN formed the density-based group, while random selection and a hexagonal-grid approximation served as uninformed spatial baselines. Let $\mathbf{x}_i$ denote the projected coordinates and $\mathbf{z}_i$ the fused, robustly scaled feature vector of candidate $c_i \in C$. Wherever a method used distance, it used Euclidean distance in the indicated projected-coordinate or fused-feature space. Table 3 summarizes the implemented settings; estimator options not listed in the notebook were left at the defaults of the software versions recorded in the run manifest.

Table 3: Implemented placement algorithms and verified principal parameters. Every method returned an ordered list of master-candidate indices that subsequently entered the same geographic feasibility-repair procedure.

Method	Category and fitting space	Verified settings and initialization	Candidate-site proposal
p-median heuristic	Direct allocation; projected geography with population demand and contextual reward	Facility pool $\min(\lvert C\rvert, \max(600, 3N))$; 3,000 sampled demand points; $\lambda = 0.25$ km; greedy addition; at most two swap passes; seeded pool and demand sampling	Selected facility candidates in greedy order, modified by accepted swaps
Maximin	Direct allocation; fused features	One uniformly seeded starting candidate; farthest-first traversal to N candidates	Candidates selected directly from C
KMeans	Fixed-count clustering; fused features	N clusters; 30 initializations; Lloyd algorithm; seeded estimator	Each fitted centroid mapped independently to its nearest candidate in fused-feature space
Approximate K-medoids	Fixed-count clustering; fused features	One seeded subsample of $\min(\lvert C\rvert, \max(700, 3N))$ candidates; N random initial medoids; at most 20 update iterations	Final medoids are sampled master candidates; all candidates labelled by their nearest returned medoid
GMM	Fixed-count probabilistic clustering; fused features	N full-covariance Gaussian components; 10 initializations; seeded estimator	Each component mean mapped independently to its nearest candidate in fused-feature space
Ward agglomerative	Fixed-count hierarchical clustering; fused features	N clusters; Ward linkage; deterministic fit	Each cluster represented by an exact or bounded-search feature-space medoid
DBSCAN	Density-based clustering; fused features	$\varepsilon = 0.85$ feature units; minimum samples = 20; Euclidean neighbourhoods; deterministic fit	Noise excluded; at most the N largest non-noise clusters represented by medoids
HDBSCAN	Hierarchical density-based clustering; fused features	Minimum cluster size = 80; minimum samples = 15; deterministic fit	Noise excluded; at most the N largest non-noise clusters represented by medoids
Random baseline	Uninformed baseline	Seeded uniform sample of N distinct master candidates	Sampled candidates returned in draw order
Hex-grid baseline	Geographic baseline; projected coordinates	Staggered lattice; spacing searched for at most 24 bisection iterations from 50 m to the maximum boundary-box extent	In-boundary lattice nodes mapped to nearest geographic candidates; excess unique candidates reduced by geographic maximin

Context-augmented population-weighted p-median heuristic. The p-median is a discrete facility-location model that selects N facilities to reduce weighted demand-to-facility distance (Hakimi, 1964; Teitz and Bart, 1968). The implemented method augmented this criterion with a small reward for urban-context priority and used restricted candidate and demand sets for computational tractability. For seed r, the facility pool $P_r \subseteq C$ contained

$$|P_r| = \min\{|C|, \max(600, 3N)\}. \tag{2}$$

Half of this pool comprised candidates with the highest context-priority scores, and the remainder was sampled without replacement from the other candidates. For the evaluated budgets, $|P_r| = 600$. The demand set $D_r \subset E$ was a seed-specific sample of 3,000 evaluation points. WorldPop weights within D_r were normalized to sum to one; if their sum had been zero, the implementation would have substituted uniform weights. Context priority $q_s \in [0, 1]$ was the weighted mean of percentile-scaled population, road, and building priorities, using weights 1.25:1:1. For selected set $S \subseteq P_r$, $|S| = N$,

the implemented objective was

$$J_r(S) = \sum_{e_i \in D_r} \widetilde{w}_i \min_{s \in S} d(e_i, s) - \frac{\lambda}{N} \sum_{s \in S} q_s, \tag{3}$$

where $\widetilde{w}_i = w_i / \sum_{e_j \in D_r} w_j$, $d(e_i, s)$ is projected distance in kilometres, and $\lambda = 0.25$ km. Starting with $S = \varnothing$, greedy addition evaluated every unselected facility and appended the candidate producing the smallest post-addition value of Equation 3. The algorithm then made at most two ordered swap passes. At each selected position, it temporarily retained the other $N - 1$ facilities, evaluated replacement by every currently unselected facility, and accepted a replacement only when it reduced J_r by more than 10^{-9}. It stopped early when a complete pass accepted no swap. Thus, population influenced demand weights, while population, roads, and buildings influenced both construction of P_r and the reward term. The method was a seeded greedy add-and-swap heuristic, not an exact p-median optimizer and not a pure population-only p-median.

Feature-space maximin. Maximin used a seeded farthest-first traversal (Gonzalez, 1985). It chose the first candidate uniformly from C. If S_k denotes the k candidates already proposed, the next proposal was

$$c_{k+1} = \arg\max_{c \in C \setminus S_k} \min_{s \in S_k} \|\mathbf{z}_c - \mathbf{z}_s\|_2. \tag{4}$$

Distances to the selected set were updated incrementally until N candidates had been proposed. Eligibility and geographic separation were not incorporated into this traversal; they were imposed by the common repair step. The method therefore promoted diversity across spatial, population, road, and building features rather than directly minimizing population-weighted geographic distance.

KMeans. KMeans partitioned all candidates into N groups by minimizing within-cluster squared Euclidean distance in fused-feature space,

$$\min_{\{a_i\}, \{\boldsymbol{\mu}_k\}} \sum_{k=1}^{N} \sum_{i : a_i = k} \|\mathbf{z}_i - \boldsymbol{\mu}_k\|_2^2, \tag{5}$$

where a_i assigns candidate i to cluster k and $\boldsymbol{\mu}_k$ is its centroid (MacQueen, 1967; Pedregosa et al., 2011). The scikit-learn implementation used 30 initializations, the Lloyd algorithm, `copy_x=False`, and the run seed as `random_state`; the initialization rule and numerical convergence settings were not overridden in the notebook and therefore followed the archived package-version defaults. Each fitted centroid was mapped independently to

$$s_k = \arg\min_{c_i \in C} \|\mathbf{z}_i - \boldsymbol{\mu}_k\|_2. \tag{6}$$

This was feature-space snapping, not geographic snapping. Independent mapping could return the same candidate for more than one centroid; duplicates were subsequently rejected and replaced during common repair.

Approximate K-medoids. The custom procedure was a single-subsample, CLARA-style approximation rather than an exhaustive Partitioning Around Medoids search (Kaufman and Rousseeuw, 1990). For each seed it sampled without replacement

$$M = \min\{|C|, \max(700, 3N)\} \tag{7}$$

candidates; $M = 700$ for all reported budgets. It initialized N distinct medoids uniformly within this sample. Each iteration assigned sampled candidate i to its nearest current medoid and updated cluster G_k to the observed candidate

$$m_k = \arg\min_{c_j \in G_k} \sum_{c_i \in G_k} \|\mathbf{z}_i - \mathbf{z}_j\|_2. \tag{8}$$

For a cluster containing more than 600 sampled candidates, the minimization in Equation 8 was restricted to 600 seeded candidate medoids while retaining all cluster members in the cost calculation. Iteration stopped when the medoid indices were unchanged or after 20 iterations. The final medoids were already master candidates and therefore required no centroid-snapping operation. For internal diagnostics, every master candidate was labelled by its nearest returned medoid before feasibility repair.

Gaussian mixture model. GMM fitted N full-covariance Gaussian components to the fused features by maximizing the mixture log-likelihood

$$\ell(\Theta) = \sum_{i=1}^{|C|} \log \left[\sum_{k=1}^{N} \pi_k \phi(\mathbf{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right], \tag{9}$$

using 10 initializations, full covariance matrices, and the run seed (Fraley and Raftery, 2002). Component initialization and numerical stopping settings were not overridden in the notebook and therefore followed the archived package-version defaults. Candidates were labelled by their fitted component assignments. Each component mean $\boldsymbol{\mu}_k$ was mapped independently to the nearest master candidate in fused-feature space using Equation 6; duplicate or geographically conflicting proposals were handled only during common repair.

Ward-linkage agglomerative clustering. Ward clustering began with singleton candidates and repeatedly merged the pair of clusters causing the smallest increase in total within-cluster sum of squares until N clusters remained (Murtagh and Contreras, 2012). If A and B are candidate clusters, the Ward merge cost is proportional to

$$\Delta(A, B) = \frac{|A||B|}{|A| + |B|} \|\bar{\mathbf{z}}_A - \bar{\mathbf{z}}_B\|_2^2. \tag{10}$$

Each final cluster was represented by its feature-space medoid. The medoid was exact for clusters of at most 600 candidates; for larger clusters, 600 possible medoids were sampled with fixed seed 0 and the candidate minimizing total distance from all cluster members was returned. The fixed seed made this bounded medoid search reproducible, and the complete Ward procedure was treated as deterministic under the fixed candidate set.

DBSCAN and HDBSCAN. DBSCAN defined the ε-neighbourhood of candidate i as

$$\mathcal{N}_\varepsilon(i) = \{j : \|\mathbf{z}_i - \mathbf{z}_j\|_2 \leq \varepsilon\}, \tag{11}$$

with $\varepsilon = 0.85$ in fused-feature units and `min_samples`$= 20$ (Ester et al., 1996). Candidates satisfying the core-point rule were joined through density reachability, while candidates not assigned to a density-connected cluster received noise label -1. HDBSCAN constructed a hierarchy from mutual-reachability structure using minimum cluster size 80 and minimum samples 15; other selection settings were the package defaults recorded with the computational environment (Campello et al., 2013). Neither density method was instructed to return N clusters. Noise-labelled candidates were excluded when forming site proposals. Non-noise clusters were ordered from largest to smallest; if more than N were found, only the N largest were represented. Each retained cluster used the same exact-or-bounded feature-space medoid rule described for Ward clustering, with fixed seed 0 for clusters exceeding 600 candidates. If fewer than N representatives survived, the common geographic maximin top-up supplied the remaining sites. Such final layouts are therefore density-informed hybrid layouts rather than unmodified DBSCAN or HDBSCAN outputs.

Random and hexagonal-grid baselines. The random baseline drew N distinct indices uniformly without replacement from the master candidate set using the run seed; it fitted no spatial, population, road, or building objective. The deterministic hexagonal-grid approximation constructed a staggered triangular lattice in projected coordinates. For horizontal spacing h, consecutive rows were separated by $h\sqrt{3}/2$ and alternating rows were offset by $h/2$. Only lattice nodes covered by the projected boundary were retained. A bisection search over $h \in [50\ \text{m}, \max(W, H)]$, where W and H are the boundary bounding-box dimensions, ran for at most 24 iterations and retained the lattice whose number of in-boundary nodes was closest to N. Each lattice node was then mapped to its nearest master candidate in projected geographic space, and duplicate candidate indices were removed. If this produced more than N unique candidates, a geographic maximin traversal with fixed seed 0 retained N of them; if fewer than N remained, the common repair procedure supplied geographic maximin top-ups.

Candidate representation and entry into feasibility repair. KMeans and GMM were the only methods whose fitted representatives were continuous feature-space centres; these centres were independently snapped to nearest master candidates in fused-feature space. Hexagonal-grid nodes were continuous geographic points and were snapped in projected coordinate space. K-medoids, Ward, DBSCAN, and HDBSCAN returned observed cluster members as medoids, while the p-median heuristic, maximin, and random baseline selected master candidates directly. Snapping did not itself enforce candidate eligibility, uniqueness, or the 0.5-km geographic separation rule. All proposed indices, including duplicate or ineligible snaps, were passed in their returned order to the common repair procedure described below.

2.6 Repeated runs and metric reporting

For each city and budget, KMeans, approximate K-medoids, GMM, random selection, feature-space maximin, and context-augmented population-weighted p-median were run with all 20 seeds from

101 to 120 (Gentle, 2003). The seed controlled the KMeans and GMM estimator initialization, the K-medoids candidate sample and initial medoids, random candidate selection, the initial maximin candidate, and the p-median facility-pool and demand-set samples. Each run continued until the method-specific stopping rule was reached: selection of N candidates for random and maximin, estimator convergence for KMeans and GMM, unchanged medoids or 20 iterations for K-medoids, and two swap passes or one complete non-improving pass for p-median. The boundary, master candidate set, evaluation set, and contextual sources remained fixed across seeds. Hexagonal-grid, Ward agglomerative, DBSCAN, and HDBSCAN were deterministic under the specified inputs and were fitted once. Their outputs were stored with seed 101 as a reference identifier only; seed 101 did not represent repeated stochastic fitting for these methods.

For each stochastic method, city, and budget, successful final-layout runs were reduced to one method-level value per metric using the median across seeds; the interquartile range was retained to describe run-to-run variability. Each deterministic method contributed its single value under the fixed candidate set. The deterministic value was compared with the stochastic median but was not interpreted as a seed distribution or as evidence of zero uncertainty. Interquartile ranges were reported descriptively and did not enter the composite score. Internal diagnostics were calculated from pre-repair labels for the six clustering methods. The retained internal-diagnostic table reports mean values across runs and the seven-budget subset $N = 5$, 10, 20, 30, 40, 50, and 60 within each city. Additional checks at $N = 70$ and $N = 80$ were used to assess ranking stability. DBSCAN and HDBSCAN values describe native clusters before top-up and repair. All comparisons were descriptive; no formal cross-city hypothesis tests were performed.

2.7 Constraint handling and optional fixed locations

Let F be a set of valid fixed locations and let S be the model-selected candidates. Fixed sites are checked against the boundary and the spacing rule before model selection. For the main multicity comparison, F was empty despite the presence of deployed monitors, because the purpose was to construct a common counterfactual benchmark rather than to redesign any operational network. In practical network expansion, existing monitors can be included in F; the implementation then retains the total target budget N and selects only the additional candidates needed, subject to the fixed-site distances.

Let $T = F \cup S$ be the combined shortlist. A common repair function enforces candidate eligibility, uniqueness, the requested count, and

$$\min_{1 \le i < j \le |T|} d(t_i, t_j) \ge d_{\min}, \tag{12}$$

where $d(\cdot, \cdot)$ is distance in kilometres. Proposals that violate a constraint are rejected, and the shortlist is completed with the feasible candidate that maximizes its distance from the sites already retained. If the target count cannot be achieved, the run is marked infeasible rather than silently relaxing $d_{\min}$. Rejections, top-ups, achieved count, and verified minimum distance are stored for each run.

2.8 Population, geographic, and urban-context evaluation endpoints and layout ranking

The primary planning objective was residential population representation, supplemented by geographic coverage and urban-context criteria. WorldPop values were used as spatial demand weights to quantify how close the modelled population would be to the nearest proposed monitor. Road and building variables characterized the urban settings represented by selected sites; they were not interpreted as emissions or pollutant measurements. The population-proximity endpoints do not measure pollutant concentration or personal exposure.

For evaluation point $e_i \in E$, let

$$\delta_i(T) = \min_{t \in T} d(e_i, t) \tag{13}$$

be its nearest-site distance. The geographic endpoints are the mean of δ_i, its 95th percentile, and its maximum. If w_i is the non-negative WorldPop density sampled at e_i, the primary endpoint is population-weighted nearest-site distance,

$$D_{\text{pop}}(T) = \frac{\sum_i w_i \delta_i(T)}{\sum_i w_i}. \tag{14}$$

For radius r, the population-coverage percentage is

$$P_{\text{pop}}(r) = 100 \, \frac{\sum_i w_i \, \mathbb{I}\{\delta_i(T) \le r\}}{\sum_i w_i}, \tag{15}$$

calculated for $r \in \{1, 2, 3\}$ km as prespecified nested proximity thresholds. The 1-km threshold provided the strictest assessment of local proximity, 2 km was the primary reporting threshold, and 3 km provided a broader assessment of population coverage. The primary 2-km threshold was informed by high-resolution sensor-network deployment practice in Kampala, where impact buffer zones of up to 2 km have been used (Lsoto et al., 2026); it was not estimated as an optimal distance from the results of this study. Reporting all three thresholds reduced dependence on a single radius. The radii describe distance to the nearest proposed monitor and were not treated as pollutant-dispersion, exposure, spatial-homogeneity, or regulatory-representativeness radii. Coverage was also calculated for evaluation points in the highest quartile of positive population-density values.

The primary comparison used three complementary final-layout endpoints. The first two were $D_{\text{pop}}(T)$ and $P_{\text{pop}}(2)$, which describe average population-weighted distance and the proportion of the weighted population within 2 km. The third was the 95th percentile of the unweighted nearest-site distances,

$$D_{95}(T) = Q_{0.95}(\delta_1(T), \ldots, \delta_{8000}(T)), \tag{16}$$

The metric directions used for comparison were lower for D_{pop} and D_{95} and higher for $P_{\text{pop}}(2)$.

2.9 Internal clustering diagnostics

Internal diagnostics were calculated for the six clustering methods before maximin top-up and feasibility repair. The Silhouette coefficient compares within-cluster cohesion with separation from

the nearest alternative cluster; higher values indicate better-defined clusters (Rousseeuw, 1987). The Davies–Bouldin index compares within-cluster dispersion with separation between cluster centres; lower values indicate better clustering (Davies and Bouldin, 1979). The Calinski–Harabasz score is the ratio of between-cluster to within-cluster dispersion, adjusted for the number of clusters and observations; higher values indicate better-separated, more compact clusters (Calinski and Harabasz, 1974). Metrics were calculated only when at least two valid clusters existed. For DBSCAN and HDBSCAN, noise-labelled candidates (-1) were excluded rather than reassigned. These diagnostics were not included in the final-layout composite ranking.

Placement approaches were ranked separately within each city and sensor budget using the stochastic medians and deterministic single estimates defined above. Six final-layout metrics entered the ranking: population-weighted nearest-site distance (weight 0.35; lower), 95th-percentile geographic distance (0.15; lower), population coverage within 2 km (0.20; higher), median population density at selected sites (0.10; higher), mean road density at selected sites (0.10; higher), and mean building coverage at selected sites (0.10; higher). The weights were study-specific choices reflecting population-oriented planning priorities, consistent with guidance on transparent composite-indicator weighting rather than prescribed numerical values (OECD/European Union/EC-JRC, 2008; Becker et al., 2017).

Each method-level metric was min–max normalized across methods within the same city–budget combination. Metrics with fewer than two valid values or no variation were excluded, and the remaining weights were renormalized to sum to one. Representative-seed selection was used only to choose coordinates for mapping and did not affect the method-level metric summaries or rankings. City-level summaries were calculated by averaging the nine budget-specific method-level metric values and composite scores and reranking methods within each city.

For each stochastic method, city, and budget, only runs with status `ok` entered representative-seed selection. Let $\mathcal{R}_{\text{ok}} \subseteq \{101, \ldots, 120\}$ be the successful seeds, let D_r be the population-weighted nearest-site distance for $r \in \mathcal{R}_{\text{ok}}$, and let $\widetilde{D}$ be their median. The representative seed was

$$r^* = \arg\min_{r \in \mathcal{R}_{\text{ok}}} |D_r - \widetilde{D}|. \tag{17}$$

Thus, the selected seed varied by method, city, and budget and was not fixed at 101. It represented a median-performing successful run rather than the run with the smallest distance or highest composite score. The run table was sorted by seed before selection, so an exact tie was resolved by retaining the smaller seed. Deterministic methods used their single fitted layout, whose stored seed value of 101 was only a reference label. Point colour represented context priority, and the method, composite score, and selected seed were stored with each mapped layout.

3 Results

Results are reported for final-layout performance, internal clustering diagnostics, and representative layouts. The final-layout comparison included 10 placement methods and 9 sensor budgets ($N = 5$, 10, 20, 30, 40, 50, 60, 70, and 80). Internal diagnostics were summarized for 6 clustering methods across the seven-budget subset through $N = 60$. The mapped output contained one representative p-median layout for each of the 36 city–budget combinations.

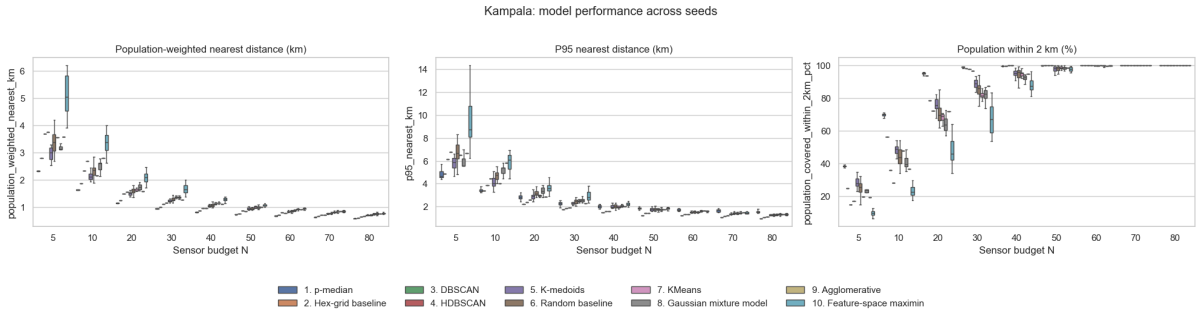
3.1 Population and geographic coverage by city and budget

Figure 2 presents population-weighted nearest-site distance, 95th-percentile geographic distance, and population coverage within 2 km for each city, method, and sensor budget. At $N = 80$, the p-median layouts placed approximately 100% of the modelled population within 2 km of a selected site in Kampala, 100% in Jinja, 98% in Nairobi, and 90% in Lagos City.

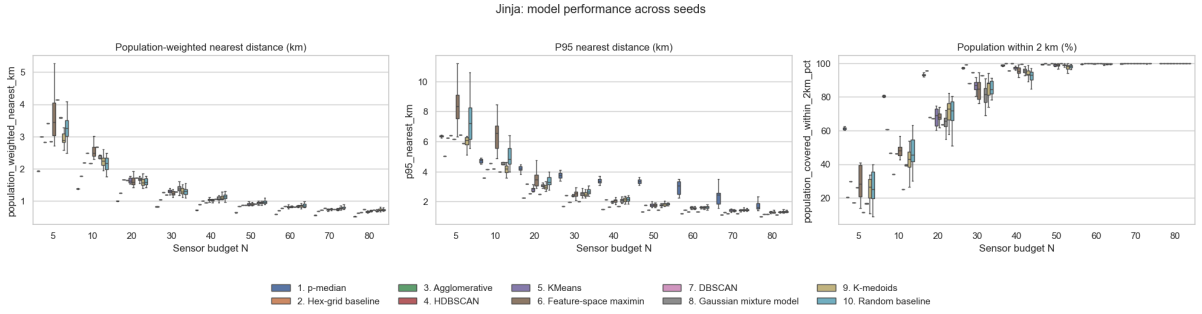
Table 4 reports city-level means obtained from the nine budget-specific summaries; the complete rankings are provided in Supplementary Table S1. The p-median ranked first and the hex-grid baseline ranked second in each city. For p-median, the mean composite score, population-weighted nearest-site distance, and 2-km population coverage were 84.01, 1.06 km, and 89.00% in Kampala; 77.13, 0.91 km, and 92.21% in Jinja; 84.61, 1.27 km, and 83.31% in Nairobi; and 86.43, 1.98 km, and 66.93% in Lagos City. Mean road density at the selected p-median sites was 16,509.7, 9,724.1, 18,083.6, and 12,773.5 m km^{-2} in Kampala, Jinja, Nairobi, and Lagos City, respectively.

Table 4: City-level mean final-layout performance for the six highest-ranked placement methods

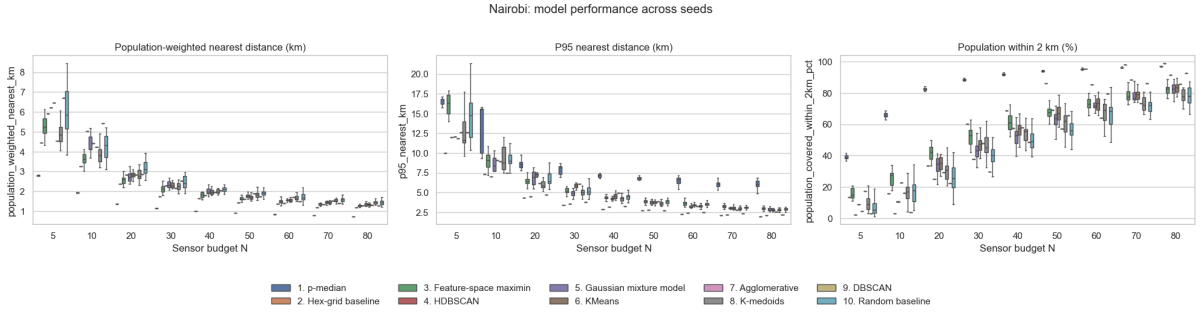
City	Rank	Method	Composite score	D_{pop} (km) ↓	$P_{\text{pop}}(2)$ (%) ↑	Mean road density (m /km^{-2})
Kampala	**1**	***p*-median**	**84.01**	**1.06**	**89.00**	**16,509.7**
Kampala	2	Hex-grid baseline	81.21	1.18	85.83	11,470.3
Kampala	3	DBSCAN	60.38	1.41	80.76	8,184.9
Kampala	4	HDBSCAN	56.87	1.47	79.30	8,110.9
Kampala	5	K-medoids	51.64	1.35	81.49	9,669.5
Kampala	6	Random baseline	47.37	1.45	79.83	10,383.0
Jinja	**1**	***p*-median**	**77.13**	**0.91**	**92.21**	**9,724.1**
Jinja	2	Hex-grid baseline	60.15	1.20	86.18	3,443.0
Jinja	3	Agglomerative	41.06	1.33	80.79	4,071.2
Jinja	4	HDBSCAN	38.81	1.43	79.20	3,558.9
Jinja	5	KMeans	38.46	1.34	80.52	4,059.4
Jinja	6	Feature-space maximin	35.75	1.44	80.37	6,613.4
Nairobi	**1**	***p*-median**	**84.61**	**1.27**	**83.31**	**18,083.6**
Nairobi	2	Hex-grid baseline	47.80	2.07	63.36	7,876.7
Nairobi	3	Feature-space maximin	36.86	2.35	55.47	10,495.4
Nairobi	4	HDBSCAN	29.49	2.60	51.17	5,706.2
Nairobi	5	Gaussian mixture model	26.09	2.60	49.52	9,299.0
Nairobi	6	KMeans	25.54	2.66	52.26	7,842.2
Lagos City	**1**	***p*-median**	**86.43**	**1.98**	**66.93**	**12,773.5**
Lagos City	2	Hex-grid baseline	55.11	2.63	44.59	3,437.2
Lagos City	3	DBSCAN	33.08	3.47	32.21	4,249.2
Lagos City	4	K-medoids	27.43	3.28	33.18	4,008.3
Lagos City	5	HDBSCAN	25.30	3.69	28.62	3,487.8
Lagos City	6	Gaussian mixture model	23.67	3.57	30.78	6,499.5



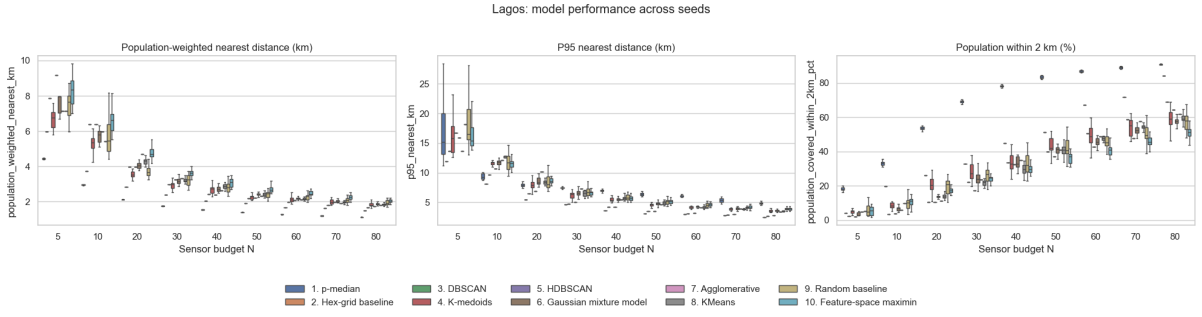
(a) Kampala.



(b) Jinja.



(c) Nairobi.



(d) Lagos City.

Figure 2: Box plot, Population-weighted distance, 95th-percentile geographic distance, and 2-km population coverage across cities and sensor budgets.

3.2 Internal clustering diagnostics

Table 5 reports mean internal diagnostics for the six clustering methods across the retained seven-budget subset through $N = 60$. Among the four methods constrained to return exactly N clusters, KMeans had the highest Silhouette and Calinski–Harabasz values and the lowest Davies–Bouldin value in each city. Its Silhouette, Davies–Bouldin, and Calinski–Harabasz values were 0.322, 0.934, and 3694.0 in Kampala; 0.376, 0.925, and 2158.3 in Jinja; 0.281, 1.237, and 866.5 in Nairobi; and 0.325, 1.036, and 1511.9 in Lagos City. GMM had the lowest Silhouette and Calinski–Harabasz values and the highest Davies–Bouldin value among the fixed-count methods in all four cities. DBSCAN diagnostics were defined only for Kampala, with values of 0.786, 0.309, and 2827.8. Checks at $N = 70$ and $N = 80$ did not materially change the method ranking.

Table 5: Mean internal clustering metrics (N = 5-80).

City	Model	Silhouette ↑	Davies–Bouldin ↓	Calinski–Harabasz ↑
Kampala	KMeans	**0.322**	**0.934**	**3694.0**
Kampala	GMM	0.039	1.733	1239.6
Kampala	KMedoids	0.227	1.257	2091.0
Kampala	Agglomerative†	0.274	1.005	3326.7
Kampala	HDBSCAN†	0.272	0.950	3087.1
Kampala	DBSCAN†	0.786	0.309	2827.8
Jinja	KMeans	**0.376**	**0.925**	**2158.3**
Jinja	GMM	0.258	1.500	1219.4
Jinja	KMedoids	0.318	1.054	1532.0
Jinja	Agglomerative†	0.348	0.962	1985.6
Jinja	HDBSCAN†	0.376	1.251	1623.0
Jinja	DBSCAN†	–	–	–
Nairobi	KMeans	**0.281**	**1.237**	**866.5**
Nairobi	GMM	0.092	1.848	434.0
Nairobi	KMedoids	0.214	1.422	687.4
Nairobi	Agglomerative†	0.246	1.310	772.8
Nairobi	HDBSCAN†	0.203	1.029	797.4
Nairobi	DBSCAN†	–	–	–
Lagos City	KMeans	**0.325**	**1.036**	**1511.9**
Lagos City	GMM	0.179	1.589	858.1
Lagos City	KMedoids	0.260	1.209	1129.0
Lagos City	Agglomerative†	0.295	1.066	1346.5
Lagos City	HDBSCAN†	0.298	1.325	1396.9
Lagos City	DBSCAN†	–	–	–

† Deterministic method under the fixed inputs; bold indicates the best result among methods constrained to return exactly N clusters. Reported values are means for N = 5-80. Density-based values describe native labels and may be undefined when a valid multi-cluster diagnostic was unavailable.

3.3 Representative p-median layouts across sensor budgets

Figure 3 presents the representative p-median layout for each city and budget. The figure contains 36 layouts covering four cities and nine budgets. The representative stochastic seeds shown in the

445 panels range from 101 to 116.

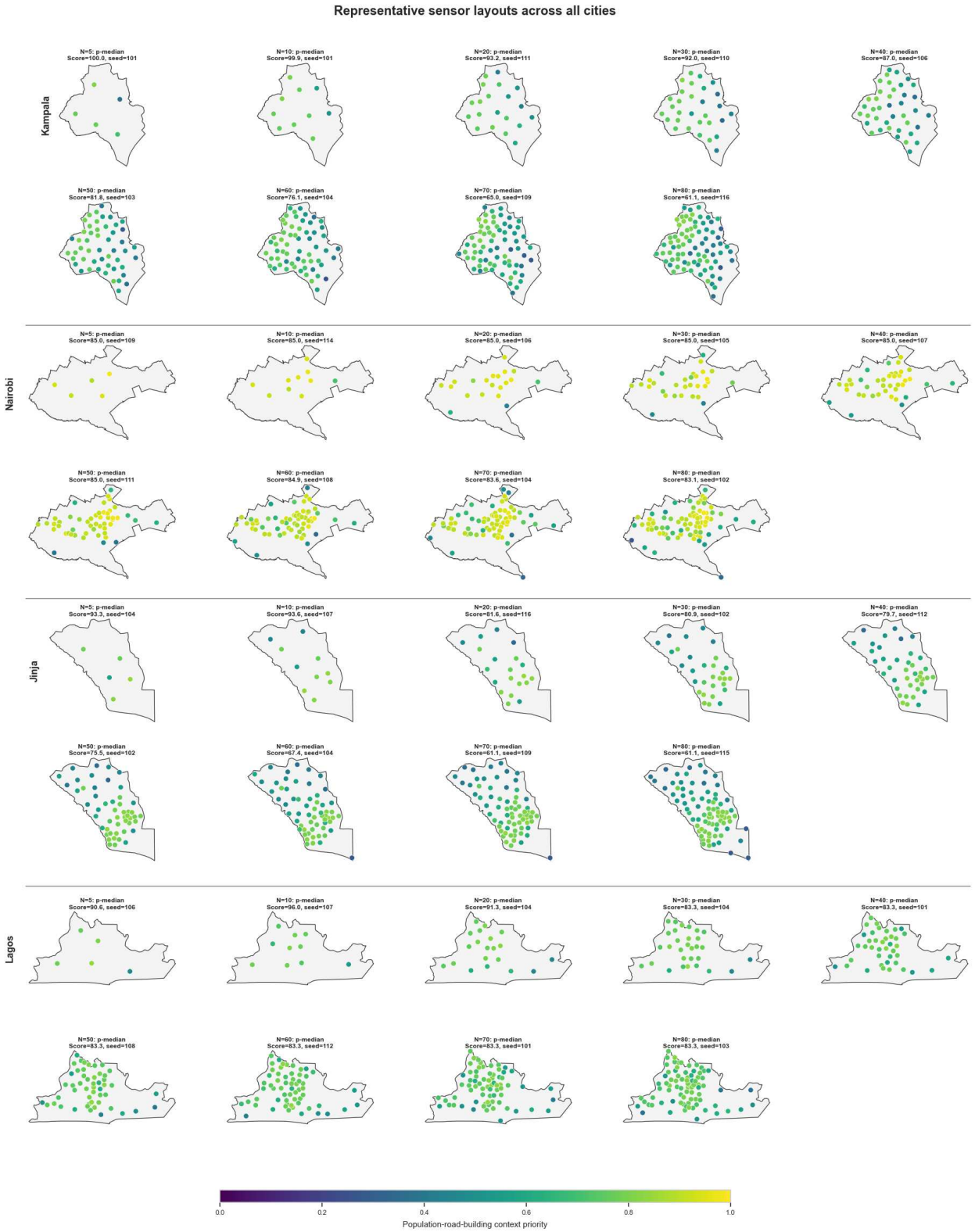


Figure 3: Representative p-median layouts for Kampala, Nairobi, Jinja, and Lagos City

4 Discussion

This study examined whether a data-light, context-aware spatial framework can support population-oriented planning of prospective air-quality sensor networks when a reliable prior pollution surface, detailed emissions inventory, or validated dispersion model is unavailable. Context-augmented population-weighted p-median performed best under the declared composite objective across the nine final-layout budgets, whereas KMeans produced the strongest mean internal diagnostics across the retained seven-budget diagnostic subset among the fixed-count clustering methods. These findings address different questions: p-median selects facilities to reduce population-weighted proximity gaps while favouring contextually informative candidates, whereas KMeans partitions the candidate feature space into compact groups. The strong p-median result is therefore partly attributable to the close alignment between its implemented objective and the population-proximity and contextual endpoints emphasized in the composite. More broadly, this reinforces the location–allocation principle that method performance should be interpreted against the planning objective being optimized (Kanaroglou et al., 2005; Boubrima et al., 2019; Mukherjee et al., 2020). Separating internal clustering quality from final-layout performance is consequently central to interpreting the comparison.

The four-city analysis showed how this distinction translated into spatial layouts. As the budget increased, additional sites reached previously unrepresented areas while retaining greater concentration in high-priority population, road, and building contexts. This concentration was more pronounced in Nairobi and Lagos City than in Kampala and Jinja. The p-median ranked first in the nine-budget mean for every city, while the hexagonal-grid baseline ranked second in all four cities. The mapped layouts in Figure 3 illustrate how the population-oriented allocation changed across the common budget range. The consistent second-place performance of the hexagonal grid also indicates that simple spatial balance can provide a competitive planning baseline. These rankings remain specific to the stated population-and-context objective and do not establish that either method is universally preferable.

The budget curves demonstrate why cities should not adopt a universal sensor count. At $N = 80$, the p-median layouts placed approximately 100% of the modelled population within 2 km in Kampala and Jinja, compared with approximately 98% in Nairobi and 90% in Lagos City. These are population-weighted estimates derived from the modelled evaluation support, not guarantees that every resident or populated location would be within 2 km of an installed sensor. The water and wetlands retained within the Lagos City boundary and Nairobi National Park primarily affect the full-boundary D_{95} metric; differences in population-weighted coverage more directly reflect the scale and spatial distribution of the modelled residential population. A city-specific minimum adequate planning budget can be defined as the smallest tested N that reaches a prespecified coverage target, but the discrete budget intervals used here identify only the first qualifying tested budget rather than an exact minimum. A finer search would be required around that point, and a strict requirement that all eligible populated locations lie within 2 km would require an explicit covering constraint rather than the present average-distance p-median objective.

The composite score expresses a particular planning preference; it is not an objective-independent definition of network quality. Population-weighted distance received the largest weight (0.35), followed by 2-km population coverage (0.20), with smaller contributions from geographic and contextual endpoints. Because each metric was normalized across methods within the same

city–budget comparison, composite values support method ranking within those comparisons but are not absolute quantities that should be compared directly between cities or budgets. Different priorities such as territorial surveillance, peripheral communities, traffic corridors, environmental justice, or pollution-field reconstruction would require different endpoints or weights and could change the method ranking. Composite scores should consequently be reported alongside population-weighted distance, threshold coverage, upper-tail geographic distance, contextual representation, runtime, and field constraints. Population proximity describes representation of modelled residential demand; it does not measure individual exposure or distributive equity.

WorldPop and OSM improve the contextual relevance of the shortlists but remain proxies. WorldPop does not fully capture recent settlement change, fine-scale population heterogeneity, commuting, or daily mobility (Tatem, 2017; WorldPop and Center for International Earth Science Information Network, 2018). OSM features distinguish road and building contexts but do not directly represent traffic volume, emissions, pollutant concentrations, installation permission, or site suitability (Haklay and Weber, 2008; OpenStreetMap contributors, 2026). OSM completeness also differed among cities; building-related features lacked usable variation in the Kampala and Lagos City caches and were excluded from the affected composite normalizations. Governments and city authorities should therefore use LOCATE to compare pre-deployment alternatives, then apply institutional and field review to assess hosts, security, power, communications, airflow, access, calibration, and maintenance (Clements et al., 2022; Bainomugisha et al., 2023).

The common candidates, budgets, contextual sources, geographic constraints, repair rules, and recorded random seeds improved comparability across methods. However, the current p-median implementation was a computational heuristic rather than an exact optimizer. It used a restricted facility pool, greedy addition, two swap passes, and a 3,000-point subset of the 8,000 evaluation locations as demand support. Because evaluation was not fully independent of optimization, the estimated advantage may be optimistic. Future comparisons should use separate demand-fitting and untouched evaluation samples, repeated candidate and evaluation realizations, larger facility pools, additional swap searches, and sensitivity analyses for the metric weights, contextual-reward parameter, coverage radius, and eligible-land mask.

The experiment further represents only one administrative boundary and one master candidate realization per city. The 20 seeds quantify algorithmic variability but not uncertainty in boundaries, geospatial inputs, or data versions. Permanent water should be treated as a hard exclusion in future population-oriented applications, while the treatment of protected areas should depend on whether the objective is residential, background, or ecological monitoring. Common absolute budgets and the 0.5-km separation rule also provide a controlled comparison rather than universally appropriate deployment standards. LOCATE should therefore be treated as a screening and prioritization stage that narrows the field-review problem, not as a substitute for pollution-based network validation or site-feasibility assessment. Post-installation studies should evaluate proposed layouts against fixed monitors, mobile-monitoring surfaces, or independently validated pollution simulations and document site displacement, rejection reasons, calibration, uptime, communications, and maintenance (Frederickson et al., 2022; Okorn and Iraci, 2024).

5 Conclusion

This study demonstrates that prospective air-quality sensor networks can be planned systematically using a population-oriented and context-aware approach before installation, without a reliable prior pollution surface, detailed emissions inventory, or validated dispersion model. LOCATE integrates projected geography, WorldPop, OSM road and building context, feasibility constraints, and explicit planning metrics in an auditable GeoAI decision-support workflow. Across Kampala, Jinja, Nairobi, and Lagos City, context-augmented population-weighted p-median remained the strongest overall method for the declared population-and-context objective over the nine budgets. KMeans was the strongest fixed-count clustering benchmark in the retained seven-budget internal-diagnostic summary. Increasing the budget from 5 to 80 sites improved coverage in every city, although gains narrowed at the larger budgets. The larger full-boundary geographic gaps in Lagos City and Nairobi partly reflect water bodies and protected non-residential land, while the population-weighted results still show that identical budgets provide unequal residential coverage. No universal urban sensor count is therefore appropriate.

As a context-aware pre-deployment tool, LOCATE can be used by municipal, metropolitan, and national governments to plan new air-quality monitoring networks or expand existing ones before instruments are installed. In any city with suitable boundary, population, and open urban-context data, it can help identify population-coverage gaps, compare the urban settings represented by alternative layouts, assess sensor-budget scenarios, and prioritize candidate areas for institutional review and field assessment. Method and network size should reflect local coverage needs, eligible land, existing networks, monitoring objectives, field constraints, and available resources. The resulting shortlists must therefore be followed by regulatory review, site verification, host engagement, and installation planning. LOCATE does not establish pollution reconstruction, individual exposure, or field feasibility; future work should validate the layouts using independent evaluation samples, alternative weights and land masks, repeated spatial realizations, post-installation pollution observations, and documented field outcomes.

CRediT authorship contribution statement

Fidel Raja Wabinyai: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. Deogratius Okedi: Writing – review & editing. Priscah Adrine: Investigation, Validation, Visualization, Writing – review & editing. George William Settala: Data curation, Investigation, Validation, Writing – review & editing. Usman Abdul-Ganiy: Writing – review & editing. Jennifer Kutesakwe: Validation, Writing – review & editing. Alex Ndyabakira: Investigation, Validation, Writing – review & editing. Sharifah Buzeki: Writing – review & editing. Sarah Zalwango: Writing – review & editing. Engineer Bainomugisha: Conceptualization, Funding acquisition, Writing – review & editing. Deo Okure: Project administration, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors acknowledge the use of ChatGPT GPT-5 (OpenAI) and Grok 4 (xAI) for assistance in editing and improving the language of the manuscript. The authors reviewed and revised all content and are solely responsible for the final version and the integrity of the work.

Funding

This work was supported by Google.org (grant 1904-57882); the EPSRC/GCRF (grant EP/T00343X/1); the Belgian Wehubit programme, implemented by Enabel (Grant Agreement BEL1707111-AP-05-2); the National Institute of Environmental Health Sciences of the U.S. National Institutes of Health (Award U01ES036147); AFRETEC; and the U.S. Mission (grant SUG50021CA3041). The content is solely the responsibility of the authors and does not necessarily represent the views of the funding organizations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank AirQo team members, collaborators, and project partners for their feedback on the study and manuscript.

Data availability

The city-boundary GeoJSON files, WorldPop dataset identifiers, candidate and evaluation point sets, run-level metrics, final site coordinates, saved OSM context layers, and analysis code are available from the corresponding author on reasonable request. The accompanying run manifest records the city, budget, method, seed, parameters, runtime, proposed and achieved site counts, native cluster count, noise fraction, coverage metrics, verified minimum distance, repair outcomes, coordinate reference systems, software and package versions, and OSM query dates. Any public archival release should preserve these materials under persistent identifiers to support exact replication.

References

Nishant Ajnoti, Hemant Gehlot, and Sachchida Nand Tripathi. Hybrid instrument network optimization for air quality monitoring. *Atmospheric Measurement Techniques*, 17:1651–1664, 2024. doi: 10.5194/amt-17-1651-2024.

Engineer Bainomugisha, Joel Ssematimba, and Deo Okure. Design considerations for a distributed low-cost air quality sensing system for urban environments in low-resource settings. *Atmosphere*, 14(2):354, 2023. doi: 10.3390/atmos14020354.

William Becker, Michaela Saisana, Paolo Paruolo, and Ine Vandecasteele. Weights and importance in composite indicators: Closing the gap. *Ecological Indicators*, 80:12–22, 2017. doi: 10.1016/j.ecolind.2017.03.056.

Ahmed Boubrima, Walid Bechkit, and Hervé Rivano. On the deployment of wireless sensor networks for air quality mapping: Optimization models and algorithms. *IEEE/ACM Transactions on Networking*, 27(4):1629–1642, 2019. doi: 10.1109/TNET.2019.2923737.

Tadeusz Calinski and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, pages 160–172, 2013. doi: 10.1007/978-3-642-37456-2_14.

Andrea L. Clements, Rachelle M. Duvall, Danny Greene, and Tim Dye. The enhanced air sensor guidebook. Technical Report EPA/600/R-22/213, U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, North Carolina, 2022. URL https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1017AWL.TXT.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

Priyanka deSouza, Ralph A. Kahn, James A. Limbacher, Eloise A. Marais, Fábio Duarte, and Carlo Ratti. Combining low-cost, surface-based aerosol monitors with size-resolved satellite data for air quality applications. *Atmospheric Measurement Techniques*, 13:5319–5334, 2020. doi: 10.5194/amt-13-5319-2020.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, 1996.

Giulia Fattoruso, Viviana Franco, Alessandro Bigi, and Guido Ghermandi. Site suitability analysis for air quality monitoring. *Atmosphere*, 11(11):1215, 2020. doi: 10.3390/atmos11111215.

Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. doi: 10.1198/016214502760047131.

Louise Bøge Frederickson, Ruta Sidaraviciute, Johan Albrecht Schmidt, Ole Hertel, and Matthew Stanley Johnson. Are dense networks of low-cost nodes really useful for monitoring air pollution? a case study in staffordshire. *Atmospheric Chemistry and Physics*, 22:13949–13965, 2022. doi: 10.5194/acp-22-13949-2022.

James E. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer, New York, 2 edition, 2003. doi: 10.1007/b97336.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

S. Louis Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3):450–459, 1964. doi: 10.1287/opre.12.3.450.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008. doi: 10.1109/MPRV.2008.80.

Jinja City Council and Ministry of Lands, Housing and Urban Development. Jinja city integrated physical development plan 2022–2032. Technical report, Jinja City Council, Jinja, Uganda, 2022. URL https://jinjacity.go.ug/node/185. Accessed: 2026-08-23.

Kampala Capital City Authority. Kampala unveils ugx 11.9 trillion strategic plan to transform the city. Kampala Capital City Authority, 2025. URL https://www.kcca.go.ug/news/1017/. Published 30 September 2025; accessed: 2026-08-23.

Pavlos S. Kanaroglou, Michael Jerrett, Jason Morrison, Bernardo Beckerman, M. Altaf Arain, Nicolas L. Gilbert, and Jeffrey R. Brook. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmospheric Environment*, 39(13):2399–2409, 2005. doi: 10.1016/j.atmosenv.2004.06.049.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

Kenya National Bureau of Statistics. 2019 kenya population and housing census, volume i: Population by county and sub-county. Technical report, Kenya National Bureau of Statistics, Nairobi, Kenya, 2019. URL https://www.knbs.or.ke/reports/kenya-census-2019/. Accessed: 2026-08-14.

Kenya Wildlife Service. Nairobi national park. Kenya Wildlife Service, n.d. URL https://kws.go.ke/park/nairobi-national-park/. Park information page; accessed: 2026-08-23.

Lagos State Resilience Office. Lagos resilience strategy. Technical report, Lagos State Government, Lagos, Nigeria, 2020. URL https://www.thegpsc.org/sites/gpsc/files/lagos_resilience_strategy.pdf. Accessed: 2026-08-23.

Dorothy Lsoto, Richard Sserunjogi, Gabriel Okello, Yaqiong Chen, Jonathan A. Patz, and James J. Schauer. A framework for interpreting air pollution sensor data: Bridging gaps in monitoring, analysis and community action. *Journal of the Air & Waste Management Association*, 2026. doi: 10.1080/10962247.2026.2635079.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

Rajib Mukherjee, Urmila M. Diwekar, and Naresh Kumar. Real-time optimal spatiotemporal sensor placement for monitoring air pollutants. *Clean Technologies and Environmental Policy*, 22(10): 2091–2105, 2020. doi: 10.1007/s10098-020-01959-z.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, 2(1):86–97, 2012. doi: 10.1002/widm.53.

Nairobi City County Government. Nairobi city county integrated development plan 2023–2027. Technical report, Nairobi City County Government, Nairobi, Kenya, 2023. URL https://nairobiassembly.go.ke/ncca/wp-content/uploads/paperlaid/2023/NAIROBI-CITY-COUNTY-INTEGRATED-DEVELOPMENT-PLAN-FOR-2023-2027-1.pdf. Accessed: 2026-08-23.

OECD/European Union/EC-JRC. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris, 2008. doi: 10.1787/9789264043466-en.

Kristen Okorn and Laura T. Iraci. An overview of outdoor low-cost gas-phase air quality sensor deployments: Current efforts, trends, and limitations. *Atmospheric Measurement Techniques*, 17:6425–6457, 2024. doi: 10.5194/amt-17-6425-2024.

Deo Okure, Sarath K. Guttikunda, Richard Sserunjogi, Priscilla Adong, Satish K. Dammalapati, Diana Lsoto, Peter Green, Engineer Bainomugisha, and Jian Xie. Integrated air quality information for kampala: Analysis of $pm_{2.5}$, emission sources, modelled contributions, and institutional framework. *Environmental Science: Atmospheres*, 5:471–484, 2025. doi: 10.1039/D4EA00081A.

OpenStreetMap contributors. Openstreetmap. https://www.openstreetmap.org, 2026. Accessed: 2026-01-29.

Fabian Pedregosa, Ga"el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and 'Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Andrew J. Tatem. Worldpop, open data for spatial demography. *Scientific Data*, 4:170004, 2017. doi: 10.1038/sdata.2017.4.

Michael B. Teitz and Polly Bart. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research*, 16(5):955–961, 1968. doi: 10.1287/opre.16.5.955.

Uganda Bureau of Statistics. National population and housing census 2024: Final report, volume 1. Technical report, Uganda Bureau of Statistics, Kampala, Uganda, 2024. URL https://www.ubos.org/nphc-2024-census-page/. Accessed: 2026-08-14.

United Nations Human Settlements Programme. World cities report 2020: Statistical annex. Technical report, UN-Habitat, Nairobi, Kenya, 2020. URL https://unhabitat.org/sites/default/files/2020/10/wcr_2020_statistical_annex_2.pdf. Reports a 2020 population estimate of 14.368 million for the Lagos urban agglomeration; accessed: 2026-08-30.

WorldPop and Center for International Earth Science Information Network. Global high resolution population denominators project. WorldPop, University of Southampton, 2018. 2020 UN-adjusted 1-km population-density rasters for Uganda, Kenya, and Nigeria; accessed: 2026-08-12.