Privacy-Preserving Smart-City Data Sharing for Urban Environmental Analytics: A k-Anonymity Case Study

Abubakari Alidu[1], Flavio DePaoli[1]
Dessislava Petrova-Antonova[2], Emil Hristov[2], Evgeny Shirinyan[2]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca
Viale Sarca 336, 20126 Milan, Italy
E-mails: a.alidu@unimib.it; flavio.depaoli@unimib.it
[2]GATE Institute, Sofia University St. Kliment Ohridski
125 Tsarigradsko Shose Blvd., 1113 Sofia, Bulgaria
E-mails: dessislava.petrova@gate-ai.eu; emil.hristov@gate-ai.eu; evgeny.shirinyan@gate-ai.eu

Keywords: urban analytics, privacy-preserving data sharing, k-anonymity, urban environmental monitoring, citizen sensing, urban data integration

Abstract

Urban analytics increasingly depends on combining sensitive municipal administrative records with heterogeneous sensing data, yet such integration is often constrained by privacy, governance, and data-sharing rules. This paper asks whether a municipality can report a monthly district-level ambient particulate-matter exposure proxy for an enrolled kindergarten cohort without sharing raw child-level records. The case study links kindergarten enrolment data in Sofia with citizen-sensed $PM_{2.5}$ and PM_{10} measurements through a three-pipeline workflow: municipality-side privacy processing, air-quality curation, and consumer-side analytics using only a released cohort and public district centroids. Air-quality data alone can describe ambient pollution, but the cohort release is required to define which children, age ranges, kindergartens, and districts are represented in the reporting population. For January 2025, the release contains 3,440 children and satisfies k-anonymity at $k = 12$ with respect to district, age range, and kindergarten, with no violating equivalence classes. Across $k \in \{5, 8, 12, 16, 20\}$, retention decreases from 97.5% to 87.3%, and represented districts decrease from 24 to 22. On the overlapping district set, district mean $PM_{2.5}$ summaries remain highly stable relative to the $k = 5$ release, with Spearman $\rho \geq 0.988$ for $k \geq 8$ and MAE below 0.3 $\mu g/m^3$. In a geocoding diagnostic, 45.9% of sampled administrative addresses were resolved under the evaluated protocol, motivating centroid linkage as a coverage-versus-precision design choice. The findings are bounded to one city, one reporting month, and a district-scale reporting task.

1. Introduction

Urban analytics increasingly relies on linking sensitive municipal records with heterogeneous sensing streams. This enables district-level reporting and planning, but introduces a central tension: reporting utility must be balanced with privacy, governance, and disclosure constraints (Peeters et al., 2024, Bernardo et al., 2024).

A particularly difficult case arises when one source is a municipal register containing person-level information. Such records may include direct identifiers, institutional assignments, and residential address strings that cannot be shared across organisational boundaries without strong safeguards. In addition to legal and governance barriers, address-based linkage is often fragile in operational settings. Administrative addresses are frequently incomplete or inconsistently formatted, making external geocoding unreliable and introducing missingness or bias into downstream analyses. As a result, even when a city has both administrative and sensing data available, turning it into a usable cross-institutional workflow remains non-trivial.

This paper addresses that challenge by proposing a privacy-preserving data-sharing workflow for urban environmental analytics. The specific analytical question is not only environmental but also cohort-specific: how can a monthly district-level ambient particulate-matter exposure proxy be reported for enrolled kindergartners when the raw municipal cohort cannot be shared? Air-quality stations and district boundaries can already produce an ambient pollution map, but they cannot identify which enrolled children, age groups, kindergartens, and districts are represented in the reporting population. Instead of transferring raw enrolment records, the municipality publishes a bounded cohort release in which direct identifiers and free-text addresses are removed, quasi-identifiers are controlled through k-anonymity, and only fields needed for downstream analysis are shared. The paper therefore quantifies how stricter disclosure controls affect retained cohort size, subgroup coverage, district representation, and district-level outputs, allowing stakeholders to judge both protection and usability for the intended task.

The paper makes three contributions. First, it presents a data-sharing workflow that combines sensitive municipal microdata with urban sensing data without releasing raw personal records. Second, it provides a measurable privacy–utility evaluation across alternative k thresholds using cohort retention, subgroup coverage, represented-district coverage, and district-level $PM_{2.5}$ summary stability. Third, it offers a reproducible city-scale case study in Sofia for privacy-governed environmental analytics, showing how district reporting can be supported through a privacy-preserving cohort release and consumer-side spatial linkage.

2. Related Work

Related work relevant to this study spans three threads that are often treated separately: release-oriented privacy protection for microdata, governance-compatible data sharing across organisations, and environmental analytics using heterogeneous sensing infrastructures. The present case sits at their intersection. Unlike work that focuses only on legal framing or only on modelling outcomes, our focus is practical: how a municipality can release a constrained cohort dataset that remains usable for district-level reporting in a cross-institutional workflow.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W2-2026-1-2026

In privacy-preserving microdata publishing, k-anonymity remains influential because it is interpretable, auditable, and easy to explain (Sweeney, 2002). Its weaknesses, including attribute disclosure in homogeneous groups, motivated extensions such as ℓ-diversity and t-closeness (Machanavajjhala et al., 2006, Li et al., 2007). Differential privacy offers stronger formal guarantees, particularly for interactive query systems, but often requires mechanism and workload choices that are less aligned with periodic municipal releases (Dwork, 2006). For this reason, many public-sector settings still rely on release-time transformations and minimum-group-size rules as a practical control layer. Our work follows that line by using k-anonymity as a declared release constraint and then quantifying its analytical consequences rather than treating compliance as sufficient evidence of utility.

In parallel, urban data-sharing research highlights that city data ecosystems are institutionally fragmented and difficult to integrate through full centralisation. Dataspace-oriented approaches and IDS-inspired architectures therefore emphasise interoperable exchange under explicit governance conditions (Halevy et al., 2006, Peeters et al., 2024, Otto and Jarke, 2019, Firdausy et al., 2022, Dam et al., 2023). This literature clarifies roles, policies, and connector-based interoperability, but less often reports how privacy transformations in one organisation propagate to downstream statistical outputs in another. The present paper complements this gap by analysing data-plane behaviour once sharing permissions are assumed to be in place, with measurable effects on retention, subgroup coverage, and district representation.

Finally, citizen-sensing and environmental-health studies show that particulate-matter exposure can be analysed at urban scale when population and monitoring data are combined, and that such linkage can inform health-relevant planning (Gauderman et al., 2004, Mueller et al., 2020). At the same time, citizen-sensing networks require curation because station availability and data quality vary over space and time. Prior analyses commonly assume that either detailed population locations are available or geocoding succeeds sufficiently for robust linkage. In municipal practice, however, address strings are often inconsistent and sharing restrictions prohibit external access to person-level location data (Bernardo et al., 2024). Our contribution is therefore not a new exposure model; it is an end-to-end, privacy-aware integration pattern that evaluates whether district-level environmental summaries remain stable as disclosure protection becomes stricter.

3. Methodology

3.1 Problem setting and use case

The problem addressed in this paper is common in municipal analytics: urban decisions often require combining datasets that are institutionally fragmented and unevenly shareable. Municipalities maintain administrative records that are highly informative for population-level analysis, while environmental conditions are observed through separate sensing infrastructures operated by research organisations, public platforms, or citizen-sensing communities. Bringing these sources together can support district-level monitoring, reporting, and planning, but it also raises privacy and governance challenges when one dataset contains person-level information (Peeters et al., 2024, Bernardo et al., 2024).

The aim is to produce monthly district-level ambient PM proxy indicators for the enrolled-child cohort, together with retained-cohort and subgroup-coverage measures. The privacy problem is not that ambient air quality cannot be computed without personal data; it is that cohort-specific reporting requires knowing which children are in scope and how the release represents them across districts, ages, and kindergartens. This creates the central integration challenge addressed in this paper: how can cross-institutional urban analytics be supported when cohort composition is needed for reporting but raw enrolment records remain non-shareable?

A second operational challenge concerns spatial linkage. In the source setting, addresses are operational administrative strings rather than validated coordinates. If precise address-level geocoding is treated as a prerequisite, unresolved or ambiguous addresses can reduce coverage and potentially distort outputs. For district-scale reporting, a more robust alternative is to separate privacy-sensitive internal processing from consumer-side analytics and use public spatial proxies where appropriate.

To address these constraints, the paper adopts a privacy-preserving workflow for cross-institutional district reporting. The stakeholder roles, data inputs, and pipeline steps are specified in the next subsection, and the workflow is evaluated in a city-scale case study in Sofia.

3.2 Data and pipeline workflow

The workflow involves three stakeholders with distinct data responsibilities. First, **Municipality** acts as a data provider for the administrative cohort data. It holds the original kindergarten enrolment records and is responsible for transforming them into a privacy-preserving release that can be shared for downstream analysis.

Second, the **Environmental monitoring platform** acts as a data provider for the air-quality data product. It curates sensor metadata and historical measurements originating from a city-wide citizen-sensing ecosystem in Sofia and publishes cleaned daily station-level aggregates for reuse.

Third, a **Healthcare authority** acts as a data consumer of the shared datasets. This stakeholder does not access raw enrolment records, identifiers, or address strings. Instead, it computes monthly ambient PM proxy indicators and district-level summaries using only the released cohort, the curated air-quality dataset, and public administrative boundary data.

The workflow comprises three connected pipelines. **Pipeline 1** (municipality-side privacy processing) ingests raw enrolment data, removes direct identifiers from the release, coarsens selected attributes, and enforces the k-anonymity release rule. **Pipeline 2** (air-quality curation) validates sensor observations and metadata, then produces daily station-level pollutant aggregates. **Pipeline 3** (consumer-side analytics) uses only provider-approved inputs to compute monthly ambient PM proxy assignments and district-level summaries.

This pipeline architecture separates provider-controlled privacy processing from consumer-side urban analytics. In dataspace terms, it assumes that access conditions and sharing permissions are handled in a control plane, while the actual datasets and derived outputs are exchanged through a data plane (Peeters et al., 2024). This paper focuses on the behaviour of the data plane once such an agreement is in place.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W2-2026-1-2026

The workflow integrates three data inputs that align with the pipeline responsibilities: municipal enrolment records, a curated citizen-sensing air-quality product, and public administrative boundaries used for district-level spatial reference.

3.2.1 Municipal kindergarten enrolment data The administrative source used in the case study is a municipality-provided kindergarten enrolment dataset for Sofia Municipality. In raw form, the dataset contains a direct child identifier, year of birth, kindergarten assignment, district label, and free-text residential address information. The address field is an operational administrative string used in municipal records, not a validated coordinate or standardised geocoding product. From an analytics perspective, these records are valuable because they define the enrolled-child cohort for which environmental reporting is required. From a privacy perspective, they are highly sensitive because they combine direct identifiers with quasi-identifying attributes and detailed location strings. The geocoding diagnostic reported later therefore characterises the evaluated sample and protocol, not the quality of all Sofia addresses.

In the proposed workflow, the raw municipal data remains under municipality control. It is used only within Pipeline 1 to construct a privacy-preserving cohort release. The downstream consumer never accesses raw identifiers or address strings.

3.2.2 Air-quality data The environmental source is derived from a city-wide citizen-sensing ecosystem in Sofia. These data provide broad spatial coverage across the city through a network of distributed monitoring stations, but station activity and measurement availability vary over time. For monthly reporting, the raw observations are therefore curated into a more stable air-quality product.

Pipeline 2 produces two core outputs for downstream use: a sensor registry containing candidate station identifiers, coordinates, pollutants, and monthly activity flags, and a daily aggregates table containing station-level $PM_{2.5}$ and PM_{10} summaries. Candidate stations are those in the Sofia registry with valid coordinates and at least one accepted PM observation in the January 2025 reporting period. Raw readings are first screened for missing timestamps, missing coordinates, and invalid negative pollutant values. Daily values are arithmetic means of accepted observations by station, pollutant, and date; monthly values are arithmetic means of valid daily station aggregates. A station is treated as active for a pollutant in the reporting month when it has at least one valid daily aggregate for that pollutant. These outputs provide the environmental input for monthly district-scale reporting. Both $PM_{2.5}$ and PM_{10} are curated by the same pipeline; $PM_{2.5}$ is used as the primary pollutant in the privacy–utility evaluation reported below.

3.2.3 Public administrative boundaries A third input to the overall workflow is a public administrative boundary dataset for Sofia Municipality. In this paper, "district" refers to one of Sofia Municipality's 24 municipal administrative districts, which are used both in the source cohort and in the reporting output. The district is the finest geography retained in the released cohort; it is not claimed to be the finest administrative unit that exists in the city. The boundary dataset version used in the workflow is the public municipal district polygon layer available at the time of the January 2025 run. Geometric centroids were computed for all district polygons and stored in WGS 84 coordinates for station selection; no correction to interior representative points was applied. Because district labels are already present in the released cohort, these points can be used for proxy assignment without requiring access to child-level coordinates or raw addresses.

3.3 Privacy-preserving cohort release

The municipality-side pipeline produces a bounded cohort release for downstream analytics. The core idea is to share only the cohort fields needed for the reporting task while excluding attributes that directly reveal identities or precise residential locations.

The release process performs four key operations. First, direct identifiers are removed from the shared dataset. Second, free-text address fields are excluded from the release. Third, no stable person-level pseudonym is required for the one-month district-level analysis reported here; released rows contain only the attributes needed to compute cohort counts and subgroup summaries. Fourth, the released cohort is filtered using k-anonymity with respect to the quasi-identifier set

$$QI = (\texttt{district}, \texttt{age\_range}, \texttt{kindergarten}).$$

The variable `age_range` is derived from year of birth using a coarse categorisation appropriate for the reporting task. Records belonging to groups smaller than the threshold k are suppressed from the released cohort. If future longitudinal workflows require a record key, any pseudonym should be project-specific, used only inside the approved analytical workflow, and not treated as anonymisation or as part of the k-anonymity guarantee.

This deliberately limits the scope of the privacy claim. The release is evaluated only with respect to the declared quasi-identifiers, and the paper reports group-size statistics and descriptive risk proxies on that basis. Table 1 gives a minimal synthetic example of the mechanism. The goal is a transparent release mechanism compatible with reproducible downstream analytics, not a complete solution to all possible disclosure risks.

Record	District	Age range	Kdg.	Class size	Action
P1	A	5–6	K1	3	keep
P2	A	5–6	K1	3	keep
P3	A	5–6	K1	3	keep
P4	B	6–7	K2	2	suppress
P5	B	6–7	K2	2	suppress

Table 1: Synthetic illustration of equivalence-class suppression at $k = 3$; Kdg. denotes kindergarten.

In the example, district, age range, and kindergarten define one equivalence class of size three and one of size two. At $k = 3$, only the first class is released. The case study uses the same rule with the actual threshold $k = 12$ and the full Sofia cohort.

3.4 Spatial linkage and exposure assignment

A key design choice in the workflow is the separation between provider-side handling of sensitive location information and consumer-side assignment of exposure. The consumer does not receive address strings or child-level coordinates. Instead, spatial linkage is performed using the district label contained in the released cohort and public district centroids derived from the administrative boundary dataset.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W2-2026-1-2026

For each district centroid, the consumer-side pipeline identifies the nearest station with valid coordinates in the curated sensor registry using great-circle distance. The nearest station is pollutant-specific: if the initially selected station has no valid $PM_{2.5}$ daily aggregate in January 2025, the assignment is updated to the nearest station active for $PM_{2.5}$ in that month; the same rule is applied to PM_{10}. Missing pollutant observations are not imputed. Monthly PM proxy indicators are computed from the valid daily pollutant aggregates associated with the selected station and joined to the retained cohort through the reporting geography. District-level summaries are then computed from the retained release for each k, so their stability is evaluated empirically rather than assumed from the spatial-linkage rule alone.

This centroid-based strategy is aligned with the intended reporting scale. The target is district-level urban environmental analytics rather than household-level exposure modelling. The resulting indicator is therefore an ambient PM proxy for cohort reporting and should not be interpreted as an estimate of personal or household exposure. District centroids trade fine spatial precision for privacy compatibility, reproducibility, and robustness, while reducing dependence on external geocoding, whose completeness and quality may be insufficient for reliable administrative linkage.

3.5 Reproducibility and data plane implementation

The implementation uses a persistent DuckDB data plane to store provider datasets and derived outputs across the three pipelines. This supports reproducibility by allowing the released cohort, curated air-quality product, and monthly reporting outputs to be regenerated from a common analytical snapshot. It also makes the privacy–utility evaluation auditable, because the same stored outputs can be reused to compute release statistics, retention values, subgroup summaries, and district-level comparisons.

This reproducibility matters because urban analytics workflows are often reused across reporting cycles, months, and stakeholder requests. A persistent data plane makes it easier to inspect what was released, what was computed, and how changes in privacy settings affect city-level indicators. In that sense, the case study is not only an environmental analytics exercise but also an example of reproducible, privacy-aware urban data sharing in practice.

4. Evaluation Design

4.1 Evaluation question and setting

The evaluation is designed to address a cohort-reporting question: *after replacing raw enrolment records with a privacy-preserving release, how much cohort coverage and district-level reporting utility remain for monthly particulate-matter reporting?* To answer it, the paper evaluates the release from four complementary perspectives: privacy compliance under the declared quasi-identifiers, retention of the released cohort under alternative anonymity thresholds, represented-district coverage and district-level environmental summaries, and diagnostics for subgroup coverage and spatial linkage robustness.

The analysis is conducted for monthly reporting for January 2025, using a fixed city cohort and a fixed environmental data product for the same month. The main operating point is $k = 12$, which is the threshold used in the case study release. To quantify the privacy–utility trade-off more broadly, the same release-and-evaluation procedure is repeated for

$$k \in \{5, 8, 12, 16, 20\}.$$

All reported metrics are computed from the reproducible data-plane outputs generated by the workflow.

The evaluation is intentionally aligned with district-level reporting objectives. It does not assess household-level exposure precision or clinical outcomes. Instead, it focuses on whether a privacy-preserving municipal release can still support district-scale environmental analytics in a reproducible and robust manner. In this paper, high utility means sufficient cohort retention, represented-district coverage, and stable district-level summaries under privacy constraints, not preservation of all fine-grained spatial detail.

4.2 Privacy metrics

Privacy is evaluated on the released cohort with respect to the declared quasi-identifier set

$$QI = (\texttt{district}, \texttt{age\_range}, \texttt{kindergarten}).$$

Under this definition, each released record belongs to an equivalence class formed by identical values of the three quasi-identifiers. For each operating point and alternative k value, the evaluation reports the number of released records, the number of equivalence classes, and descriptive statistics of class sizes, including the minimum, median, mean, and maximum observed group sizes.

To make the release threshold more interpretable, the paper also reports a simple prosecutor-style risk proxy based on observed group size, approximated as $1/\text{group size}$. When the release satisfies k-anonymity under the declared quasi-identifiers, the maximum value of this proxy is bounded by $1/k$. This quantity is used solely as a simple descriptive indicator derived from the released data; it is not intended as a comprehensive adversarial risk model.

4.3 Utility metrics for downstream analytics

The utility evaluation focuses on the intended downstream use of the released cohort: district-level environmental reporting. Two types of utility measures are used.

First, *retention* is defined as the fraction of records that remain in the released cohort after privacy filtering relative to a fixed baseline cohort. The baseline is the deduplicated municipal cohort restricted to records with non-missing quasi-identifier values, so that variation across k reflects only the effect of the disclosure-control rule rather than unrelated cleaning steps. Formally,

$$\text{Retention}(k) = \frac{n_k}{n_{\text{baseline}}}, \tag{1}$$

where n_k is the number of records retained under threshold k.

Second, the evaluation measures whether district-level environmental summaries remain stable as the privacy threshold becomes stricter. The main summary output is the district mean $PM_{2.5}$ value computed from the consumer-side monthly PM proxy assignments. $PM_{2.5}$ is used as the primary pollutant for the privacy–utility comparison; PM_{10} is curated by the same

pipeline and follows the same station-selection and monthly aggregation logic. Stability is assessed relative to the $k = 5$ release using two complementary measures: Spearman rank correlation, which captures whether districts remain similarly ordered from lower to higher $PM_{2.5}$, and mean absolute error (MAE), which captures the magnitude of change in district mean $PM_{2.5}$ values.

Because stronger privacy filtering can suppress all cohort records in some districts, utility metrics are computed on the overlapping district set represented in both the $k = 5$ reference release and the threshold-k release. The overlap size is reported alongside the stability metrics, so that high agreement is not interpreted independently of lost district coverage. The $k = 5$ release is the lowest-threshold comparator available in the shared evaluation, not a raw-data ground truth. A complete unsuppressed comparator is left for future internal aggregate validation, where it can be computed without disclosing child-level records.

4.4 Subgroup and spatial diagnostics

Overall retention can conceal uneven effects across the city. To make these representativeness issues visible, the evaluation also reports subgroup retention by district and by age range. These diagnostics show whether stronger privacy thresholds disproportionately reduce coverage for particular districts or demographic subgroups. When discussing the lowest-retention districts, attention is restricted to districts with sufficiently large baseline counts in order to avoid unstable percentages driven by very small denominators.

A final diagnostic addresses the robustness of the workflow's spatial linkage strategy. Since the consumer pipeline relies on district centroids rather than address-level geocoding, the evaluation includes a separate internal comparison between centroid-based locations and third-party geocoded points derived from sampled administrative addresses. This diagnostic is reported only in aggregated form and serves two purposes: it quantifies the spatial distortion introduced by centroid-based linkage and measures the success rate of external geocoding for the sampled administrative addresses. Together, these quantities help interpret the operational trade-off between fine spatial precision and reproducible, privacy-compatible city reporting.

5. Results

This section reports the empirical behaviour of the proposed workflow for the January 2025 city-scale case study and evaluates whether the privacy-preserving release remains usable for district-level environmental analytics. The results answer five questions: how the released cohort is constructed from the municipal source, whether the release satisfies the declared privacy rule at the operating point, how retention, represented-district coverage, and district-level $PM_{2.5}$ stability change across alternative k thresholds, which subgroups are most affected by filtering, and how robust the centroid-based spatial linkage strategy is in practice.

5.1 Released cohort construction

The raw municipal enrolment file is transformed into the released cohort used by the downstream consumer. The raw dataset contained 4,003 rows. After initial cleaning, 3 records with invalid identifiers were removed, leaving 4,000 valid rows. Deduplication was then enforced to produce a fixed baseline cohort of 3,681 records, one row per child. This deduplicated cohort serves as the reference population for all subsequent retention calculations. Applying the privacy filter at the operating point $k = 12$ suppressed 241 additional records that belonged to equivalence classes smaller than the threshold. The resulting released cohort therefore contains 3,440 children. Table 2 summarises the construction of the released cohort.

This illustrates the effect of the provider-side release pipeline. The consumer does not operate on the full municipal dataset, but on a reduced, privacy-filtered cohort whose size is determined by source data quality, deduplication, and the anonymity threshold. In reporting terms, this released cohort defines the population that can be included in downstream district-level analytics without exposing raw personal records.

Step	Rows	Change
Raw municipal input	4,003	n/a
After identifier cleaning	4,000	-3
Deduplicated baseline cohort	3,681	-319
Released cohort at $k = 12$	3,440	-241

Table 2: Cohort construction and release size for the evaluated Sofia case.

5.2 Privacy compliance at the operating point

The released cohort is evaluated to check whether it satisfies the declared privacy rule. Privacy is assessed with respect to the quasi-identifier set

$$QI = (\texttt{district}, \texttt{age\_range}, \texttt{kindergarten}).$$

Under this definition, the 3,440 released records are partitioned into 145 observed equivalence classes. The minimum observed class size is 12, which matches the configured release threshold, and there are no violating groups in the released cohort. Group sizes are not concentrated entirely at the lower bound: the median class size is 19, the mean is 23.72, and the largest class contains 93 records.

For interpretability, a simple prosecutor-style risk proxy is computed as the reciprocal of the observed class size. Under this proxy, the maximum observed value in the released cohort is $1/12 \approx 0.083$. The record-weighted mean of this reciprocal class-size proxy is $145/3440 \approx 0.042$, while the unweighted mean across the 145 observed equivalence classes is 0.053. These quantities do not constitute a full adversarial disclosure model, but they provide a transparent description of the release under the declared quasi-identifiers.

This confirms that the provider release pipeline produces an auditable cohort release satisfying the stated privacy rule at the chosen operating point. Downstream district-level analytics can therefore be based on measurable group-size protection rather than informal minimisation alone.

5.3 Privacy–utility trade-offs across k

The main experiment of the study quantifies how reporting utility changes as the anonymity threshold increases. Table 3 reports the released cohort size, retention relative to the fixed deduplicated baseline, overlap in represented districts, and

https://doi.org/10.5194/isprs-annals-XII-4-W2-2026-1-2026

district-level $PM_{2.5}$ stability relative to the $k = 5$ reference release.

Retention declines monotonically as k increases. At $k = 5$, 3,590 children remain in the released cohort, corresponding to a retention of 97.5% relative to the baseline. At the operating point $k = 12$, 3,440 children remain, corresponding to 93.5% retention. At $k = 20$, the released cohort decreases to 3,212 children, or 87.3% retention. This makes the cost of stronger privacy thresholds directly visible: more restrictive anonymity requirements suppress a larger share of the municipal cohort.

District-level environmental summaries remain comparatively stable under this stronger filtering. Using district mean $PM_{2.5}$ as the main downstream reporting quantity, Spearman rank correlation relative to the $k = 5$ reference is at least 0.988 for all evaluated thresholds $k \geq 8$ on the overlapping district set. MAE in district mean $PM_{2.5}$ remains below 0.3 μg/m^3 for all reported settings. These results indicate that, for January 2025, the district ordering and magnitudes of the summary output change only slightly as more records are suppressed.

This stability must be read alongside district overlap. The comparison includes 24 districts at $k \in \{5, 8\}$, 23 at $k \in \{12, 16\}$, and 22 at $k = 20$. The results therefore support a qualified conclusion: privacy filtering reduces cohort size and can remove some districts from the release, but among the districts that remain represented, district-level $PM_{2.5}$ summaries are highly stable in this case study.

k	n	**Retention**	$\lvert\mathcal{D}_k\rvert$	**Spearman** ρ	**MAE**
5	3,590	0.975	24	1.000	0.000
8	3,534	0.960	24	0.999	0.169
12	3,440	0.935	23	0.990	0.261
16	3,282	0.892	23	0.988	0.280
20	3,212	0.873	22	0.998	0.070

Table 3: Privacy–utility trade-offs across anonymity thresholds. Retention is relative to the fixed deduplicated baseline cohort ($n = 3{,}681$). District $PM_{2.5}$ stability is computed relative to the $k = 5$ release on the overlapping district set $\mathcal{D}_k$; MAE is reported in μg/m^3.

These results allow privacy settings to be interpreted through measurable reporting consequences. Stakeholders can see how many records remain available, how many districts remain represented, how strongly district-level $PM_{2.5}$ summaries change, and which subgroups lose coverage under stricter thresholds.

5.4 Subgroup coverage effects

Overall retention does not fully capture the release's impact on the city. To assess representativeness more explicitly, subgroup retention was examined by district and by age range. The strongest variation appears across districts. At $k = 12$, district-level retention has a median of 0.928, but the lower tail is substantially weaker, with a 5th percentile of 0.623. By contrast, age-range retention is more concentrated at the same threshold, with a median of 0.958 and a 5th percentile of 0.813.

The lowest-retention districts at $k = 12$ are shown in Table 4, restricting attention to districts with baseline counts of at least 30 in order to avoid unstable percentages driven by very small denominators. Bankya is fully suppressed at the operating point, while Pancharevo, Novi Iskar, Kremikovtsi, and Ilinden all show noticeably reduced retention. This means that the privacy filter does not simply reduce the cohort uniformly across the city; it also changes which districts remain visible in downstream reporting.

District	**Baseline** n	**Released** n	**Retention**
Bankya	32	0	0.000
Pancharevo	52	32	0.615
Novi Iskar	57	38	0.667
Kremikovtsi	51	35	0.686
Ilinden	86	60	0.698

Table 4: Lowest district retention at $k = 12$ among districts with baseline count $n \geq 30$.

This result is important because district-level reporting is interpreted both geographically and statistically. A privacy configuration that preserves summary stability on represented districts may still disproportionately reduce visibility for smaller or less represented districts. Retention by subgroup should therefore be considered alongside headline privacy–utility curves when selecting a release threshold.

5.5 Spatial utility and geocoding robustness

For the spatial-linkage diagnostic, the consumer pipeline uses district centroids instead of address geocoding. A separate internal diagnostic compared centroid-based locations with third-party geocoded points obtained from a simple random sample of 1,000 deduplicated complete-case administrative records with non-empty address strings. Address strings were trimmed, obvious repeated whitespace was removed, and Sofia/Bulgaria context was appended before geocoding; no manual correction of individual addresses was performed. A geocoding attempt was counted as successful only when the service returned a point that could be interpreted as being in Sofia Municipality. Distances were computed between the district centroid and returned point after projecting coordinates to a metric coordinate reference system for distance calculation.

Only 459 of the 1,000 sampled addresses were successfully geocoded, corresponding to a success rate of 45.9% for this sample and protocol. This should not be read as a general property of all Sofia addresses. Rather, it indicates that relying on external geocoding for this dataset would leave many sampled records unresolved before privacy filtering. On the successfully geocoded subset, the median distance between the centroid-based location and the geocoded point is 4.06 km, the mean distance is 5.75 km, and the 95th percentile is 14.44 km. These values show the expected coverage-versus-precision trade-off: centroid assignment gives deterministic district coverage but should not be interpreted as household-level exposure.

These results support public administrative centroids as a compromise between precision and robustness. The workflow does not aim to preserve household-level location accuracy or support individual exposure inference. Instead, it prioritises reproducible district-level linkage that is privacy-compatible under the stated release constraints when raw municipal addresses cannot be shared and geocoding may be incomplete under the available protocol.

6. Discussion

The case study shows that district-scale environmental reporting can continue when sensitive municipal data are shared only

through a bounded release. Here, measurable consequences mean retained records, represented districts, subgroup coverage, equivalence-class sizes, and any distortion in released reporting outputs. These quantities complement, rather than replace, legal and organisational compliance checks: they make visible what a selected privacy setting does to the intended analytical task.

Anonymity thresholds behave as practical design parameters. Increasing k improves the minimum observed equivalence-class size but reduces the number of retained records and represented districts. In this study, district mean $PM_{2.5}$ summaries remain highly stable on overlapping districts, with only small changes in rank order and magnitude across the evaluated thresholds. This stability should not be interpreted as preservation of complete reporting utility: stricter thresholds still reduce cohort size, suppress some equivalence classes, and can remove districts or subgroups from the released cohort.

The approach is transferable when three conditions hold: the sensitive provider can prepare the release internally, the downstream task can be expressed at an agreed reporting geography, and public spatial proxies plus curated environmental data are sufficient for that geography. The numerical results themselves are less transferable. Retention, suppressed districts, geocoding success, and station assignment depend on local cohort composition, address quality, administrative geography, sensor density, and the chosen k threshold.

Some limitations follow from these design choices. The study covers one city and one reporting month, so seasonal patterns and different cohort compositions remain to be tested. The privacy claim is limited to k-anonymity over the declared quasi-identifiers and does not provide differential-privacy guarantees or protection against all background-knowledge attacks. Centroid linkage supports district reporting but should not be interpreted as household-level exposure modelling. The geocoding diagnostic is based on one sampled address set and protocol, so its aggregate results should be interpreted as protocol-specific rather than as general address-quality estimates. Finally, the shared evaluation uses $k = 5$ as its lowest-threshold reference rather than a complete unsuppressed cohort. Future work will extend the evaluation with an internal aggregate comparison between the deduplicated complete-case cohort and the $k = 12$ release without disclosing child-level records. Such a comparison would quantify the effect of suppression on child-weighted citywide PM proxies, represented-district coverage, and age-group summaries. Extending the analysis across multiple months will also help determine whether the observed privacy–utility behaviour remains stable under seasonal changes in cohort composition and sensor availability.

7. Conclusions

This paper presented a privacy-preserving data-sharing workflow for urban environmental analytics in which sensitive municipal administrative records are not transferred to downstream analysts in raw form. Instead, the workflow combines a municipality-produced cohort release, a curated citizen-sensing air-quality product, and a consumer-side analytics pipeline that generates district-level monthly indicators from approved inputs only.

The city-scale case study shows that this design can support urban reporting under explicit disclosure control. At the evaluated operating point, the January 2025 release satisfies the declared k-anonymity rule for district, age range, and kindergarten, while retaining 3,440 children from a deduplicated baseline of 3,681. Across alternative anonymity thresholds, stronger privacy protection reduces overall cohort retention and district coverage. At the same time, district mean $PM_{2.5}$ summaries remain highly stable on the overlapping district set, with Spearman $\rho \geq 0.988$ for $k \geq 8$ and MAE below 0.3 $\mu g/m^3$. The spatial diagnostic further shows that centroid-based linkage sacrifices fine-grained precision but avoids dependence on incomplete address-level geocoding for district reporting.

Together, these results suggest that privacy settings for municipal analytics can be selected and justified using measurable privacy–utility consequences alongside legal and organisational compliance requirements. Cohort retention, subgroup coverage, district representation, equivalence-class sizes, and output stability provide a practical basis for evaluating whether a release remains fit for its intended reporting purpose.

Acknowledgements

This research was supported by the GATE project, funded by the Horizon 2020 WIDESPREAD-2018–2020 TEAMING Phase 2 programme under Grant Agreement No. 857155; the programme "Research, Innovation and Digitalisation for Smart Transformation" 2021–2027 (PRIDST), under Grant Agreement No. BG16RFPR002-1.014-0010-C01; the DS4SSCC–DEP project, funded by the European Union's Digital Europe Work Programme 2021–2022 under Grant Agreement No. 101123342; and the DataPACT project, funded by the Horizon Europe programme under Grant Agreement No. 101189771.

References

Bernardo, B. M. V., Mamede, H. S., Barroso, J. M. P., dos Santos, V. M. P. D., 2024. Data governance & quality management—Innovation and breakthroughs across different fields. *Journal of Innovation & Knowledge*, 9(4), 100598.

Dam, T., Klausner, L. D., Neumaier, S., Priebe, T., 2023. A survey of dataspace connector implementations. *Proceedings of the 2nd Italian Conference on Big Data and Data Science (ITADATA 2023)*, CEUR Workshop Proceedings, 3606, CEUR-WS.org, Naples, Italy, 1–12. Available at: `https://ceur-ws.org/Vol-3606/paper41.pdf`.

Dwork, C., 2006. Differential privacy. *International Colloquium on Automata, Languages and Programming (ICALP)*.

Firdausy, D., De Alencar Silva, P., van Sinderen, M. J., Iacob, M. E., 2022. Towards a reference enterprise architecture to enforce digital sovereignty in international data spaces. *24th IEEE International Conference on Business Informatics (CBI)*, 117–125.

Gauderman, W. J., Avol, E., Gilliland, F., Vora, H., Thomas, D., Berhane, K., McConnell, R., Kuenzli, N., Lurmann, F., Rappaport, E., Margolis, H., Bates, D., Peters, J., 2004. The Effect of Air Pollution on Lung Development from 10 to 18 Years of Age. *New England Journal of Medicine*, 351(11), 1057–1067.

Halevy, A., Franklin, M., Maier, D., 2006. Principles of dataspace systems. *Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 1–9.

Li, N., Li, T., Venkatasubramanian, S., 2007. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*, 106–115.

Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M., 2006. ℓ-diversity: Privacy beyond k-anonymity. *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 24.

Mueller, N., Rojas-Rueda, D., Khreis, H., Cirach, M., Andrés, D., Ballester, J., Bartoll, X., Daher, C., Deluca, A., Echave, C., Milà, C., Márquez, S., Palou, J., Pérez, K., Tonne, C., Stevenson, M., Rueda, S., Nieuwenhuijsen, M., 2020. Changing the Urban Design of Cities for Health: The Superblock Model. *Environment International*, 134, 105132.

Otto, B., Jarke, M., 2019. Designing a Multi-Sided Data Platform: Findings from the International Data Spaces Case. *Electronic Markets*, 29, 561–580.

Peeters, B., Pezuela, C., Stolwijk, C., Alonso, D., Curry, E., Berkers, F., Lamerichs, G., Korhonen, H., Suksi, J., Martin, M., Frigeri, M., Musidlowska, M., Razzaq, M. A., Haque, R., Rogotis, S., Castellvi, S., Guggenberger, T., 2024. Data space design principles. Interim report, Data Spaces Support Centre (DSSC). Interim report reflecting Version 2.0 of the design principles; part of the DSSC project (Grant Agreement No. 101083412).

Sweeney, L., 2002. K-Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 557–570.