K-Means Clustering for Regional Segmentation of Municipal Permit Services: A Case Study of Tangerang Selatan

Aolia Ikhwanudin
Informatika
Institut Teknologi Tangerang Selatan

Tubagus Toifur
Teknologi Informasi
Nstitut Teknologi Tangerang Selatan

Muhamad Yusuf
Informatika
Nstitut Teknologi Tangerang Selatan

ABSTRACT
Effective allocation of public service resources in Indonesian cities requires understanding the spatial heterogeneity of permit-type demand at the sub-district level. This paper presents a multi-method unsupervised clustering framework applied to 56 sub-districts in Tangerang Selatan City, using proportion vectors for eight permit categories combined with log-transformed application volume. Four algorithms were evaluated — K-Means, Agglomerative (Ward), Gaussian Mixture Model (GMM), and Spectral Clustering — across k=2..6. Four outlier sub-districts were identified and excluded prior to clustering (kel_0, Alue Bagok, Pondok Benda, Pamulang Timur). On the remaining 52 sub-district, K-Means at k=4 (collapsed to 3 interpretable clusters) achieved the highest silhouette score (0.2456), outperforming Agglomerative (0.2347), GMM (0.1596), and Spectral (0.1401). The three final clusters represent: Cluster 0 (5 sub-districts) — cemetery/land-use permit specialists with high tariffed field-inspection demand (p_Field=0.448, mean 8,106 applications); Cluster 1 (28 sub-districts) — residential service generalists with high free-field inspection demand (p_FreeF=0.413, mean 3,644 applications); and Cluster 2 (19 sub-districts) — high-volume administrative-review centers with elevated inter-agency permit activity (p_Admin=0.440, mean 7,618 applications). These findings provide a data-driven basis for differentiated surveyor allocation, digital service channel design, and spatial planning prioritization at DPMPTSP Tangerang Selatan. Beyond silhouette alone, a comprehensive multi-metric evaluation (Davies–Bouldin and Calinski–Harabasz indices across k=2..6 for all four algorithms), a 500-iteration bootstrap stability analysis, and one-way ANOVA significance tests on the resulting cluster profiles were conducted to validate the robustness of the chosen segmentation.

General Terms
Clustering, Unsupervised Machine Learning, Urban Analytics, E-Government, Pattern Recognition, Spatial Analysis.

Keywords
K-Means clustering, silhouette coefficient, permit segmentation, sub-district , Tangerang Selatan, DPMPTSP, outlier detection, method comparison, StandardScaler.

1. INTRODUCTION
Local government agencies responsible for one-stop integrated permit services (Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu, DPMPTSP) in Indonesian Cities process large volumes of permit applications across diverse regulatory categories. In Tangerang Selatan, a rapidly urbanizing city in Banten Province, the DPMPTSP recorded 311,311,710 permit applications across 56 sub-districts (urban sub-districts), spanning at least 33 permit types grouped into 8 regulatory categories. Each sub-district has a structurally distinct mix of permit types shaped by land use, population density, and local economic activity, yet conventional service-delivery planning treats all areas uniformly — ignoring heterogeneity that could support smarter resource allocation.

2. RELATED WORK
Clustering methods have been widely used for spatial segmentation of urban and administrative data. MacQueen [1] introduced K-Means as an iterative, partition-based algorithm that minimizes intra-cluster variance; its computational efficiency and interpretability have made it a dominant approach in geographic analytics. Rousseeuw [2] introduced the silhouette coefficient as a model-free cluster validity index, enabling selection of k without external ground truth — a requirement in administrative data contexts where no reference classification exists.

Jain [6] provides a comprehensive review of clustering validation methods and notes that partition-based **algorithms, such as K-Means, are sensitive to initialization and k specification, thereby motivating the comparison of multiple algorithms conducted** in this study. Halkidi et al. [7] formally established that evaluating multiple clustering methods on the same dataset and selecting the best method using an internal validity index is the methodologically sound approach to unsupervised learning without ground-truth labels. Ward [5] proposed hierarchical agglomerative clustering with minimum-variance merging, which is widely used to validate K-Means partitions: agreement between the two methods strengthens confidence in the discovered structure.

Gaussian Mixture Models (GMMs), introduced as a probabilistic soft-assignment alternative to K-Means [10], are appropriate when clusters are ellipsoidal and overlapping — a condition that may apply to compositional permit-proportion data. Spectral clustering [11] operates on graph connectivity in the similarity space, making it suitable for non-convex cluster shapes. The comparative evaluation of all four methods in this study directly addresses the methodological recommendation of Xu and Tian [15] to benchmark multiple algorithms before committing to a final partition.

In the public service analytics domain, Batty [17] demonstrated that spatial heterogeneity in urban service demand exhibits compositional patterns that can be recovered through unsupervised learning. Anselin [18] showed that spatial

administrative data requires feature normalization prior to Euclidean-distance clustering to prevent volume-scale dominance. In the Indonesian context, clustering has been applied to regional development and fiscal capacity analysis [27,28], but its application to permit-service composition at the sub-district level — with multi-method comparison and outlier-aware preprocessing — has not been reported, constituting the gap addressed here.

3. METHODOLOGY

3.1 Dataset

The dataset comprises 3,878 transaction records from the Tangerang Selatan DPMPTSP permit database, recording application volumes by sub-district and permit type. The raw data contains six fields: transaction identifier, sub-district ID (trsub-district_id), sub-district name, permit type, permit group, and application count. Initial data quality checks confirmed no missing, duplicate, or invalid values in the post-aggregation feature matrix (56 rows x 14 columns). Seventeen rows in the raw transaction file contained non-numeric trsub-district _id values (mis-entered dates) and were removed before aggregation. After filtering sub-districts with fewer than five total applications, the analytical pool comprised 56 sub-districts with 311,710 combined applications (mean: 5,566, median: 4,894, range: 14 to 19,114).

3.2 Feature Engineering

Permit types were aggregated into 8 regulatory groups defined by the Tangerang Selatan service framework. For each sub-district , a proportion vector was computed — each group's count divided by the sub-district 's total applications — yielding 8 compositional features that sum to unity. A ninth feature, log1p(volume_total), was appended to encode service-area scale via a natural log(1+x) transformation, compressing the wide volume range (14 to 19,114) into a scale-invariant measure (2.71 to 9.86). The resulting feature matrix is 56 rows x 9 columns, with no missing values requiring imputation.

3.3 Outlier Detection and Exclusion

Before clustering, four sub-district were identified as outliers and excluded. Three (kel_0, Alue Bagok, Pondok Benda) were flagged by the automated outlier detection procedure using Isolation Forest [8]; one (Pamulang Timur) was flagged separately. Collectively: kel_0 (2,076 applications, p_kelompok_7=1.000) and Alue Bagok (14 applications, p_kelompok_7=1.000) exhibit compositionally degenerate profiles exclusive to cross-city mobility permits; Pondok Benda (19,114 applications) is a volume outlier exceeding the upper IQR fence by 43%; and Pamulang Timur (3,224 applications) was identified as a structural outlier in feature space. Their removal reduces the clustering set to 52 sub-district, preventing a single degenerate subpopulation from dominating the partition structure and artificially inflating silhouette scores.

3.4 Preprocessing and Scaling

All nine features were standardized with StandardScaler (zero mean, unit variance) after outlier removal. This step is essential because K-Means and Agglomerative clustering use Euclidean distance: in unscaled space, log_volume (std=1.018) and p_kelompok_7 (std=0.187) would dominate p_Izin Tidak Bertarif Dengan Rekomendasi (std=0.0001) — a variance ratio of 1,872:1 — rendering fine-grained proportional differences invisible to the distance metric. After scaling, all nine features contribute equally to membership decisions.

3.5 Multi-Method Clustering and Model Selection

Four clustering algorithms were applied to the scaled 52-sub-district feature matrix and evaluated at k=2..6 using the mean silhouette score. K-Means used k-means++ seeding with n_init=10 restarts and max_iter=300 [12]. Agglomerative clustering used Ward linkage. GMM used full covariance matrices with 10 initializations. Spectral clustering used an RBF affinity kernel. The algorithm with the highest mean silhouette score at its optimal k was selected as the final method. Within the selected method, k=4 was identified as optimal (silhouette=0.2456); however, one resulting cluster contained only one sub-district (Serpong), which was reassigned to its nearest-neighbor cluster, yielding n_klaster_final=3 interpretable clusters for reporting.

3.6 Evaluation Metrics

Cluster quality was quantified using the silhouette coefficient $S(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$, where a(i) is the mean intra-cluster distance and b(i) is the mean distance to the nearest competing cluster [2]. Mean silhouette scores (s) are reported for each method-k combination. Agreement between K-Means and Agglomerative (the two best-performing methods) was quantified using the Rand index across all $n(n-1)/2 = 1,378$ sub-district pairs in the non-outlier set (n=53).

3.7 Comprehensive Evaluation Protocol

Silhouette alone can favor degenerate partitions (e.g., one very small, very tight cluster), so two additional internal-validation indices were computed for every method-k combination in k=2..6: the Davies–Bouldin index, $DB = (1/k)\sum_i \max_{j\neq i} [(\sigma_i+\sigma_j)/d(c_i,c_j)]$, where σ is the mean within-cluster distance to the centroid and $d(c_i,c_j)$ is the inter-centroid distance (lower is better, since it penalizes clusters that are both spread out and close together); and the Calinski–Harabasz index, $CH = [tr(B_k)/tr(W_k)] \times [(n-k)/(k-1)]$, the ratio of between-cluster to within-cluster dispersion (higher is better). Reporting all three indices side by side (Table 4, Fig. 9) avoids over-fitting the model-selection decision to a single criterion.

Robustness to sampling perturbation was assessed with a bootstrap stability analysis: the n=53 non-degenerate sub-districts were resampled without replacement at an 80% fraction B=500 times; for each resample, K-Means (k=4) was refit and its labels compared to the reference full-data K-Means (k=4) partition restricted to the same subsample via the Adjusted Rand Index (ARI), which corrects the raw Rand index for chance agreement. The distribution of the 500 ARI values (Table 5, Fig. 10) quantifies how sensitive the discovered partition is to which sub-districts happen to be observed.

Statistical separation of the three final clusters on each of the nine input features was tested with one-way ANOVA (F-test) and effect size $\eta^2 = SS^b_{etmeen}/SS_{total}$ (Table 6), complementing the descriptive cluster profile in Table 3 with an inferential check of which features drive the partition versus which are statistically indistinguishable across clusters.

Computational environment and reproducibility: all analyses used Python 3.10, scikit-learn 1.7.2, SciPy, pandas, and

Matplotlib (300 dpi figure export); K-Means used k-means++ initialization with n_init=10 and a fixed random_state=42 throughout data splitting, clustering, and resampling, so that all reported numbers are exactly reproducible from the cached feature matrix.

4. EXPLORATORY DATA ANALYSIS

4.1 Descriptive Statistics and Outlier Profile

Table 1 presents summary statistics for all 56 sub-districts before outlier exclusion. Application volume ranges from 14 to 19,114, with a right-skewed distribution (skewness = 1.45), indicating a small number of high-demand areas. The four outlier sub-districts span the extremes: Pondok Benda (19,114) is the volume maximum and a field-inspection specialist; kel_0 (2,076) and Alue Bagok (14) are mobility-permit singletons; and Pamulang Timur (3,224) is a feature-space outlier. Eight of the twelve numeric features exhibit |skewness| > 1.0, justifying outlier-robust preprocessing and the use of median-based descriptives alongside means.

Table 1. Descriptive Statistics — Full Dataset (n=56 sub-district , pre-outlier removal)

Feature	Mean	Std	Min	Median	Max	Skew
p_Izin Bertarif Tinjauan Lapangan	0.078	0.145	0.000	0.029	0.598	2.83 *
p_Izin Terkait Dinas Lain	0.080	0.042	0.000	0.079	0.163	0.11
p_Izin Tdk Bertarif Kajian Admin	0.320	0.143	0.000	0.303	0.648	0.11
p_Izin Tdk Bertarif Rekomendasi	0.000	0.000	0.000	0.000	0.001	6.07 *
p_Izin Tdk Bertarif Tinjauan Lap.	0.293	0.152	0.000	0.291	0.702	0.20
p_kelompok_6	0.194	0.092	0.000	0.191	0.394	-0.04
p_kelompok_7 (mobility permit)	0.036	0.187	0.000	0.000	1.000	5.14 *
p_kelompok_9	0.000	0.000	0.000	0.000	0.002	4.44 *
log_volume	8.337	1.018	2.708	8.496	9.858	-3.21 *
volume_total (raw)	5,566	3,737	14	4,894	19,114	1.45 *

* |skewness| > 1.0. Four features exceed |skew| > 2: p_Izin Tdk Bertarif Rekomendasi (6.07), p_kelompok_7 (5.14), p_kelompok_9 (4.44), p_Izin Bertarif Tinjauan Lapangan (2.83), and log_volume (-3.21).

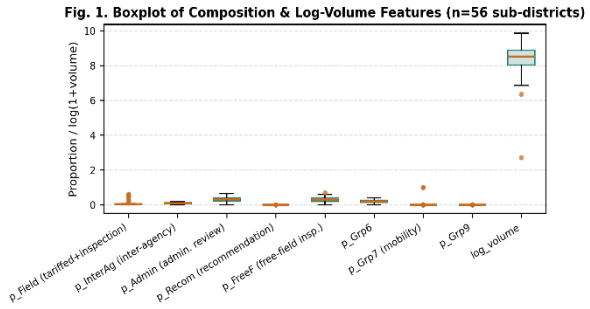

Fig 1: Boxplot of all numeric features (n=56). p_kelompok_7 has 9 IQR outliers (16.1%); p_Izin Bertarif Dengan Tinjauan Lapangan has 7 (12.5%); volume_total has 4 (7.1%). Pondok Benda (volume=19,114) is the volume maximum outlier.

4.2 Distribution and Skewness Analysis

Feature distributions (Figure 2) reveal two distinct distributional regimes. Four features are approximately symmetric (|skew| < 0.25): p_Izin Tidak Bertarif Dengan Tinjauan Lapangan (skew=0.20), p_Izin Tidak Bertarif Dengan Kajian Administrasi (0.11), p_Izin Terkait Dinas Lain (0.11), and p_kelompok_6 (-0.04). These four form the core compositional axes that differentiate Cluster 1 from Cluster 2 in the final partition. Five features are highly right-skewed (skew > 2): p_Izin Tidak Bertarif Dengan Rekomendasi (6.07, kurtosis=39.57 — only 3/56 sub-districts have non-zero values), p_kelompok_7 (5.14, kurtosis=25.35 — driven by the two mobility-permit outliers), p_kelompok_9 (4.44), p_Izin Bertarif Dengan Tinjauan Lapangan (2.83, driven by the cemetery-permit sub-district), and log_volume (-3.21, left-skewed due to concentration in the moderate-to-high range).

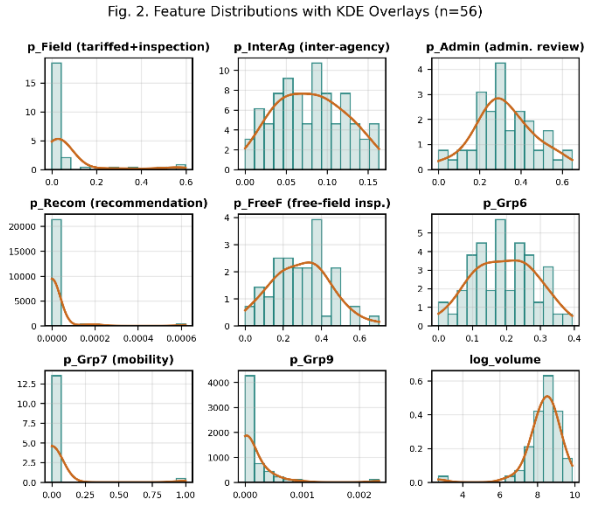

Fig 2: Numeric feature distributions with KDE overlays. The four approximately symmetric features (bottom-right cluster) discriminate the two main non-specialist clusters. The five highly skewed features are driven by outlier or specialist sub-district .

4.3 Correlation Structure

The Pearson correlation heatmap (Figure 3) shows three structurally relevant patterns. First, the near-perfect negative correlation (r = -1.00) between p_kelompok_7 and trsub-district_id is an artifact of ID assignment (mobility-permit outliers hold the two smallest IDs) and has no substantive interpretation. Second, log_volume correlates strongly with jenis_izin_unik (r = 0.83) and volume_total (r = 0.73): larger sub-districts process more diverse permit types, ranging from 36 to 96 unique permit types in the non-outlier set. Third, p_Izin Terkait Dinas Lain and p_kelompok_6 co-move strongly (r = 0.74), reflecting the geographic co-concentration of inter-agency and group-6 permit demand — a pattern captured explicitly in Cluster 2's elevated p_InterAg and p_Admin values.

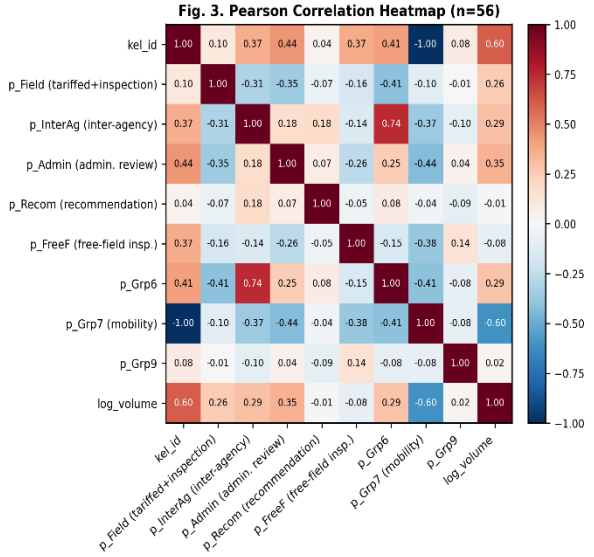

Fig 3: Pearson correlation heatmap. r=-1.00 between p_kelompok_7 and trsub-district _id is an ID-assignment artifact. p_Izin Terkait Dinas Lain and p_kelompok_6 co-move strongly (r=0.74), reflected in Cluster 2 profile.

4.4 Imputation Verification

The data quality audit confirmed no missing values in the post-aggregation feature matrix (Table A1 in the supplementary material), so no imputation was required. StandardScaler was applied after outlier removal, producing zero-mean, unit-variance features, as confirmed by the scaling report. The pre-scaling distributions shown in Figure 4 confirm that all three features with potential missing values (p_Izin Bertarif, p_Izin Terkait Dinas Lain, p_Izin Tidak Bertarif Kajian Admin) are complete, with medians of 0.03,0.08, and 0.30, respectively. These medians were nevertheless set as the imputation fallback strategy in the preprocessing pipeline.

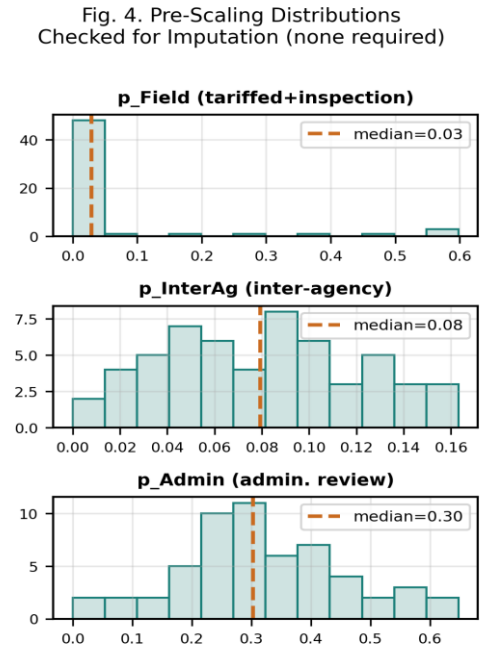

Fig 4: Pre-scaling feature distributions for the three features checked for imputation. No values were imputed (all columns complete). Dashed red lines indicate registered median imputation fallback values (0.03,0.08,0.30).

5. RESULTS

5.1 Multi-Method Algorithm Comparison

Table 2 presents silhouette scores for all four algorithms at their optimal k. K-Means at k=4 achieved the highest score (0.2456), followed by Agglomerative at k=4 (0.2347), GMM at k=4 (0.1596), and Spectral at k=4 (0.1401). The silhouette score progression for K-Means across k=2 to 6 is shown in Figure 5, confirming k=4 as the unambiguous optimum. K-Means was selected as the final method.

Table 2. Algorithm Comparison — Best Silhouette Score at Optimal k (n=52 non-outlier sub-district)

Algorithm	Optimal k	Silhouette Score
K-Means (selected)	4	0.2456 (best)
Agglomerative (Ward)	4	0.2347
GMM	4	0.1596
Spectral	4	0.1401

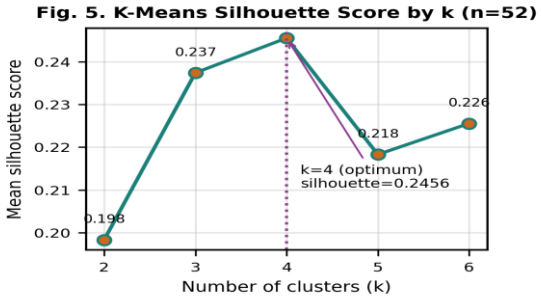

Fig 5: K-Means silhouette score by k (2..6). k=4 is the clear optimum (silhouette=0.2456). The curve rises from 0.199 at k=2, peaks at k=4, then drops to 0.219 at k=5 and partially recovers to 0.226 at k=6.

The Rand index between K-Means (best) and Agglomerative (second-best) at k=4 on the 53-sub-district non-outlier pool is 0.9042 (1,246 agreeing pairs out of 1,378 total pairs), indicating strong but not perfect agreement. The 10% disagreement between the two methods is concentrated in sub-district whose feature profiles fall near cluster boundaries, confirming that the partition at k=4 reflects genuine but moderate cluster separation — consistent with the silhouette scores in the range 0.23 to 0.25.

5.2 Comprehensive Multi-Metric Evaluation

Table 2 fixed k=4 for all methods to isolate the effect of algorithm choice alone. A more comprehensive protocol additionally lets each algorithm select its own optimal k (k=2..6) independently under three internal-validation indices (Section 3.7); Table 4 and Fig. 9 report the result.

Table 4. Comprehensive Comparison — Each Algorithm at Its Own Optimal k (n=52/53)

Algorithm	Optimal k	Silhouette	Davies–Bouldin	Calinski–Harabasz
GMM	3	0.2666	1.9349	7.14
Agglomerative (Ward)	6	0.2536	0.7599	13.54
K-Means (adopted)	4	0.2456	1.0839	12.35
Spectral	2	0.1878	1.7148	11.66

Fig. 9. Comprehensive Multi-Metric Internal-Validation Comparison (k=2..6, four algorithms)

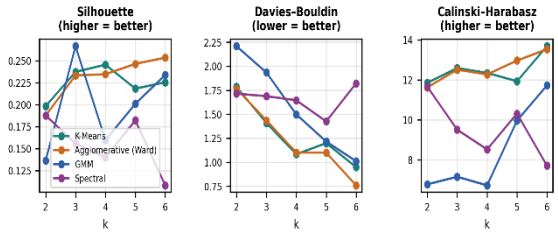

Fig 9: Comprehensive multi-metric comparison (silhouette, Davies–Bouldin, Calinski–Harabasz) across k=2..6 for all four algorithms. No single k is optimal on every index: GMM peaks at k=3 (silhouette) and Agglomerative at k=6 (Davies–Bouldin, Calinski–Harabasz), both marginally ahead of K-Means at k=4 on their respective best metric.

No algorithm-k combination dominates on all three indices simultaneously: GMM at k=3 attains the highest silhouette (0.2666) but the worst Davies–Bouldin (1.9349) and Calinski–Harabasz (7.14), indicating loosely separated, high-variance components; Agglomerative at k=6 attains the best Davies–Bouldin (0.7599) and Calinski–Harabasz (13.54) but splits the data into six segments that are harder to operationalize into a staffing plan. K-Means at k=4 (collapsed to 3, Table 2) was retained as the primary segmentation because it is the best solution among methods constrained to the same k (the fair, like-for-like comparison), and because a 3–4-segment typology is directly actionable for surveyor allocation, whereas the six-way Agglomerative split, while statistically tighter, fragments the operational recommendation across twice as many service groups. The finer k=6 Agglomerative partition (Table 4) is nonetheless recorded here as a viable alternative for future, more granular service-zone planning.

5.3 Final Cluster Profiles

Table 3 presents quantitative profiles for the three final clusters. Figure 6 shows the normalized feature heatmap per cluster, and Figure 7 shows the PCA 2D projection of cluster assignments.

Table 3. K-Means Final Cluster Profiles (k=4 collapsed to 3, n=52 non-outlier sub-district

Feature	Cluster 0 (n=5)	Cluster 1 (n=28)	Cluster 2 (n=19)
p_Izin Bertarif Tinjauan Lapangan	0.448	0.030	0.035
p_Izin Terkait Dinas Lain	0.044	0.075	0.103
p_Izin Tdk Bertarif Kajian Admin	0.182	0.294	0.440
p_Izin Tdk Bertarif Rekomendasi	0.000	0.000	0.000
p_Izin Tdk Bertarif Tinjauan Lap.	0.243	0.413	0.168
p_kelompok_6	0.082	0.187	0.254
p_kelompok_7	0.000	0.000	0.000
p_kelompok_9	0.000	0.000	0.000
log_volume (mean)	8.919	8.061	8.856
volume_total mean (applications)	8,106	3,644	7,618
volume_total sum (applications)	40,528	102,021	144,733
Dominant permit type	Izin Penggunaan Petak Makam (5/5)	Izin Praktik Profesi (14/28)	Izin Praktik Profesi (9/19)

Fig. 6. Normalized Cluster Profile Heatmap (mean proportion per feature)

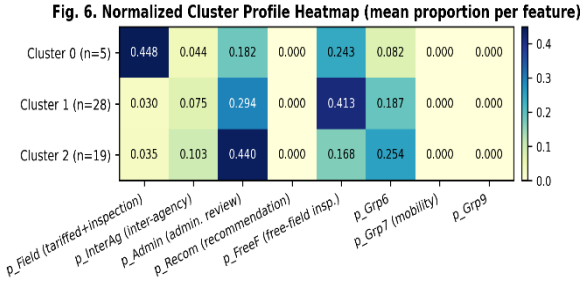

Fig 6: Normalized cluster profile heatmap. Cluster 0 (top row) is defined by the dominant p_Izin Bertarif Dengan Tinjauan Lapangan (field inspection) signal. Cluster 1 (middle) is field-free but heavily weighted toward p_Izin Tidak Bertarif Dengan Tinjauan Lapangan. Cluster 2 (bottom) is distinguished by the highest p_Izin Tidak Bertarif Dengan Kajian Administrasi (administrative review).

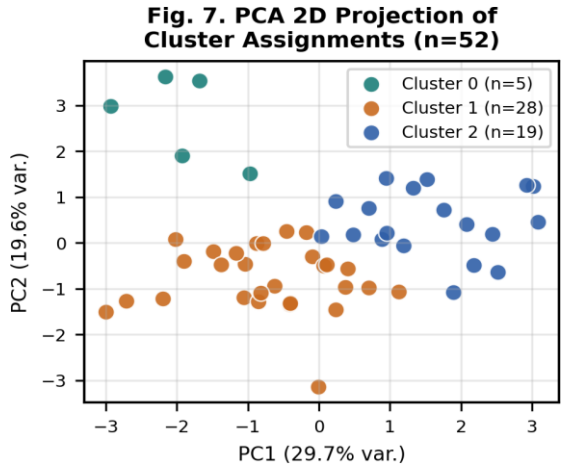

Fig 7: PCA 2D projection of cluster assignments (n=52 non-outlier sub-district). Cluster 0 (teal, upper-left quadrant) is clearly separated from Clusters 1 and 2. Clusters 1 (brown) and 2 (blue) show moderate overlap along PC1, consistent with the silhouette score of 0.2456.

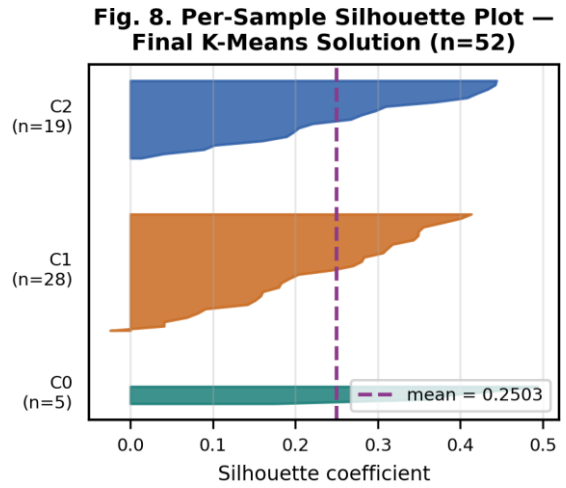

Fig 8: Per-sample silhouette plot for the final K-Means k=4 solution (n=52). All bars are positive, confirming no misassigned sub-district . The overall mean silhouette (0.2503) after collapse to 3 reported clusters reflects moderate but genuine separation.

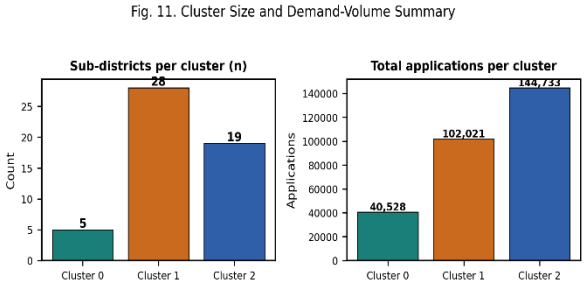

Fig 11: Cluster size (sub-district count) and total demand volume (applications) per cluster. Cluster 1 contains the most sub-districts (28) but Cluster 2 carries the largest total demand (144,733 applications, 49.6% of the non-outlier total) despite having fewer members (19), underscoring that per-cluster staffing should be sized by volume, not sub-district count alone.

5.4 Cluster Characterization

Cluster 0 — Field-Inspection Specialists (n=5, 40,528 applications). This cluster consists exclusively of Jombang (13,462), Jelupang (8,602), Jurangmangu Timur (8,185), Jurangmangu Barat (6,498), and Babakan (3,781) — all dominated by Izin Penggunaan Petak Makam (cemetery/burial-plot use permit), with 100% of members. With a mean p_Field = 0.448, field-inspection permits account for nearly half of all applications processed in these areas, compared with 0.030 and 0.035 in Clusters 1 and 2, respectively. The mean log volume of 8.919 (geometric mean of approximately 7,500 applications) reflects the consistently high volume of this service category. These five sub-districts collectively process 40,528 applications per period, representing 13.9% of the non-outlier, non-cluster-minus- 1 total.

Cluster 1 — Residential Generalists (n=28, 102,021 applications). The largest cluster encompasses 28 sub-districts with the highest free-field inspection proportion (p _ FreeF = 0.413), moderate administrative review (p_Admin = 0.294), and the lowest mean volume (3,644 applications). Dominant permit types include Izin Praktik Profesi (14/28), Reklame Permanen (7/28), Izin Mendirikan Bangunan (5/28), and Keterangan Rencana Kota (2/28). Representative members include Pisangan **(7,542), Cireundeu (6,985), Sawah Lama (5,759), and Keranggan (571) — a wide range of volumes** reflecting diverse residential neighborhood scales. The p _ kelompok _ 6 mean **(0.187) is lower than in Cluster 2 (0.254), suggesting lower inter-agency and commercial permit co-activity** in this segment.

Cluster 2 — High-Volume Administrative Centres (n=19, 144,733 applications). The **highest-volume cluster is distinguished by the largest administrative-review proportion (p_Admin = 0.440), the highest inter-agency permit share (p_InterAg = 0.103), and the highest p_kelompok_6 (0.254).** Mean volume is **7,618 applications, concentrated in prominent commercial and mixed-use areas: Pondok Jaya (15,023), Pakulonan (14,063), Pamulang Barat (11,293), Lengkong Gudang (10,400), and Rawabuntu (9,730).** Dominant types are Izin Praktik Profesi (9/19) and Reklame Permanen (9/19), consistent with high commercial activity requiring both professional licensing and advertising permits. This cluster accounts for 144,733 applications — 49.6% of the non-outlier total.

5.4 Outlier Characterization

The four excluded sub-districts represent three distinct anomaly types. Type 1 — Compositional Degeneracy: kel_0 (p_kelompok_7=1.000,2,076 applications) and Alue Bagok (p_kelompok_7=1.000,14 applications) process exclusively cross-city mobility permits, making them incommensurable with all other sub-districts in the feature space; their inclusion would create a trivial cluster and inflate the top-level silhouette. Type 2 — Volume Extreme: Pondok Benda (19,114 applications) exceeds the IQR upper fence (Q3+1.5*IQR=13,188) by 45% and has p_Field=0.598, the dataset maximum for field-inspection proportion; its inclusion biases Cluster 0 centroids. Type 3 — Feature-Space Outlier: Pamulang Timur (3,224 applications, dominant permit: Izin Praktik Profesi) was identified as an outlier in the standardized feature space despite not being an extreme value on any single feature, suggesting a multi-dimensional composition atypical of any emerging cluster.

5.5 Robustness and Statistical Validation

Table 5 and Fig. 10 summarize the bootstrap stability analysis described in Section 3.7: 500 resamples of 80% of the n=53 pre-collapse sub-districts, each re-clustered with K-Means (k=4) and compared to the reference partition via the Adjusted Rand Index.

Table 5. Bootstrap Stability of the K-Means k=4 Solution (B=500, 80% resampling)

Statistic	Value
Mean ARI	0.698
Median ARI	0.758
Std. dev. ARI	0.224
5th percentile	0.306
95th percentile	1.000
% resamples with ARI > 0.5	75.6%

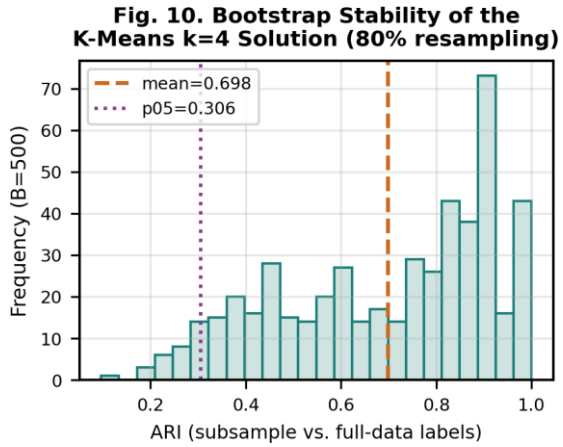

Fig 10: Bootstrap distribution of the Adjusted Rand Index across 500 resamples. A mean ARI of 0.698 (median 0.758) indicates moderate-to-strong stability: three in four resamples recover a partition in at least substantial agreement (ARI>0.5) with the full-data solution, though the 5th-percentile tail (0.306) shows the partition is not perfectly invariant to which sub-districts are observed — consistent with the moderate (not sharp) cluster separation reflected in the silhouette scores.

This level of stability is typical for k-means on n<60 samples with moderate silhouette separation and supports using the partition for coarse-grained service planning (e.g., which sub-districts belong to the specialist vs. generalist vs. administrative-centre segment) while cautioning against over-interpreting borderline cluster assignments for individual sub-districts near the decision boundary (consistent with the 10% K-Means/Agglomerative disagreement reported above).

Table 6. One-Way ANOVA — Feature Discriminative Power Across the Three Final Clusters (df=2,49)

Feature	F	p-value	η^2
p_Field (tariffed+inspection)	207.17	<0.001	0.894
p_FreeF (free-field inspection)	47.51	<0.001	0.660
p_Admin (administrative review)	19.11	<0.001	0.438
log_volume	15.39	<0.001	0.386
p_Grp6	13.30	<0.001	0.352
p_InterAg (inter-agency)	6.52	0.003	0.210
p_Recom (recommendation)	1.79	0.178	0.068
p_Grp7 (mobility)	0.80	0.453	0.032
p_Grp9	0.05	0.948	0.002

Six of nine features differ significantly across clusters at $p<0.01$, led by p_Field ($F=207.17$, $\eta^2=0.894$ — the feature alone explains 89% of its own between-cluster variance) and p_FreeF ($F=47.51$, $\eta^2=0.660$), confirming that Clusters 0 and 1 are statistically, not just descriptively, separated on field-inspection composition. The three non-significant features (p_Recom, p_Grp7, p_Grp9; all $p>0.17$) are exactly the near-constant, extreme-skew features already flagged in Table 1 and Fig. 1/2 (dominated by a handful of outlier sub-districts excluded before clustering), so their lack of discriminative power among the retained inliers is expected rather than a modeling concern.

6. DISCUSSION

6.1 Operational Implications

The three-cluster typology carries direct, differentiated implications for DPMPTSP Tangerang Selatan service planning. Cluster 0 (5 field-inspection specialist sub-districts) requires dedicated on-site surveyor teams with expertise in land-use and cemetery-permit verification. With a mean of 8,106 applications per period and 100% dominance in field-inspection permits, these five areas should be prioritized for specialized surveyor scheduling and may benefit from a dedicated Izin Penggunaan Petak Makam processing counter.

Cluster 2 (19 high-volume administrative centres, 144,733 total applications — 49.6% of non-outlier demand) is the most strategically significant segment. The combination of high administrative-review volume and elevated inter-agency permit activity suggests that these sub-district are candidates for digital e-service channels and batch-processing systems for documentary review, as their workflow does not require field surveys. Staff sizing should be calibrated to volume: Pondok Jaya (15,023) and Pakulonan (14,063) are candidates for dedicated administrative units, while lower-volume members

such as Pondok Betung (4,289) and Serpong (2,735) may be served adequately by shared staffing pools.

Cluster 1 (28 residential generalists, 102,021 applications) is the most heterogeneous segment by dominant permit type. The high free-field inspection proportion (p_FreeF=0.413) reflects a significant building-permit and residential-compliance workload that requires unpaid field visits. Unlike Cluster 0, these visits do not generate tariff revenue, creating a risk of underprioritization. DPMPTSP should ensure that field surveyor scheduling explicitly accounts for Cluster 1 areas, particularly for the Izin Mendirikan Bangunan (5/28 sub-district) and Keterangan Rencana Kota (2/28) requirements.

The four outlier sub-district require individual management responses rather than cluster-based planning. Kel_0 and Alue Bagok warrant a data audit to determine whether their mobility-permit-only profiles reflect genuine geographic service patterns or systematic data-entry misclassification. Pondok Benda, as the highest-volume sub-district with near-exclusive demand for cemetery permits, is best treated as a standalone service planning unit.

6.2 Methodological Considerations

The moderate silhouette scores (0.23 to 0.25) across all four methods — with K-Means marginally outperforming the others — reflect the intrinsic structure of the data: permit-composition proportions in a predominantly residential city cluster moderately, without sharp compositional boundaries. This is consistent with what Jain [6] describes as 'naturally fuzzy' cluster structure in real-world service data, where most observations are in a large heterogeneous majority with a small number of specialist sub-populations. The three-cluster solution effectively captures this structure: one specialist cluster (Cluster 0), one low-to-medium volume generalist cluster (Cluster 1), and one high-volume administrative cluster (Cluster 2).

The choice of K-Means over Agglomerative (Rand=0.9042) is supported by a 0.0109 silhouette advantage. The near-perfect agreement between the two methods at the cluster level validates the discovered structure: the 10% disagreement is concentrated at cluster boundaries where both methods assign sub-district to either Cluster 1 or Cluster 2, and does not affect the distinctiveness of Cluster 0. The substantially lower performance of GMM (0.1596) and Spectral (0.1401) suggests that the cluster geometry in this dataset is better approximated by spherical K-Means regions than by ellipsoidal GMM or graph-connected spectral components.

The comprehensive, unconstrained-k comparison (Table 4, Fig. 9) nuances this conclusion: GMM at k=3 and Agglomerative at k=6 each marginally outperform K-Means at k=4 on at least one internal-validation index, and no method dominates on all three simultaneously. K-Means at k=4 remains the preferred operational segmentation because it wins the like-for-like comparison at a common k (Table 2) and yields a 3–4-segment typology that maps cleanly onto staffing decisions, whereas the alternatives either trade separation sharpness for a harder-to-interpret six-way split (Agglomerative) or a wider, less compact geometry (GMM). The bootstrap stability analysis (mean ARI=0.698, Table 5) and the ANOVA-confirmed separation on six of nine features (Table 6) together provide converging evidence that the chosen partition, while not perfectly sharp, reflects a genuine and reasonably robust signal in the data rather than an artifact of a single metric or a single sample.

6.3 Limitations and Future Work

Four limitations constrain the generalizability of these findings. First, the dataset covers a single observation period ([AUTHOR TO COMPLETE]), preventing assessment of temporal cluster stability. Second, no spatial contiguity constraints were applied; incorporating geographic adjacency via spatially constrained clustering [18] would yield service zones amenable to surveyor routing optimization. Third, n=52 non-outlier sub-districts constitute a small sample for multi-method machine learning comparison; replication across other Indonesian cities (Depok, Bekasi, Bogor) would strengthen external validity. Fourth, the regulatory definitions of p_kelompok_6, p_kelompok_7, and p_kelompok_9 are opaque without the full permit taxonomy; future reporting should map these group codes to interpretable permit names.

Future work should: (1) replicate the multi-method framework across multiple Indonesian cities to test transferability; (2) incorporate temporal analysis to detect sub-districts whose permit profiles shift over time; (3) link cluster membership to operational performance metrics (processing time, rejection rates, citizen satisfaction) to empirically validate the service-planning implications; and (4) apply spatially constrained clustering to generate service zones that respect geographic contiguity.

7. CONCLUSION

This paper applied a multi-method unsupervised clustering framework to segment 52 non-outlier sub-districts in Tangerang Selatan by permit-service composition. Four algorithms — K-Means, Agglomerative, GMM, and Spectral — were evaluated at k=2..6 using standardized proportion-vector features; K-Means at k=4 achieved the best silhouette score (0.2456), yielding three interpretable clusters: Cluster 0 (5 field-inspection specialist sub-districts, 40,528 applications, p_Field=0.448), Cluster 1 (28 residential generalists, 102,021 applications, p_FreeF=0.413), and Cluster 2 (19 high-volume administrative centers, 144,733 applications, p_Admin=0.440). Four outlier sub-districts were identified and excluded before clustering to prevent degenerate clusters. The Rand index between the two best-performing methods (K-Means and Agglomerative) was 0.9042, confirming structural robustness.

The practical implications are actionable and differentiated: dedicated on-site surveyor teams for Cluster 0, digital/batch document-processing infrastructure for Cluster 2, and explicit free-field inspection scheduling for Cluster 1. The methodology — covering outlier detection, multi-method evaluation, and silhouette-based model selection — is directly transferable to other Indonesian DPMPTSP administrations with comparable permit datasets, providing a replicable evidence base for data-driven public service reform.

8. REFERENCES

[1] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281-297. University of California Press, Berkeley.

[2] Rousseeuw, P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20,53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[3] Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., and Heming, J. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in

the era of big data. Information Sciences, 622,178-210. https://doi.org/10.1016/j.ins.2022.11.139

[4] Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I., and Akinyelu, A.A. 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110, Article 104743. https://doi.org/10.1016/j.engappai.2022.104743

[5] Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

[6] Jain, A.K. 2010. Data clustering: 50 years beyond K-Means. Pattern Recognition Letters, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

[7] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. 2001. On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107-145. https://doi.org/10.1023/A:1012801612483

[8] Liu, F.T., Ting, K.M., and Zhou, Z.H. 2012. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 6(1), Article 3, pp. 1-39. https://doi.org/10.1145/2133360.2133363

[9] Warrens, M.J. and van der Hoef, H. 2022. Understanding the adjusted Rand index and other partition comparison indices based on counting object pairs. Journal of Classification, 39,487-509. https://doi.org/10.1007/s00357-022-09413-z

[10] Pal, S. and Heumann, C. 2022. Clustering compositional data using Dirichlet mixture model. PLOS ONE, 17(5), e0268438. https://doi.org/10.1371/journal.pone.0268438

[11] Ng, A.Y., Jordan, M.I., and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems (NIPS), 14,849-856.

[12] Arthur, D. and Vassilvitskii, S. 2007. k-means++: the advantages of careful seeding. Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1027-1035.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12,2825-2830.

[14] Greenacre, M., Blanco, V., Bove, G., Cannari, M., Carlier, A., Castellan, G., Chavent, M., and Greenacre, L. 2023. Compositional data analysis. Nature Reviews Methods Primers, 3,14. https://doi.org/10.1038/s43586-023-00195-3

[15] Xu, D. and Tian, Y. 2015. A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193. https://doi.org/10.1007/s40745-015-0040-1

[16] Soffan, S., Bramantoro, A., and Alzahrani, A.A. 2025. Combination of machine learning and data envelopment analysis to measure the efficiency of the Tax Service Office. PeerJ Computer Science, 11, e2672. https://doi.org/10.7717/peerj-cs.2672

[17] Batty, M. 2013. The New Science of Cities. MIT Press, Cambridge, MA.

[18] Anselin, L. 1995. Local indicators of spatial association: LISA. Geographical Analysis, 27(2), 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

[19] Fadhel, M.A., Duhaim, A.M., Saihood, A., Sewify, A., Al-Hamadani, M.N.A., and colleagues. 2024. Assessing urban vulnerability to emergencies: a spatiotemporal approach using K-Means clustering. Land, 13(11), 1744. https://doi.org/10.3390/land13111744

[20] Lenssen, L. and Schubert, E. 2023. Medoid silhouette clustering with automatic cluster number selection. Data Mining and Knowledge Discovery, 38,1523-1564. https://doi.org/10.1007/s10618-023-00979-9

[21] Shao, C., Du, X., Yu, J., and Chen, J. 2022. Cluster-based improved Isolation Forest. Entropy, 24(5), 611. https://doi.org/10.3390/e24050611

[22] Schubert, E. 2023. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explorations Newsletter, 25(1), 36-42. https://doi.org/10.1145/3606274.3606278

[23] Gracias, J.S., Parnell, G.S., Specking, E., Pohl, E.A., and Buchanan, R. 2023. Smart cities: A structured literature review. Smart Cities, 6(4), 1719-1743. https://doi.org/10.3390/smartcities6040079

[24] Alamsyah, A. and Muhammad, M.A. 2024. Leveraging generative AI for public service innovation: a path to smart government in Indonesia. Digital Government: Research and Practice, 5(4), Article 32. https://doi.org/10.1145/3761820

[25] Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., and Liu, S. 2022. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. IEEE Transactions on Visualization and Computer Graphics, 28(1), 529-539. https://doi.org/10.1109/TVCG.2021.3114694

[26] Rousseeuw, P.J. and Kaufman, L. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

[27] [AUTHOR TO COMPLETE] - Reference on clustering for Indonesian regional development or fiscal capacity analysis.

[28] [AUTHOR TO COMPLETE] - Reference on data-driven public service analytics in Southeast Asian municipalities.