Understanding postal delivery areas in the Republic of Korea using multiple unsupervised learning approaches

Keejun Han1 | Yeongwoong Yu1 | Dong-gil Na1 | Hoon Jung1 | Younggyo Heo2 | Hyeoncheol Jeong2 | Sunguk Yun2 | Jungeun Kim2

1Intelligent Convergence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea
2Department of Computer Science and Engineering, Kongju National University, Cheonan, Republic of Korea

Correspondence
Jungeun Kim, Kongju National University, Cheonan, Republic of Korea. Email: jekim@kongju.ac.kr

Funding information
Institute for Information and Communications Technology Promotion, Grant/Award Number: 1711125741; Institute of Information & communications Technology Planning & Evaluation (IITP); Korea government (MSIT), Grant/Award Number: 2018-0-01664

Abstract
Changes in household composition and the residential environment have had a considerable impact on the features of postal delivery regions in recent years, resulting in a large increase in the overall workload of domestic postal delivery services. In this paper, we provide complex analysis results for postal delivery areas using various unsupervised learning approaches. First, we extract highly influential features using several feature-engineering methods. Then, using quantitative and qualitative cluster analyses, we find the distinctive traits and semantics of postal delivery zones. Unsupervised learning approaches are useful for successfully grouping postal service zones, according to our findings. Furthermore, by comparing a postal delivery region to other areas in the same group, workload balancing was achieved.

KEYWORDS
clustering, feature engineering, postal delivery management, unsupervised learning, workload balancing

1 | INTRODUCTION

Recently, the Republic of Korea has experienced rapid growth of single-person households in urban areas and retirement homes in rural areas. This has inevitably increased the overall workload of postal delivery services. According to Kim [1], mail carriers in the Republic of Korea work on average 2745 h annually, 693 h higher than other regular paid occupations. Mail carriers in other wealthy countries work approximately 900 h more per year than waged employees. This indicates that, on a scale of 8 h each weekday, they work for approximately 87 to 123 days more than other waged employees in the Republic of Korea and other developed countries, respectively.

In terms of labor intensity, Borg's rating for mail carriers, which is a measurement for evaluating physical labor intensity, is evaluated as 14 of 20, indicating that the work requires high physical activities [2]. The rising death rate of mail carriers in the Republic of Korea over the last decade has been attributed to this severe job [3].

To solve this, a system that measured the workload of each mail carrier for reallocation was proposed in 2017 [4]. The method considered various factors of mail delivery services, such as daily mail volume per person and delivery site type, which affects delivery intensity, and linearly aggregated them into a numerical value termed delivery workload value. More than 40 workload elements were used to compute the delivery labor intensity as precisely as possible; however, the system failed owing...
to two major flaws: the curse of too many features and the potential for system exploitation.

First, the system designed a fixed formula that measured the labor intensity that considered more than 40 features that were determined by conducting multiple surveys and focused interviews involving mail carriers. Because of the human-centric approach, the system took into account a variety of factors that postal carriers believed influenced their labor intensity. The significance of the features, however, was disregarded, potentially resulting in a discrepancy between the predicted delivery workload value and the real labor settings. Because some features are recorded manually and the delivery is not totally digitalized, quantifying the human-centric variables is difficult.

Second, the purpose of the system was to measure the labor intensity of each mail carrier and use the result to restructure the carrier service by reallocation or additional staffing. However, rather than functioning as an indicator to reduce the burden of postal carriers with high labor intensity, the system assigned more responsibilities to mail carriers with relatively lower labor intensity as a side consequence. Because of this unforeseen misuse, the system should not provide any human-related scores that could be abused.

To overcome the aforementioned limitations, we propose a novel framework that provides complex analysis results on postal delivery areas by exploiting various unsupervised learning approaches. Every single unit allotted to each mail carrier is referred to as a postal delivery area. Individual postal delivery zones have different delivery circumstances depending on whether they are urban or rural. Owing to the high number of buildings and residents in cities, huge amounts of mail are delivered. However, in a rural area, the overall distance between delivery points can be longer, causing frequent movements during delivery, although the number of mails is relatively smaller. The necessity to demonstrate the uniqueness and difference of individual delivery areas by comparing them to their similar group is addressed by this imbalance among postal delivery zones. We also omit the labor intensity value from available features to minimize misuse, forcing us to employ unsupervised machine learning algorithms using unlabeled data. To our knowledge, this is the first study to use unsupervised learning algorithms to examine postal delivery areas.

Contributions made through this paper are as follows.

- We discovered important features that differentiate postal delivery service areas in the Republic of Korea using several feature selection and extraction methods.
- We further unveiled the hidden patterns in the postal delivery area dataset by performing multiple clustering algorithms, which focused more on the postal delivery area, not mail carriers.
- We uncovered the intrinsic traits and semantic clusters of postal service areas through extensive quantitative and qualitative investigation, with each cluster representing distinct geographical characteristics in the Republic of Korea.

The remainder of this paper is organized as follows: Section 2 reviews related studies. Section 3 describes the dataset of the postal delivery area. Section 4 discusses data preprocessing and feature engineering. Section 5 introduces unsupervised methods used in this paper. Sections 6 and 7 show our findings by quantitative and qualitative analysis. Section 7 suggests implications based on the findings, and Section 9 presents our conclusion.

2 | RELATED WORK

This section reviews several relevant studies on workload identification of postal delivery services. Thereafter, related work is introduced on delivery optimization to facilitate Logistic 4.0, which is described in Section 2.2.

2.1 | Workload estimation for postal delivery services

The need for automating delivery processes and measuring work intensity has gained widespread attention since Lee and Kim [5] suggested that automation would relieve fatigue and heavy workload on mail carriers. Sun and others [6] conducted a study to standardize delivery workload for mail carriers. They divided delivery areas into three categories: apartment, multiplex house, and detached house. They also divided delivery sites into three categories: apartment, multiplex house, and detached house. Individual mail carriers’ workloads were calculated by putting various weights on different types of delivery locations and points. However, Yang and others [7] specified the mail carriers task into three sub-tasks: preparation, the movement toward the assigned delivery area, and delivery. To compute the final workload of the mail carriers, they allocated varied times to preparation and movement tasks from 36 specified delivery kinds, as well as delivery duties from 100 predetermined delivery sites. Lee and others [8] introduced postal delivery management systems, which digitalized postal delivery area information, in various countries: New Zealand, Japan, and the United States. By information digitalization, those services measured the work intensity for each postal delivery area and equalized the
intensity among the delivery areas. However, the services offered statistical figures only, not the machine learning results.

There was an alternative attempt exploiting crowd shipping [9], outsourced parcel delivery service to occasional carriers. Regular mail carriers may be relieved of their duties as a result of this method. The challenge of distributing work to the audience, however, was complex and difficult. Instant delivery [10] and order consolidation [12] would be applied for employing occasional carriers; however, heuristic algorithms were yet mainly proposed. Our research, however, used unsupervised learning methodologies to take a more data-centric approach to workload balancing.

In the above studies [6–8], they defined the limited number of postal delivery area types before measuring the work intensity; however, approximately 500 post offices in the Republic of Korea exist, implying that the number of types would be oversimplified. However, before measuring the work intensity, Lee and others [13] defined the standard postal delivery service processes and then assigned weights on each process. The final work intensity for individual mail carriers was obtained using a linear aggregation of multiplying the completed individual processes and their weights. Kim and others [14] validated those work time measures by comparing the workload estimation and actual work time of mail carriers. In Park and Park [15], predefining subprocesses for the postal delivery service were chosen, and the initial system architecture and user interfaces to visualize the workload intensity measurement were given. Postal delivery is done by contracted mail carriers or part-time flexibles. Kim and others [16] further considered those carriers for workload estimation by differentiating sub-tasks for the flexible and regular carriers, respectively.

In this paper, instead of the conventional approaches, we do not directly measure the workload estimation of individual mail carriers by the work time measures since it can be misused during workload balancing. Alternatively, we investigate the circumstances in postal delivery areas. Furthermore, we adopted multiple unsupervised learning methods to discover hidden patterns from various datasets with a data-driven view [17].

2.2 Delivery service optimization

Advances in digital technologies have enabled an efficient work balancing and standardization of logistic information, causing a new era of Logistic 4.0 [18]. Logistic 4.0 stands in a broader sense of Industry 4.0, including five functional areas valid across businesses, namely, data collection and processing, assistance systems, networking and integration, decentralization and service orientation, and self-organization and autonomy [19]. A delivery management platform that provides analytic results for optimal delivery is a representative example of digital transformation for logistic data to enhance the productivity of the delivery and pickup service so that the limited resources can be fully exploited to meet the customers’ needs and reduce costs.

Initially, Wong and Beasley [20] proposed a heuristic algorithm for dividing the depot area into subareas, from an initial allocation of customers to subareas followed by customer interchanges to improve the subareas. In practice, however, most of the service zones were manually defined by referring to the subdivisions designed by other operators for similar services, such as ZIP codes or city districts. These caused highly unbalanced workloads due to the geographical variants of the zones [21]. Wong [22] introduced some practical approaches for delivery zone partitioning and addressed again that parcel pickup and delivering are vital components of national and international transportation.

Moreover, Carlsson [23] proposed an algorithm to solve vehicle routing problems where vehicle depot locations are fixed, and client locations in a service region are unknown and assumed to be independent. The algorithm was designed to be flexible and applicable to other load balancing tasks. Haugland and others [24], under the circumstance of designing districts for vehicle routing problems with stochastic demand, exploited variants of tabu search and multistart heuristic to design delivery districts. Schneider and others [25] focused on the influence of time window constraints on delivery routing and addressed that considering geographical aspects in districts is significant for generating high-quality territories, whereas explicitly incorporating time window characteristics and historical demand data does not lead to a perceptible improvement of the solution quality. Parariani and others [21] proposed a decision support system that partitioned package delivery zones by exploiting an advanced multi-attribute clustering algorithm, a variant of the $k$-means algorithm to match between customers and clusters. Barzegar and others [26] exploited expert knowledge of seasoned individuals, such as taxi drivers to find low-traffic routes in big cities by constructing a knowledge base in a form of ontology.

Last-mile delivery using unmanned robots and urban air mobility to carry items to their final destination has been highlighted as a viable metropolitan freight transportation strategy as a result of the global Covid-19 outbreak [11]. Some of the attempts include drones for parcel delivery [27], integration with public transport [28], and autonomous vehicles [29]. Despite these advancements, last-mile delivery remains the most
expensive section of the supply chain, as it presents couriers with various operational obstacles, such as discovering the location in enormous buildings that require a security pass [9].

However, the literature above is either too complex or time-consuming in large-scale real-world applications. More so, they are irrelevant for delivery optimization tasks in the Republic of Korea with rapid changes in urban and rural delivery environments due to massive urban regeneration and development, causing a frequent workload balancing. Furthermore, the literature has approached delivery optimization as a routing problem. To the best of our knowledge, this is the first work that analyzes postal delivery areas with unsupervised learning approaches.

3 CHARACTERISTICS OF THE DATASET

The dataset used in this paper is the nationwide postal delivery area information of the Republic of Korea. It was initially sampled from PostNet, a postdelivery management portal, internally used in Korea Post¹. The dataset contains the postal delivery area information extracted from nine provinces in the Republic of Korea from January to December 2020.

Table 1 summarizes the profiles of the dataset used for our analysis. Post office ID + postal delivery area ID is a unique Identifier for the postal delivery area. There may be potential conflicts between postal delivery area IDs. However, putting the IDs together prevents these conflicts. Mail bundling is a task to group mails and parcels to improve delivery efficiency. Total delivery time includes the actual delivery time, movement to delivery points, and extra activities like delivery notice and pickup. The volume of batch delivery heavily affects the delivery workload. Although the number of mails is high, the delivery time is relatively short if the mails are sorted in batches since mail batches are delivered at once for the same delivery point. Total delivery distance is a sum of the distance from the first to last delivery points. It is recorded manually by checking the odometer of the delivery vehicle after delivery. In addition, building types are a crucial factor for measuring the delivery workload.

For data preprocessing, we eliminated missing values. Since the variances of the attributes were generally large, we conducted a min-max normalization to rescale the values of the attributes from 0 to 1, to prevent data overflow or underflow during the analysis because some attributes had relatively low nonzero ratios.

4 FEATURE ENGINEERING

Feature engineering is the process of transforming the given data into features that state the problem of the predictive model well to increase model accuracy and reduce complexity. It is divided into both feature selection and extraction. The feature selection selects features among $N$ features, whereas feature extraction creates $m$ new features that best represent $N$ features ($n$ and $m \leq N$).

We adopted two methods for feature selection: mean absolute deviation (MAD) and localized feature selection based on scatter separability (LFS) [30]. MAD is the average distance between each feature value and its mean, meaning that the indicator represents the distribution of the data. We only chose features valued at 0.05 or above. LFS was derived from the fact that feature sets between two clusters need not be the same. For instance, a document related to sports news would be highly classified into a set of clusters with common vocabularies, such as FIFA and ball. Meanwhile, a document related to tech blogs would be classified with common vocabularies, such as Apple and IBM. LFS selects the features based on the differences between clusters. By conducting MAD, the volume of regular mail, the volume of registered parcel, the volume of batch delivery of regular mail, and the volume of registered mail were significant in the descending order of MAD values. By conducting LFS, the volume of regular mail came first, followed by the volume of regular mail, the volume of registered mail, preparation + mail bundling, and the volume of registered parcel, with the value of 0.05 or above.

In addition, we conducted two feature extractions: principal component analysis (PCA) and independent component analysis (ICA). PCA transforms high-dimensional data into low-dimensional data, mapped into a new coordinate by arranging the axis of the variance as main components in descending order, from the largest to smallest. ICA predicts the independent component to maximize the statistical independence of the predicted component. We used PCA with the value of 4 for $n$ components and FastICA functions from sklearn for PCA and ICA, respectively.

Figure 1A and 1B show feature extraction results for PCA and ICA, respectively. In these figures, the $y$-axis represents $N$ original features and the $x$-axis represents $m$ (i.e., 4 in the both cases of PCA and ICA) newly created features. For PCA, the volume of regular mail, volume of registered parcel, volume of batch delivery of regular mail, and volume of express mail service (EMS) were shown significantly with the value of 0.5 or above. For ICA, the volume of batch delivery of regular mail, the volume of registered parcel, and the volume of EMS were shown considerably.

¹https://www.koreapost.go.kr/
After feature engineering, the selected or extracted features are used in the clustering phase. Here, four representative clustering algorithms were adopted: (i) \( k \)-means clustering, (ii) \( k \)-medoids clustering, (iii) hierarchical clustering, and (iv) Gaussian mixture model (GMM) to discover hidden patterns in the postal delivery area dataset.

\( k \)-means clustering is a popular unsupervised learning method. It forms clusters by repeatedly updating a set of center points using the average coordinates of the points assigned to each cluster until the center points do not change. Thus, this algorithm minimizes the distance difference between the center points and objects.

\( k \)-medoids clustering compensates for the shortcomings of the \( k \)-means algorithm, which are sensitive to initial centroids, noises, and outliers. Unlike the \( K \)-means algorithm, which uses the mean of the cluster as the center point, the \( k \)-medoids algorithm employs the median of the cluster as the center point, so it is more robust for noises and outliers. However, it is time-consuming compared with the \( k \)-means algorithm because the volume of the dataset increases due to its higher computational complexity.

### Table 1

<table>
<thead>
<tr>
<th>Attribute name</th>
<th>Data type</th>
<th>Average</th>
<th>Variance</th>
<th>Nonzero ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post office name</td>
<td>String</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Post office ID + postal delivery area ID</td>
<td>String</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Preparation + mail bundling</td>
<td>Double</td>
<td>0.9253</td>
<td>0.0299</td>
<td>99.93</td>
</tr>
<tr>
<td>Delivery + movement for delivery + extra work for delivery</td>
<td>Double</td>
<td>4.4388</td>
<td>3.0951</td>
<td>98.31</td>
</tr>
<tr>
<td>Mail bundling + extra work for mail not delivered</td>
<td>Double</td>
<td>1.0068</td>
<td>2.1262</td>
<td>99.98</td>
</tr>
<tr>
<td>The volume of batch delivery of regular mail</td>
<td>Double</td>
<td>140.6842</td>
<td>14 464.2054</td>
<td>88.79</td>
</tr>
<tr>
<td>The volume of batch delivery of registered mail</td>
<td>Double</td>
<td>28.1047</td>
<td>6085.1050</td>
<td>94.24</td>
</tr>
<tr>
<td>The volume of batch delivery of registered parcel</td>
<td>Double</td>
<td>5.3441</td>
<td>184.9478</td>
<td>95.32</td>
</tr>
<tr>
<td>The volume of batch delivery of express mail service (EMS)</td>
<td>Double</td>
<td>0.2305</td>
<td>0.6937</td>
<td>78.99</td>
</tr>
<tr>
<td>The volume of regular mail</td>
<td>Double</td>
<td>476.6402</td>
<td>72 774.1372</td>
<td>89.29</td>
</tr>
<tr>
<td>The volume of registered mail</td>
<td>Double</td>
<td>49.8830</td>
<td>862.8748</td>
<td>95.92</td>
</tr>
<tr>
<td>The volume of registered parcel</td>
<td>Double</td>
<td>21.7761</td>
<td>320.0171</td>
<td>96.77</td>
</tr>
<tr>
<td>The volume of express mail service (EMS)</td>
<td>Double</td>
<td>1.1216</td>
<td>7.0033</td>
<td>93.32</td>
</tr>
<tr>
<td>Total delivery distance</td>
<td>Double</td>
<td>69.2295</td>
<td>5189.3886</td>
<td>95.35</td>
</tr>
<tr>
<td>The number of super high-rise apartments in the postal delivery area</td>
<td>Double</td>
<td>99.6224</td>
<td>1 193 860.5820</td>
<td>15.41</td>
</tr>
<tr>
<td>The number of high-rise apartments in the postal delivery area</td>
<td>Double</td>
<td>518.2203</td>
<td>2 565 253.9410</td>
<td>42.17</td>
</tr>
<tr>
<td>The number of midrise apartments in the postal delivery area</td>
<td>Double</td>
<td>386.8931</td>
<td>1 231 481.0650</td>
<td>49.24</td>
</tr>
<tr>
<td>The number of low-rise apartments in the postal delivery area</td>
<td>Double</td>
<td>129.1428</td>
<td>167 379.8749</td>
<td>44.65</td>
</tr>
<tr>
<td>The number of detached houses in the postal delivery area</td>
<td>Double</td>
<td>579.1593</td>
<td>1 295 503.2860</td>
<td>90.22</td>
</tr>
<tr>
<td>The number of attached houses in the postal delivery area</td>
<td>Double</td>
<td>1216.8137</td>
<td>87 411.2646</td>
<td>77.89</td>
</tr>
<tr>
<td>The number of offices in the postal delivery area</td>
<td>Double</td>
<td>64.7854</td>
<td>87 411.2646</td>
<td>77.89</td>
</tr>
<tr>
<td>The number of studios in the postal delivery area</td>
<td>Double</td>
<td>156.8918</td>
<td>358 465.4691</td>
<td>39.14</td>
</tr>
<tr>
<td>Total number of buildings in the postal delivery area</td>
<td>Double</td>
<td>3151.5292</td>
<td>37 937 413.6400</td>
<td>93.71</td>
</tr>
</tbody>
</table>
FIGURE 1  Feature extraction results (Number of main components is 4): (A) principal component analysis (PCA) result and (B) independent component analysis (ICA) result
A parametric approach for unsupervised learning assumes that data are generated from several distributions with parameters. Among them, the GMM is the most popular and assumes that data are generated from \( k \) normal distributions. There are two parameters in GMM: (i) weights, indicating where an object probabilistically belongs among \( k \) normal distributions, and (ii) mean and variance of each normal distribution. The first parameter is called the latent variable and is obtained through the expectation-maximization algorithm. Upon completing the parameter estimation, clustering can be performed by classifying the data into appropriate normal distribution where the data belong.

Hierarchical clustering can be divided into two types: (i) agglomerative hierarchical clustering and (ii) divisive hierarchical clustering. Here, we used an agglomerative clustering algorithm for hierarchical clustering. The agglomerative hierarchical clustering sequentially and hierarchically integrates similar data or clusters. Unlike the \( k \)-means, \( k \)-medoids, and GMM, hierarchical clustering can learn without predetermining the number of clusters.

### 6 | QUANTITATIVE CLUSTER ANALYSIS

The quantitative cluster analysis was conducted for two main objectives: determining the number of clusters (\( k \)) and comparing the performance of feature-engineering methods and clustering algorithms.

Determining the number of clusters is an important issue in cluster analysis since the number of clusters significantly affects clustering quality. An increase in the number of clusters without penalty decreases the number of errors. For example, when the number of clusters equals data objects, every data object belongs to its cluster, and no error exists between clusters and data objects. Contrarily, a decrease in the number of clusters increases information loss. For example, when the number of clusters equals one, every data object belongs to the same cluster, and information loss is maximized. In such an extreme case, we cannot find any cluster tendency.

The silhouette coefficient, a widely used measure of the similarity of a data object to its cluster, was used for the quantitative cluster analysis. It ranges from \(-1\) to \(1\), where the higher value represents better clustering quality. We also used elbow graphs, dendrograms, and Dunn indexes to assess the clustering results. The results from those metrics were removed, however, because they revealed a similar pattern to the silhouette coefficient finding.

Table 2 shows the results of the quantitative cluster analysis using the silhouette coefficient. We considered 16 combinations depending on the clustering algorithm adopted and the feature-engineering method used. Then, we varied the number of clusters for each case. First, the values of the silhouette coefficient tend to be highest when the number of clusters is 5, 6, or 8. For clustering algorithms, the silhouette coefficient is highest when the \( k \)-means algorithm is used, then seconded by hierarchical clustering algorithm. For feature-engineering methods, the silhouette coefficient is higher when feature extraction methods, such as PCA or ICA, are used rather than feature selection methods, such as MAD or LFS. Among 16 combinations, the best result was recorded in the use of the \( k \)-means algorithm, ICA feature extraction method, and five clusters. In this case, the value of the silhouette coefficient was 0.359. The second best case was when the \( k \)-means algorithm and PCA feature extraction method were used with six clusters. Here, the value of the silhouette coefficient was 0.358. For each clustering algorithm, the values of the highest and second-highest silhouette coefficients were marked in bold.

### 7 | QUALITATIVE CLUSTER ANALYSIS

In addition to quantitative analysis using the evaluation metric, such as the silhouette coefficient, the qualitative analysis also discusses the characteristics and semantics of discovered clusters since it allows us to better understand the postal delivery area data. To this end, we visualized two cases with the best clustering results in quantitative evaluation using box plots to show each feature distributed in different clusters. In this quantitative cluster analysis, we have three objectives: (i) feature analysis by cluster, (ii) regional distribution by cluster, and (iii) comparison of two cluster sets derived by two clustering algorithms.

Figure 2 shows the box plot visualization of the results of \( k \)-means clustering. Here, the ICA feature extraction method was used with five clusters. Interestingly, for the feature distribution, \( \text{delivery total time, preparation+mail bundling, delivery+movement for delivery + extra work for delivery, and total delivery distance} \), which have been used as traditional indicators for the characteristics of postal delivery area data, showed little difference by clusters compared with other features.

For the characteristics of each cluster, cluster 1 is a group of postal delivery areas where \( \text{the volume of batch delivery of EMS, volume of EMS, number of super high-rise apartments, and number of high-rise apartments} \) show low values. Cluster 2 is a group of postal delivery areas where \( \text{mail bundling + extra work for mail not delivered, volume of regular mail, and volume of registered mail show high values, whereas volume of batch delivery of registered mail, volume of batch delivery of registered parcel,} \)
TABLE 2  Results of quantitative cluster analysis

<table>
<thead>
<tr>
<th></th>
<th>$k = 5$</th>
<th>$k = 6$</th>
<th>$k = 7$</th>
<th>$k = 8$</th>
<th>$k = 9$</th>
<th>$k = 10$</th>
<th>$k = 11$</th>
<th>$k = 12$</th>
<th>$k = 13$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$-means</td>
<td>MAD 0.347</td>
<td>0.318</td>
<td>0.303</td>
<td>0.302</td>
<td>0.305</td>
<td>0.299</td>
<td>0.278</td>
<td>0.276</td>
<td>0.272</td>
</tr>
<tr>
<td></td>
<td>LFS 0.331</td>
<td>0.295</td>
<td>0.288</td>
<td>0.276</td>
<td>0.282</td>
<td>0.275</td>
<td>0.251</td>
<td>0.248</td>
<td>0.248</td>
</tr>
<tr>
<td></td>
<td>PCA 0.344</td>
<td><strong>0.358</strong></td>
<td>0.324</td>
<td>0.309</td>
<td>0.310</td>
<td>0.312</td>
<td>0.307</td>
<td>0.284</td>
<td>0.284</td>
</tr>
<tr>
<td></td>
<td>ICA <strong>0.359</strong></td>
<td>0.296</td>
<td>0.295</td>
<td>0.303</td>
<td>0.306</td>
<td>0.310</td>
<td>0.295</td>
<td>0.295</td>
<td>0.277</td>
</tr>
<tr>
<td>$k$-medoids</td>
<td>MAD 0.246</td>
<td>0.252</td>
<td>0.245</td>
<td>0.236</td>
<td>0.201</td>
<td>0.272</td>
<td>0.242</td>
<td>0.237</td>
<td>0.236</td>
</tr>
<tr>
<td></td>
<td>LFS 0.266</td>
<td>0.228</td>
<td>0.185</td>
<td>0.199</td>
<td>0.244</td>
<td>0.232</td>
<td>0.217</td>
<td>0.200</td>
<td>0.217</td>
</tr>
<tr>
<td></td>
<td>PCA 0.233</td>
<td>0.247</td>
<td><strong>0.279</strong></td>
<td>0.263</td>
<td>0.250</td>
<td>0.236</td>
<td>0.247</td>
<td>0.238</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ICA 0.223</td>
<td>0.230</td>
<td>0.256</td>
<td><strong>0.291</strong></td>
<td>0.237</td>
<td>0.204</td>
<td>0.235</td>
<td>0.208</td>
<td>0.222</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>MAD 0.299</td>
<td>0.292</td>
<td>0.251</td>
<td>0.254</td>
<td>0.228</td>
<td>0.227</td>
<td>0.228</td>
<td>0.230</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td>LFS 0.288</td>
<td>0.221</td>
<td>0.220</td>
<td>0.228</td>
<td>0.207</td>
<td>0.196</td>
<td>0.199</td>
<td>0.193</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>PCA 0.310</td>
<td><strong>0.319</strong></td>
<td>0.238</td>
<td>0.251</td>
<td>0.217</td>
<td>0.224</td>
<td>0.215</td>
<td>0.218</td>
<td>0.221</td>
</tr>
<tr>
<td></td>
<td>ICA <strong>0.312</strong></td>
<td>0.274</td>
<td>0.236</td>
<td>0.241</td>
<td>0.217</td>
<td>0.221</td>
<td>0.223</td>
<td>0.224</td>
<td>0.209</td>
</tr>
<tr>
<td>GMM</td>
<td>MAD 0.231</td>
<td>0.191</td>
<td>0.156</td>
<td>0.151</td>
<td>0.150</td>
<td>0.146</td>
<td>0.132</td>
<td>0.119</td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td>LFS 0.213</td>
<td>0.213</td>
<td>0.228</td>
<td>0.219</td>
<td>0.133</td>
<td>0.157</td>
<td>0.141</td>
<td>0.138</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>PCA 0.185</td>
<td><strong>0.238</strong></td>
<td>0.185</td>
<td>0.171</td>
<td>0.159</td>
<td>0.149</td>
<td>0.119</td>
<td>0.141</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>ICA <strong>0.275</strong></td>
<td>0.214</td>
<td>0.216</td>
<td>0.238</td>
<td>0.251</td>
<td>0.243</td>
<td>0.244</td>
<td>0.235</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Note: For each clustering algorithm, the values of the highest and second-highest silhouette coefficients were marked in bold.
Abbreviations: GMM, Gaussian mixture model; ICA, independent component analysis; MAD, mean absolute deviation; PCA, principal component analysis; LFS, localized feature selection.

FIGURE 2  Box plot visualization of $k$-means clustering results: (A) delivery total time, (B) preparation + mail bungling, (C) delivery + movement for delivery + extra work for delivery, (D) mail bundling + extra work for mail not delivered, (E) the volume of batch delivery of regular mail, (F) the volume of batch delivery of registered mail, (G) the volume of match delivery of registered parcel, (H) the volume of match delivery of express mail service (EMS), (I) the volume of regular mail, (J) the volume of registered mail, (K) the volume of registered parcel, (L) the volume of EMS, (M) total delivery distance, (N) the number of super high-rise apartments in the postal delivery area, and (O) the number of high-rise apartments in the postal delivery area.
volume of batch delivery of EMS, and volume of EMS show low values. In cluster 3, the values of the volume of batch delivery of regular mail, volume of batch delivery of registered mail, volume of regular mail, and volume of registered mail are low; however, the value of the volume of registered parcel is the highest compared with other clusters. In cluster 4, both volumes of batch delivery of regular mail and batch delivery of registered mail show the highest values and volumes of regular mail and registered mail show the second-highest values. Cluster 5 is the most distinctive group, such that the volume of batch delivery of EMS, volume of EMS, number of super high-rise apartments, and number of high-rise apartments have uniquely high values compared with other clusters.

It is difficult to explain all clusters only by the difference in regional distribution since the characteristics of postal delivery areas can be different even in the same administrative district. However, some clusters show distinct regional differences. Cluster 1 has many postal delivery areas located in small- and medium-sized administrative districts, such as Geochang, Yeongcheon, Miryang, Jeonju, Nonsan, and Geumsan compared with other clusters. This resonates with describing cluster 1 with low values of the volume of batch delivery of EMS, volume of EMS, number of super high-rise apartments, and number of high-rise apartments. However, cluster 5 has many postal delivery areas located in Seoul or metropolitan administrative districts, such as Seoul Gangnam, Seoul Seocho, Seoul Central, Seoul Dobong, and Seongnam Bundang, compared with other clusters. This agrees with describing cluster 5 with extremely high values of the volume of batch delivery of EMS, the volume of EMS, number of super high-rise apartments, and the number of high-rise apartments.

Figure 3 shows the box plot visualization of hierarchical clustering results. Here, the PCA feature extraction method was used with six clusters. Contrary to the results of the k-means clustering, only the total delivery distance shows the difference by clusters for feature extraction compared with other features. The rest of the features show a distinctly different distribution in at least one cluster.

Furthermore, we found interesting results for the characteristics of other clusters. The characteristics of the results of clusters 1, 2, 5, and 6 of the hierarchical clustering are similar to the characteristics of clusters 1, 2, 4, and 5 of the result of k-means clustering. From a close investigation of the postal delivery areas belonging to each cluster, the results of clusters 1, 2, 5, and 6 of the hierarchical clustering highly overlap with those of clusters 1, 2, 4, and 5 of the k-means clustering. Also, the result of cluster 3 of the k-means clustering highly overlaps with the union of those of clusters 3 and 4 in hierarchical clustering. Considering that the number of clusters in hierarchical clustering is one more than that in the k-means clustering, this is an interesting finding arising due to the hierarchical nature of the adopted clustering algorithm.

8 | IMPLICATION

Our findings show that our clustering approach provides practical design principles to mitigate workload imbalance among postal delivery areas. Manual redistribution of the delivery workload is not the best way to handle the workload balancing since it can be misused for productivity evaluation. The results of feature engineering show that several features need to be carefully considered crucial for differentiating between clusters.

The number k is a sensitive measure, as it may potentially affect additional staffing or workload redistribution policy. If k is too small, postal delivery areas that belong to minor but distinct clusters would be wrongly included in major clusters, thus, potentially overlooking them during the decision-making process for workload balancing. However, if it is too large, the data-driven approach becomes ineffective. It is worth noting that 5, 6, and 8 are the best values for k by our quantitative analysis; however, the final value should be carefully determined after the consensus among the postal delivery stakeholders is reached.

Our qualitative analysis shows a distinct regional difference between clusters 1 and 5 representing rural and urban postal delivery areas, respectively. However, the other clusters are more subtle to be distinguished. Thus, a radar chart, which displays multivariate data in the form of multiple quantitative features represented on axes starting from the same point, is suitable for displaying a postal delivery area. With the charts, comparison can be made quickly and outliers, which further need to be investigated in terms of workload estimation, become easily noticeable [31]. Moreover, we proposed a general description of each cluster by qualitative analysis to enhance the understanding of clustering results for end-users who handle additional staffing.

Finally, the clustering findings can be used by end-users in charge of workload rebalancing as a useful signal. In the post office in Gapyeong, South Korea, for example, there are 18 postal delivery regions. Prior to the unsupervised approach, all postal delivery zones were regarded similarly when computing workload scores, although each area had its unique perspective, resulting in often unjust workload redistribution findings. With the proposed approach, each area’s relative work intensity is measured by comparing the area with the average value of the cluster to which the area belongs. We can acquire the following benefits as a result of doing
so. First, we can obtain the highest work intensity delivery zones for each post office. Second, as long as two post delivery zones in separate post offices belong to the same cluster, it becomes possible to compare them. It should be stressed that human managers should make the final choice on workload rebalancing and additional manpower, as the clustering results might be highly distorted by sudden changes in delivery data. Nonetheless, the suggested method yields encouraging results in terms of maximizing the potential of key components of data-centric workload estimate for improving postal last-mile delivery.

9 | CONCLUSION

In this study, we investigated the postal delivery area dataset in the Republic of Korea and differentiated the areas by conducting multiple clustering methods. By grouping similar areas, a data-driven approach for workload estimation becomes possible by comparing one area to other similar areas. Because the data for a postal delivery area consist of various features that affect the postal delivery, we conducted data preprocessing and feature engineering to discover how each feature affects determining the postal delivery area of a cluster.

It is effective in discovering more latent information by exploiting graph clustering methods in recent studies [32,33]. To do so, the data can be converted into a graph, with nodes representing postal delivery zones and edges constructed by comparing the similarity of two nodes. We will continue to work on this in the future.

From our quantitative and qualitative analyses, we conclude that unsupervised learning methods are well fitted for grouping postal delivery areas. Furthermore, workload balancing can be performed by comparing one area to others in the same cluster. The identified clusters were differentiated by not only regional features but also several intrinsic features, indicating that the comparison between the postal delivery area should not be performed using an ad hoc approach.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.
Evaluating 


Keejun Han received his BS degree in Information Technology from the Australian National University, Canberra, Australia, in 2010. In 2013 and 2019, he earned his MS and Ph.D. degrees in Knowledge Service Engineering from KAIST in Daejeon, Republic of Korea. He worked for Samsung Electronics in Suwon, Republic of Korea, from 2019 until 2021. He has been a researcher at Daejeon’s Electronics and Telecommunications Research Institute (ETRI) since 2021. His research interests include information retrieval, recommender systems, and big data analysis on logistic data.

Yeongwoong Yu received his PhD degree in Industrial Engineering from Pusan National University, Busan, Republic of Korea, in 2017. He has worked as a researcher at ETRI in Daejeon, Republic of Korea, since 2011. Intelligent logistics process management, postal logistics, logistics process mining, and GIS-based logistics process execution are some of his research interests.

Dong-gil Na received his BS, MS, and Ph.D degrees in Industrial Engineering from Chonbuk National University, Chonbuk, Republic of Korea. He has been with the ETRI, Daejeon, Republic of Korea, since 2005, where he is presently a chief researcher, in the postal & logistics technology research center. Big data analysis on logistic data and recommendation systems are two of his research interests.

Hoon Jung received his PhD degree in Industrial Engineering from University of Missouri, Columbia, USA, in 2001. In 2002, he became a member of ETRI. Since 2017, he has served as the head of the Postal Logistics Technology Research Center, where he is in charge of drone delivery research and development.

Younggyo Heo received the BS degree in Computer Science and Engineering from Kongju National University (KNU), Chungnam, Republic of Korea. He is a research associate at the Data Intelligence Laboratory, KNU. His research interests include data mining and graph mining.

Hyeonchoel Jeong received the BS degree in Computer Science and Engineering from Kongju National University, Chungnam, Republic of Korea. He is a research associate at the Data Intelligence Laboratory, KNU. His research interests include artificial intelligence and machine learning.

Sunguk Yun received the BS degree in Computer Science and Engineering from Kongju National University, Chungnam, Republic of Korea. He is currently pursuing an MS degree at the Department of Software, KNU. His research interests include artificial intelligence and machine learning.

Jungeun Kim received the PhD degree in knowledge service engineering from KAIST. He is an assistant professor of computer science and engineering with Kongju National University (KNU). Before joining KNU, he was a senior researcher of the artificial intelligence research laboratory with Electronics and Telecommunications Research Institute (ETRI). His research interests include data mining, artificial intelligence, big data analysis with distributed processing platforms, and open data platforms.

How to cite this article: K. Han, Y. Yu, D. Na, H. Jung, Y. Heo, H. Jeong, S. Yun, and J. Kim, Understanding postal delivery areas in the Republic of Korea using multiple unsupervised learning approaches, ETRI Journal 44 (2022), 232–243. https://doi.org/10.4218/etrij.2021-0407