Constructing Transit Origin-Destination Matrices using Spatial Clustering

Luo, Ding; Cats, Oded; van Lint, Hans

DOI
10.3141/2652-05

Publication date
2017

Document Version
Final published version

Published in
Transportation Research Record

Constructing Transit Origin–Destination Matrices with Spatial Clustering

Ding Luo, Oded Cats, and Hans van Lint

So-called tap-in–tap-off smart card data have become increasingly available and popular as a result of the deployment of automatic fare collection systems on transit systems in many cities and areas worldwide. An opportunity to obtain much more accurate transit demand data than before has thus been opened to both researchers and practitioners. However, given that travelers in some cases can choose different origin and destination stations, as well as different transit lines, depending on their personal acceptable walking distances, being able to aggregate the demand of spatially close stations becomes essential when transit demand matrices are constructed. With the aim of investigating such problems using data-driven approaches, this paper proposes a *k*-means-based station aggregation method that can quantitatively determine the partitioning by considering both flow and spatial distance information. The method was applied to a case study of Haaglanden, Netherlands, with a specified objective of maximizing the ratio of average intra-cluster flow to average inter-cluster flow while maintaining the spatial compactness of all clusters. With a range of clustering of different *K* performed first using the distance feature, a distance-based metric and a flow-based metric were then computed and ultimately combined to determine the optimal number of clusters. The transit demand matrices constructed by implementing this method lay a foundation for further studies on short-term transit demand prediction and demand assignment.

Transit demand studies and models form an essential part of any transit planning process. The purpose of such research is to estimate and evaluate passenger demand by using models and by collecting and analyzing data pertaining to current and future transit needs (*1*). Traditionally, a sequential four-step process has been extensively used in both academia and practice to estimate the aggregated travel demand for a number of traffic analysis zones that are predetermined on the basis of geographical and socioeconomic factors. The share of transit demand is then computed at the step of modal split or mode choice, using discrete choice models. This four-step method provides researchers and practitioners with a straightforward way to obtain transit demand when such demand can hardly be observed directly, although the results cannot always be as accurate as is desirable. However, with a smart card automated fare collection (AFC) system being adopted by more and more transit agencies, a new type of data source is rapidly becoming available. AFC systems record individual travelers' boarding and alighting information, greatly facilitating research on passenger travel patterns that can support transit network planning, behavioral analysis, and transit demand estimation and forecasting (*2*).

A large amount of research effort has been directed to transit origin–destination (O-D) estimation, especially for cases where only entry information or exit information is available, or even where neither of them is available. Different methods have been proposed to infer O-D matrices for transit journeys with limited boarding or alighting information (*3–10*), and these methods can be categorized on the basis of their estimation assumptions, including walking distances (buffer zones), transfer times, and last-destination assumptions (*11*). Along with an increase in the number of methods, the importance of evaluation and validation of the O-D estimation methods and results has also been highlighted in a series of studies (*12–16*). Recently, Alsger et al. used a high-quality data set containing accurate boarding and alighting information to validate a multi-leg journey inference method, where the alighting information is assumed to be unknown for validation's sake (*11*).

The aforementioned O-D estimation studies focused on attaining a more accurate journey inference; less effort was directed toward stop or station aggregation while constructing demand matrices. Stop or station aggregation, in this context, means that transit users' activities of originating from, or alighting at, an individual stop can be virtually associated with an area that covers a number of adjacent transit stops or stations (*17*). In this sense, demand at a more aggregate level can be of more practical use to both transit researchers and practitioners. McCord et al. pointed out that the size of stop-to-stop O-D matrices makes it difficult to synthesize important flow patterns and to estimate stop-to-stop O-D passenger flows accurately (*18*). By grouping transit stops, however, the estimation, analysis, and communication of passenger flows can be improved. Further, understanding transit demand at an aggregate level was also motivated by Lee et al., who highlighted that the ability to define a specific land use type and the temporal characteristics related to passengers' activities can be enhanced through stop aggregation (*17*). The aggregation of stops is also in line with the analysis and modeling of transit users' stop choice behavior, which has been recently explored by Hassan et al. and Nassir et al. (*19, 20*). The rationale is that in reality, travelers are very often capable of choosing from a set of origin and destination stations that are within acceptable walking distance. As a result of such behavior, different choices in transit services (modes and lines) can be characterized in terms of travel demand from one area to another.

A limited number of studies involving stop aggregation can be identified in the current literature. For instance, Chu and Chapleau used the term "anchor points" to define the places that a person repeatedly visits, which usually include residence, work, or study locations (*21*). They performed spatial aggregation by grouping stops within 50 m of each other to form a new node. A so-called stop-aggregation model was later proposed by Lee et al. and applied to studying the metro

Department of Transport and Planning, Faculty of Civil Engineering and Geoscience, Delft University of Technology, Stevinweg 1, P.O. Box 5048, 2600 GA Delft, Netherlands. Corresponding author: D. Luo, D.Luo@tudelft.nl.

Transportation Research Record: Journal of the Transportation Research Board, No. 2652, 2017, pp. 39–49.
http://dx.doi.org/10.3141/2652-05

transit of the Minneapolis–Saint Paul metropolitan area in Minnesota (*22, 23*). This model aims to develop a generalized definition of a "stop" that more closely matches the nature of locations serving as passenger origins and destinations. An aggregated area around a transit stop or station can thus be defined by three parameters: (*a*) distance or proximity, measured by using Euclidean and network distances in geographic information systems; (*b*) text in the description of the stop, queried using database tools in SQL; and (*c*) the catchment area, which is defined as how a stop is related to the land uses surrounding it. Alsger et al. simply aggregated the estimated O-D trips according to the 1,515 zones in the Brisbane Strategic Transport Model to provide an overview of the results (*12*). McCord et al. proposed two computationally efficient heuristic algorithms to aggregate bus stops at the route level to reduce the size of the O-D matrix for improved estimation, analysis, and communication (*18*). More recently, Tamblay et al. developed a methodology to estimate a zonal O-D matrix for a transit system (*24*). On the basis of a stop-to-stop O-D matrix created with smart card data from Santiago de Chile, a logit model was constructed to compute the probability that an observed trip using transit stops k and l (as the boarding and alighting points, respectively) was originated in zone i and ended in zone j. In the methodology of Tamblay et al., a zonal system must be predefined and a survey is required to help identify the model parameters.

Unsupervised learning techniques have recently been employed to investigate spatial travel pattern and demand, given their natural advantages in solving clustering problems (*6, 25, 26*). One of the successful applications turns out to be the identification of individual transit riders' spatial and temporal travel patterns using the density-based spatial clustering of applications with noise (DBSCAN) algorithm (*6, 25, 26*). This specific algorithm stands out in its flexibility; it does not require predetermining the number of clusters, and it identifies arbitrarily shaped clusters while it is being implemented. Ma et al. first applied this algorithm to examine the spatial travel pattern of transit users in Beijing after inferring individuals' journey chains (*6*). Adopting the approach proposed by Ma et al., Kieu et al. later performed a two-level approach that relied on the standard DBSCAN algorithm to reveal both spatial and temporal travel patterns in Southeast Queensland, Australia (*26*). They further improved the efficiency of this approach by using the existing knowledge of individual travel patterns while clustering the studied journey, developing a so-called weighted-stop DBSCAN (WS-DBSCAN) algorithm (*25*).

The above-mentioned studies highlight the importance of stop or station aggregation in analyzing AFC data. In many cases where transit services are well provided in both urban and suburban areas, in terms of the number of stations and lines—for example, in Haaglanden, the conurbation surrounding The Hague in the Netherlands (Figure 1) (dots represent stations)—the data on traveler O-D flow from one area to another, both of which areas contain a group of bus or tram stations, would be much more usable in transit modeling, prediction, and management than the data at the individual stop level. This paper proposes a k-means-based station aggregation methodology that can in four steps quantitatively determine the clustering by considering both flow and spatial distance information.

This method was applied to a case study of Haaglanden, Netherlands, by specifying a criterion that the ratio of average intra-cluster flow to average inter-cluster flow should be maximized while maintaining the spatial compactness of all clusters. In the first step, a number of different clustering scenarios were obtained by implementing the standard k-means algorithm with the geodesic distance considered as the only feature. Then, two metrics that represented spatial distance and passenger flow, respectively, were computed, and finally integrated to determine the optimal number of clusters. The proposed data-driven method allowed researchers to obtain clusters that are on one hand sufficiently large to enable the consideration and modeling of travel alternatives between parts of the network, and on the other hand are compact enough to include only viable alternatives and support fine-grained demand estimation. Unlike the standard DBSCAN algorithm that can identify some points as "noise"—points that are then not included in any of the clusters—the proposed method ensures that all travel information is retained in the O-D matrix attained.

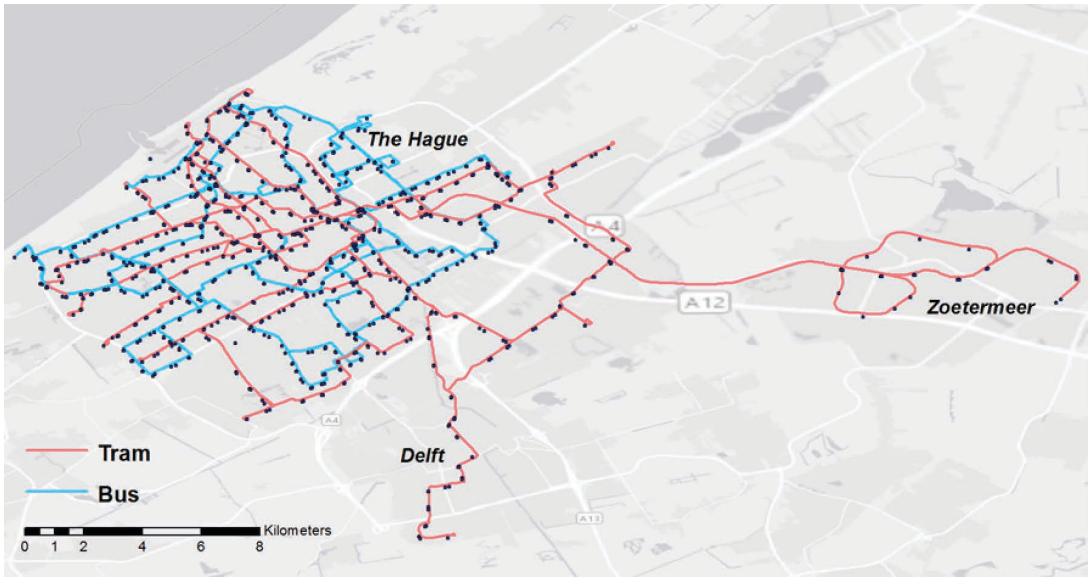


FIGURE 1 Tram and bus lines operated by HTM in Haaglanden, Netherlands.

The next section of this paper presents the data preparation for the proposed methodology from two aspects, including an introduction of the Dutch smart card data and the construction of valid multi-leg journeys. That is followed by a section that details the proposed four-step *k*-means-based methodology. The results are then presented, with both spatial and temporal variability analyses included. The paper concludes with a discussion and suggestions for future research.

DATA PREPARATION

Data Overview

The smart card data set used in this study comes from Haaglanden, the conurbation surrounding The Hague in the Netherlands. It contains data collected from both buses and trams that are operated by Haagsche Tramweg Maatschappij (HTM), the local transit operating company. As Figure 1 depicts, the transit network organized by HTM in this area consists of 12 tram and eight bus lines serving 931 stations. A more detailed description of the Dutch smart card system, the OV-Chipkaart, can be found in a study conducted by Van Oort et al. (*27*). The original transaction data are from March 2015 and contain 8,177,434 records (i.e., tap-in–tap-out pairs), covering 31 days.

Construction of Multi-Leg Journeys

Multi-leg journeys using this data set were constructed by Bagherian et al. (*28*). The procedure started with excluding data that contain missing values (e.g., records that were missing tap-in or tap-out or had no line identifier). Three types of transactions were subsequently removed, including those with identical location of tap-in and tap-out; those between stops *i, j* with $t^o_{n,i} - t^c_{n,j} < \gamma^{\text{leg}}_{\min}$, where $t^o_{n,i}$ and $t^c_{n,j}$ represent the tap-in and tap-out time of a card of identification (ID) *n* at station *i* and *j,* respectively, and $\gamma^{\text{leg}}_{\min}$ denotes the minimum duration of a leg; and those with abnormally long durations $t^o_{n,i} - t^c_{n,j} > \gamma^{\text{leg}}_{\max}$, where $\gamma^{\text{leg}}_{\max}$ denotes the maximum duration. After this step, transactions were grouped using card ID for each day in the analysis period, and within each group the transactions were also sorted using a check-in time-stamp. Finally, an iterative procedure chained transactions forming a journey if $t^c_{r_n} - t^o_{r-1_n} < \gamma^{\text{transfer}}$, where $t^c_{r_n}$ and $t^o_{r-1_n}$, respectively, represent the tap-in time of transaction *r* and the tap-out time of transaction $r-1$; γ^{transfer} denotes the time interval between two successive legs with same card ID. In this case study, $\gamma^{\text{leg}}_{\min}$, $\gamma^{\text{leg}}_{\max}$, and γ^{transfer} were set to be 1 min, 60 min, and 35 min, respectively. Once a journey was formed, transfer times and the number of transfers were also computed, and the journey was added to the database. Consequently, 6,255,798 journeys in the analysis period were generated. The output of this procedure was a database of the identified journeys, including an ID, date, number of transfers, and a list of details (line ID, tap-in time and location, and tap-out time and location) for all legs.

An issue regarding the validity of journey identification was observed while analyzing the data set—some journeys are unreasonably long (i.e., several hours). These "noise"-inferred multi-leg journeys were presumably caused by short activity chaining and were removed from the data set by adopting a threshold of a maximum of 90 min for a journey. This value was determined on the basis of the maximum time that a person needs to spend reaching his or her destination in Haaglanden. As a result of this cleaning process, 14,794 journeys were removed, which left 99.76% of the original records for further analysis.

k-MEANS-BASED STATION AGGREGATION METHOD

A Four-Step *k*-Means-Based Method

As one of the simplest and most popular clustering techniques, the *k*-means algorithm has been extensively leveraged in various fields since it was first proposed in 1967 (*29*). Given a set of *n* observations ($\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$), each of which is a *d*-dimensional real vector, this clustering algorithm aims to partition the *n* observations into $K (\leq n)$ mutually exclusive and collectively exhaustive clusters $\mathbf{C} = \{C_1, C_2, \ldots, C_K\}$. It iteratively determines the center $\boldsymbol{\mu}_i$ for each cluster C_i and assigns each observation to a cluster whose center is closest to the observation. This iterative clustering process terminates when the assignments no longer change, which can be described to minimize the within-cluster sum of squares (the sum of distance functions of each observation in the cluster C_i to the center $\boldsymbol{\mu}_i$):

$$\underset{\mathbf{C}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 \tag{1}$$

Details on the implementation of the *k*-means algorithm can be found in Tou and Gonzalez (*30*). The algorithm's main disadvantage is that the number of clusters, *K,* must be supplied as a parameter. In this study, a four-step *k*-means-based station aggregation method is proposed, in which a quantitative way to determine the optimal *K* is incorporated (Figure 2).

The method starts with finding the best clustering on the basis of a chosen measure for each *K,* and then continues with the computation of two metrics that are related to spatial distance and passenger flow. In the final step, two metrics are integrated for the determination of the optimal number of clusters, *K**. Such a method is flexible as it can accommodate different formulations of both metrics and final integration function in order to cater different purposes pertaining to the construction of transit O-D matrix. The essential idea, however, is to maximize either the intra-cluster or the inter-cluster flow while maintaining the spatial compactness of all clusters simultaneously.

k-Means-Based Clustering

Given that the clusters of transit stations should be spatially compact, the geodesic distance between points, which can be calculated based on their coordinates, was used as the only feature in the *k*-means clustering. While the *k*-means algorithm was implemented, a set of *K* points were input as the initial cluster centers so that the algorithm could proceed with iterations. Because the result of the *k*-means algorithm can vary given different initial centers, a common way to obtain better and reproducible results is to perform the algorithm a number of times with different initial centers and then select the initial centers that produce the optimal clustering in terms of the adopted measure. In this study, a measure called sum of the squared error (SSE) was employed to help select the initial centers because it can reflect the quality of a clustering—the lower SSE is, the better the clustering. The SSE was defined as follows for the current case:

$$\text{SSE}(K) = \sum_{i=1}^{K} \sum_{x \in C_i} d^2_{\mu_i, x} \tag{2}$$

where $d_{\mu_i, x}$ denotes the geodesic distance between a station and the cluster center to which it belongs. The *k*-means algorithm was

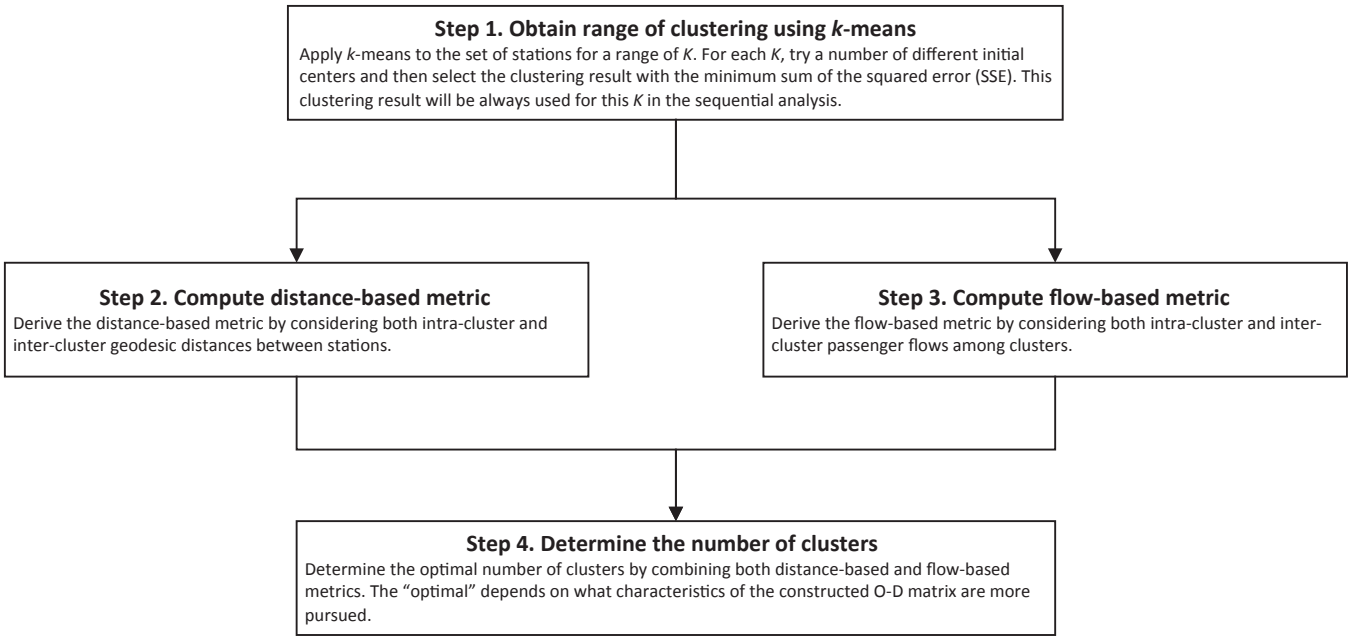


FIGURE 2 Steps of the proposed *k*-means-based station aggregation methodological framework.

programmed in Python 2.7 and the implementation process was as follows. A number of randomly generated sets of initial centers were tested for each K (ranging from 2 to 30) in the current study, and the initial center scenario that resulted in the minimum SSE was eventually selected and fixed for this K in the sequential analysis. Given the particular spatial distribution of stations in the study area (i.e., most stations are in the core area of The Hague, with others scattered in relatively isolated areas), six subareas, including Delft, Zoetermeer, and areas to the northeast, northwest, southwest, and southeast of The Hague were set up. When the number of clusters K was larger than five, two initial center points were randomly generated from the Delft and Zoetermeer subareas; the rest would also be generated from The Hague subareas. By doing so, the efficiency of implementing the *k*-means algorithm for a great number of iterations was dramatically improved. It is still worth mentioning that it can be time consuming to complete all clustering experiments for large values of K (>25). This issue can be further resolved by optimizing the *k*-means program. After obtaining all clustering results for different Ks, the subsequent analysis was performed with MATLAB (*31*).

Distance-Based Metric

The construction of the distance-based metric adopted the approach proposed by Ray and Turi (*32*). It examined the spatial compactness of a clustering by taking into consideration both intra-cluster and inter-cluster distance measures. The former computes the square of distance between a point and its cluster center, and then takes the average of all of them, denoted by D^{intra}:

$$D^{\text{intra}} = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i} d^2_{\mu_i,x} \tag{3}$$

where N is the number of stations in the study area.

The inter-cluster distance measure, D^{inter}, on the other hand, only takes the square of minimum distance between cluster centers because as long as the minimum of such distance is maximized, the others will by definition be larger than it. This measure is defined as follows:

$$D^{\text{inter}} = \min d^2_{\mu_i,\mu_j} \qquad \forall j \neq 1 \tag{4}$$

The two measures are then combined by taking the ratio as follows:

$$\tau = \frac{D^{\text{intra}}}{D^{\text{inter}}} \tag{5}$$

where τ denotes the final distance-based metric. To obtain the optimal number of clusters in terms of spatial compactness, τ is minimized; the intra-cluster distance measure D^{intra} in the numerator should be minimized, and the inter-cluster distance measure D^{inter} in the denominator should be maximized.

Flow-Based Metric

The passenger flow at the station level can first be derived from the original data set and then aggregated on the basis of a specific clustering. The flow-based metric provides additional information that can be used to determine the optimal number of clusters. Intuitively, total intra-cluster flow decreases as the number of clusters grows, given the constant total flow over the entire study period. More flows are naturally assigned to the inter-cluster group (Figure 3*b*).

When considering the flow information, it is possible to seek to maximize the total inter-cluster flow over the total intra-cluster

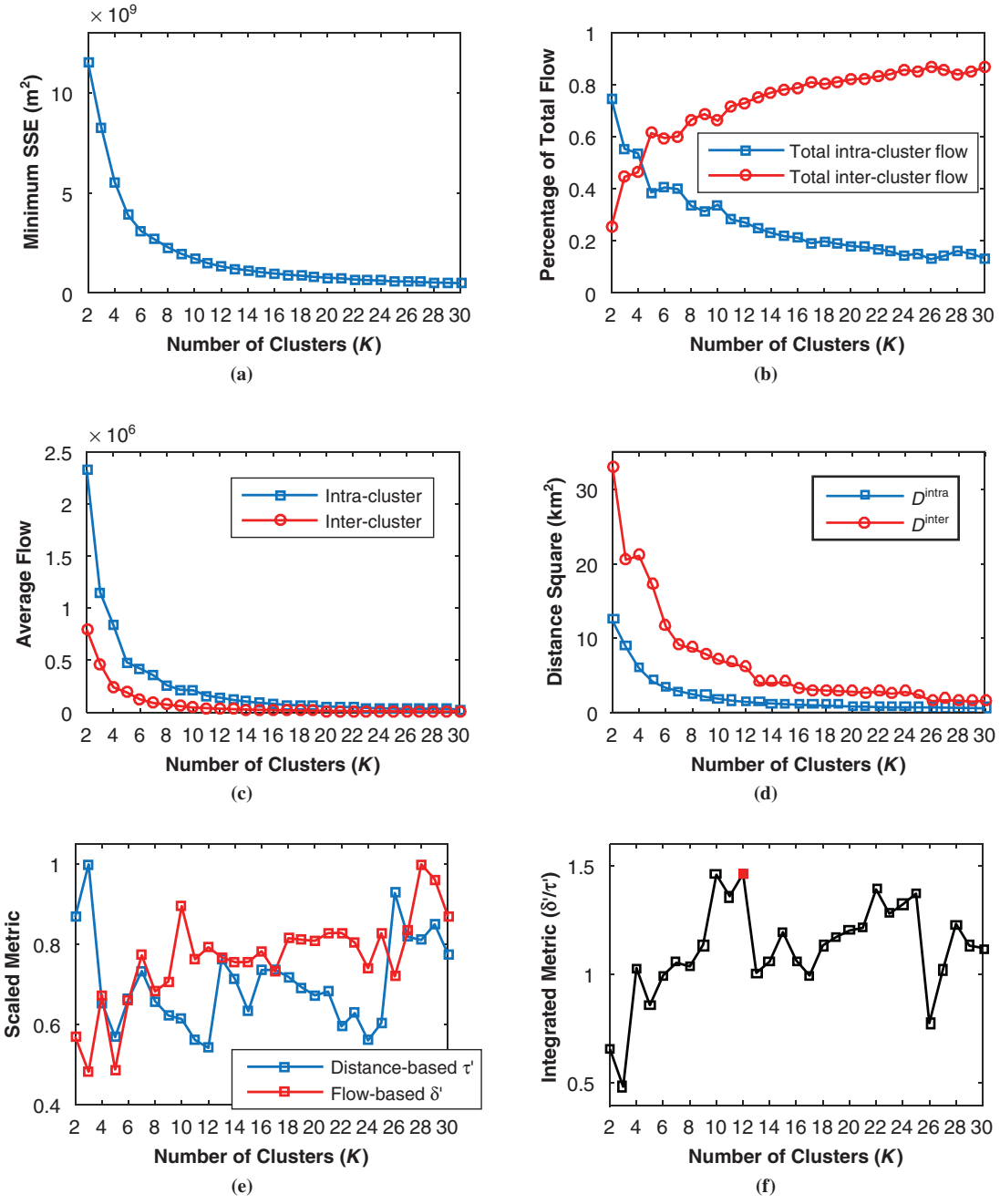


FIGURE 3 Intra- and inter-cluster flows: (*a*) SSE decreases exponentially as the number of clusters increases, (*b*) variation in both total intra-cluster and total inter-cluster flows, (*c*) variation in both average intra-cluster and average inter-cluster flows, (*d*) intra-cluster and inter-cluster flow measures, (*e*) illustration of two scaled metrics, and (*f*) integrated metric that reaches the maximum value when the number of the cluster is equal to 12.

one, or vice-versa, depending on the application and the analysis objectives. An argument in favor of the former is that it leads to more flow being assigned as inter-cluster (nondiagonal) elements in the O-D matrix. In contrast, by making the intra-cluster flow more significant, most self-contained and coherent clusters in terms of travel demand (diagonal elements) can be obtained, which is more desirable from a planning perspective. In the case of Haaglanden, Netherlands, the second option was eventually adopted and the following two flow measures were proposed:

$$F^{\text{intra}} = \frac{1}{K}\sum_{i=1}^{K}\sum_{x_m,x_n\in C_i} f_{x_m,x_n} \tag{6}$$

$$F^{\text{inter}} = \frac{1}{K^2-K}\sum_{i=1}^{K}\sum_{j=i}^{K}\sum_{x_m\in C_i,x_n\in C_j,\forall i\neq j} f_{x_m,x_n} \tag{7}$$

where f_{x_m,x_n} denotes the passenger flow from station x_m to station x_n and K denotes the number of clusters. Essentially, F^{intra} and F^{inter}

represent the average intra-cluster and average inter-cluster flow, respectively (Figure 3*c*). To combine two measures, the ratio of F^{intra} to F^{inter} is adopted and defined as follows:

$$\delta = \frac{F^{\text{intra}}}{F^{\text{inter}}} \tag{8}$$

where δ denotes the flow-based metric. To obtain most self-contained clusters, δ should be maximized so that the average intra-cluster flow is as significant as possible.

Determination of the Number of Clusters

To determine the optimal number of clusters with both distance-based and flow-based metrics, different objective functions can be formulated. Because, in the current case, it was desired to (*a*) obtain clusters that are as spatially compact as possible, which can be achieved by minimizing τ, and (*b*) attain an intra-cluster flow as strong as possible, which can be achieved by maximizing δ, a straightforward way that takes the ratio of δ to τ was adopted. A scaling procedure was applied to both metrics before taking the ratio so that their magnitudes were comparable.

$$X' = \frac{X}{X_{\max}} \tag{9}$$

After applying the scaling procedure, the optimal number of clusters K^* was attained:

$$\underset{K\in[K_{\min},K_{\max}]}{\arg\max}\ \frac{\delta'_K}{\tau'_K} \tag{10}$$

where δ'_K and τ'_K denote the scaled flow-based and distance-based metrics for the K clustering, respectively.

RESULTS AND ANALYSIS

Results

Figure 4 shows the clustering results determined for each K, which ranged from 2 to 30 in this study on the basis of the calculation of SSE. The different clusters are illustrated with various combinations of colors and markers without the underlying transit network included. The variation in SSE is presented in Figure 3*a*. It can be seen that as the number of clusters increases, more clusters are generated, mainly within The Hague area. The SSE does not decrease linearly as K increases. Instead, a sharp drop can be observed in Figure 3*a* when K is approaching 8; then the decline becomes increasingly flat as K grows.

Figure 3*d* reveals that both the intra-cluster and inter-cluster distance measures show a pattern of decrease as K grows, although the intra-cluster one is smoother than its counterpart. Two scaled metrics are plotted together in Figure 3*e* for the sake of comparison. No specific patterns are very clear for both metrics, but when K is equal to 5, 12, and 24, the distance-based metric reaches some local minimums. The flow-based metric exhibits an overall growing pattern.

The integrated metric that takes the ratio of scaled flow-based to scaled distance-based metric is displayed in Figure 3*f*. The optimal number of clusters in terms of the integrated index in this case turns out to be 12 (highlighted in Figure 3*f*), although there is only a slight difference between $K10$ and $K12$, and $K22$, $K23$, $K24$, and $K25$ are also close. Detailed results and analysis of this optimal clustering are presented in Figure 5, including the spatial outcome and the number of stations contained in each cluster. The bar chart shows that the number of stations contained in the more isolated parts of the network (Clusters 1, 2, and 7) is significantly lower than in other parts of the network. This is arguably attributed to the low density of stations in these areas. Within the core area of The Hague, stations are more evenly distributed in different clusters, although there are still more stations assigned in Cluster 5 and Cluster 8.

Aggregated passenger demand at the cluster level is shown in Figure 5*d* and Figure 5*e,* with the former specifying all the numbers while the latter offers a visualization through a chord chart. Apparently, Cluster 5 accounts for the most demand because it contains all stations around the central station of The Hague with connections to train services. It is followed by Cluster 11 and Cluster 12. Cluster 11 covers the area of Den Haag HS station, which is the second-biggest train station in The Hague; Cluster 12 covers the main commercial area. Clusters 1, 3, 4, 7, and 10, however, show relatively low demand for HTM services, which can be partially explained by the presence of competing transit operators (i.e., bus and train). Furthermore, the low demand of Cluster 4 and 10 can be attributed to the relatively lower-density residential areas and lower overall transit market share. Cluster 8 exhibits a higher demand, presumably caused by the presence of regional and national institutions (e.g., museums, theater, stadium, and embassies) that attract visitors.

Spatial Variability Analysis

The spatial variability of all individual clusters is illustrated in Figure 5*c* through box plots of the spatial distance between stations in a cluster. This is important in the sense that travelers are assumed to be able to reach alternative stations within a cluster easily using nonmotorized modes, primarily walking and cycling. If the spatial variability of a cluster is too large, then some stations within the cluster are too far from each other and would not likely be considered as alternatives by travelers. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that are not considered outliers, and the outliers are plotted individually using the + symbol. All clusters' median station-to-station distances are less than 2 km, but Clusters 1, 2, and 7 show larger variability because of their larger spatial extents. Besides these three spatially isolated clusters, Cluster 3 and Cluster 9 also show more variability than the average. This is because these two clusters were generated with more scattered stations; they do not show the most desired round shapes as a result of the *k*-means algorithm. For example, some stations in the west of Cluster 9, as well as some in the southwest of Cluster 3, are more distant from the majority of the stations. This is admittedly one of the drawbacks of adopting the *k*-means algorithm.

Temporal Variability Analysis

As can be seen in Figure 6*a,* the transit demand in Haaglanden revealed clear within-day and across-day patterns during the regular operating time. To examine the influence of time-dependent passenger flow on

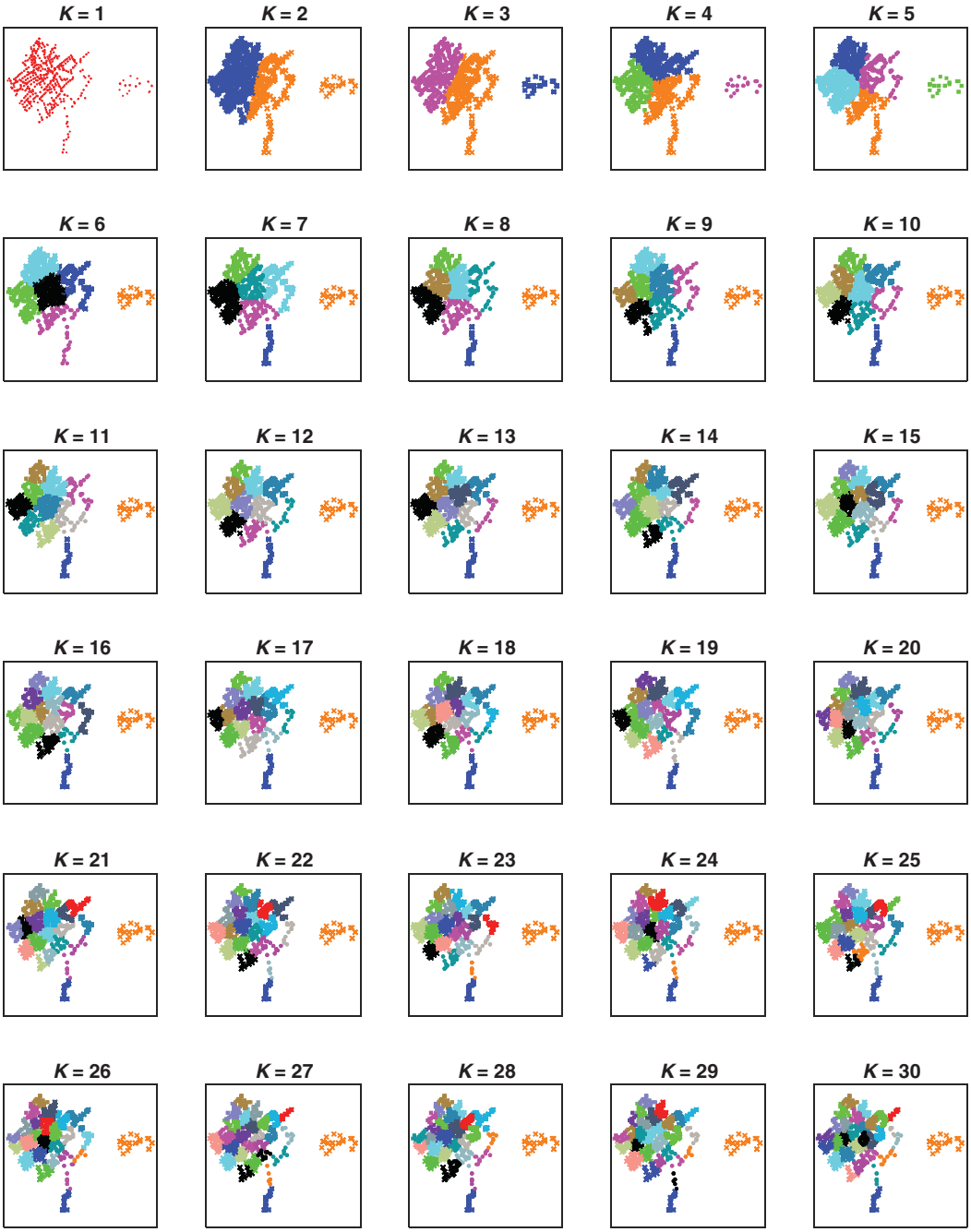
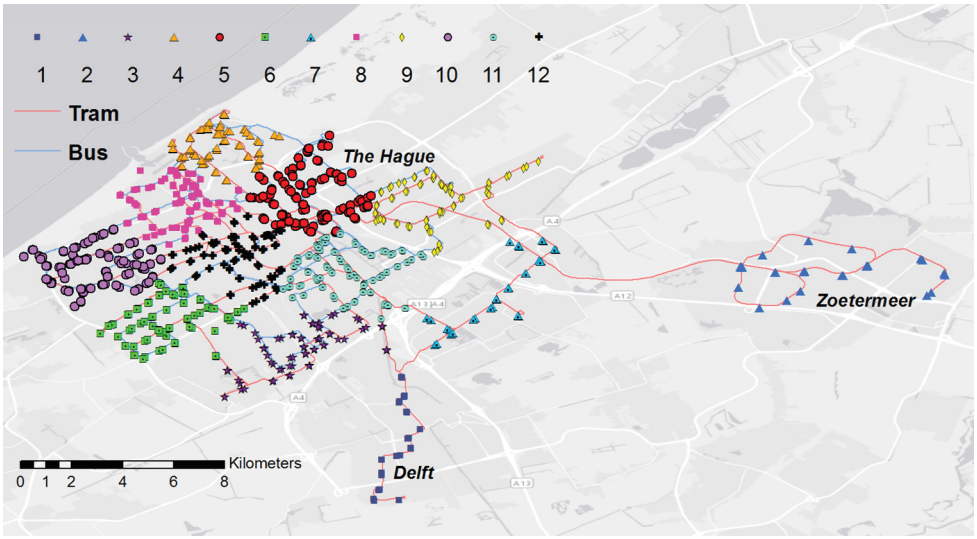
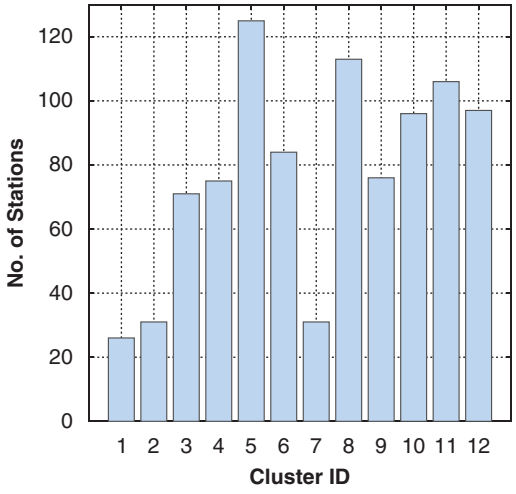


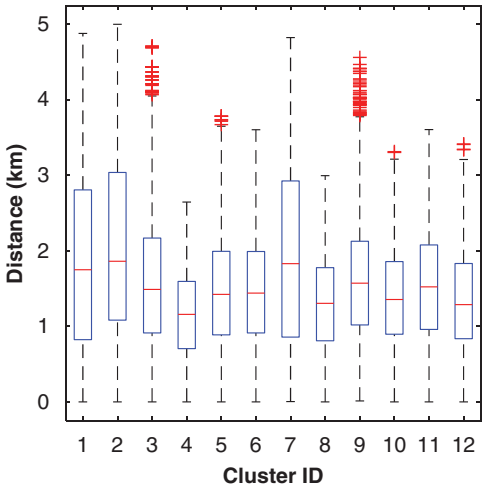
FIGURE 4 Illustration of clustering with different *K*.



(a)



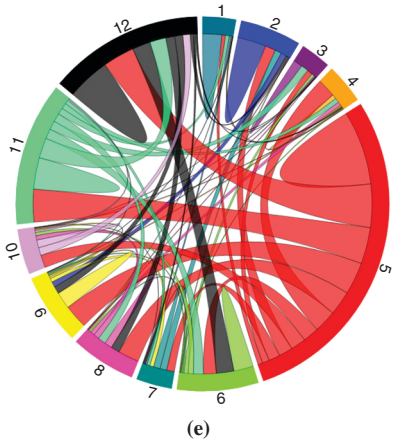
(b)



(c)

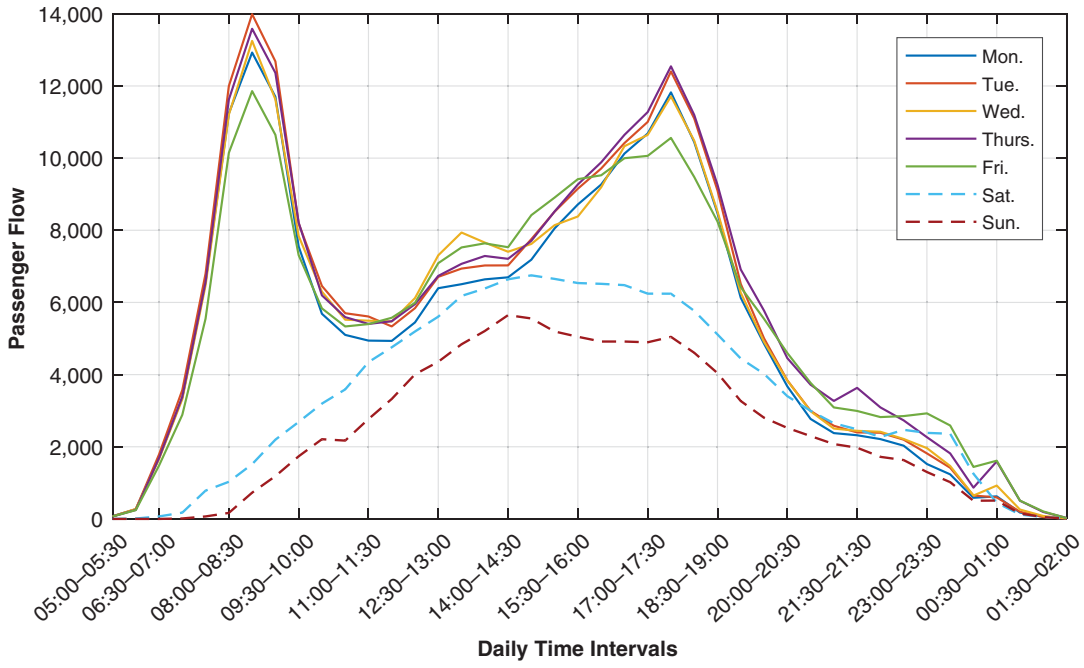
OD	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12
Cluster 1	121177	0	1546	5850	23001	0	13751	228	2711	0	22223	0
Cluster 2	0	178816	0	0	66471	3935	40878	3897	35243	2419	0	15844
Cluster 3	1069	0	58100	1232	35804	16170	0	4360	4071	4873	45274	1661
Cluster 4	5966	0	1187	28283	117664	4957	0	23469	7608	4253	28012	5818
Cluster 5	30268	87636	44599	141126	466499	100297	70813	153711	202626	96113	228770	272318
Cluster 6	0	4668	25186	6124	80867	122933	3605	7819	7006	32956	62252	110581
Cluster 7	10912	42561	0	0	48622	2801	42323	2783	23480	1794	18607	9671
Cluster 8	212	4435	4631	26360	132465	7182	3003	49039	7032	36745	47954	53815
Cluster 9	2279	34005	3808	7319	150812	6457	23575	8935	105280	9430	6548	46555
Cluster 10	0	2551	5646	4222	79136	19509	1850	28635	7529	48343	10960	48864
Cluster 11	40727	0	55729	36417	206610	62525	25834	52833	7128	7718	196911	126246
Cluster 12	0	20964	1856	9366	198927	115379	13404	57849	51066	45848	99807	272591

(d)

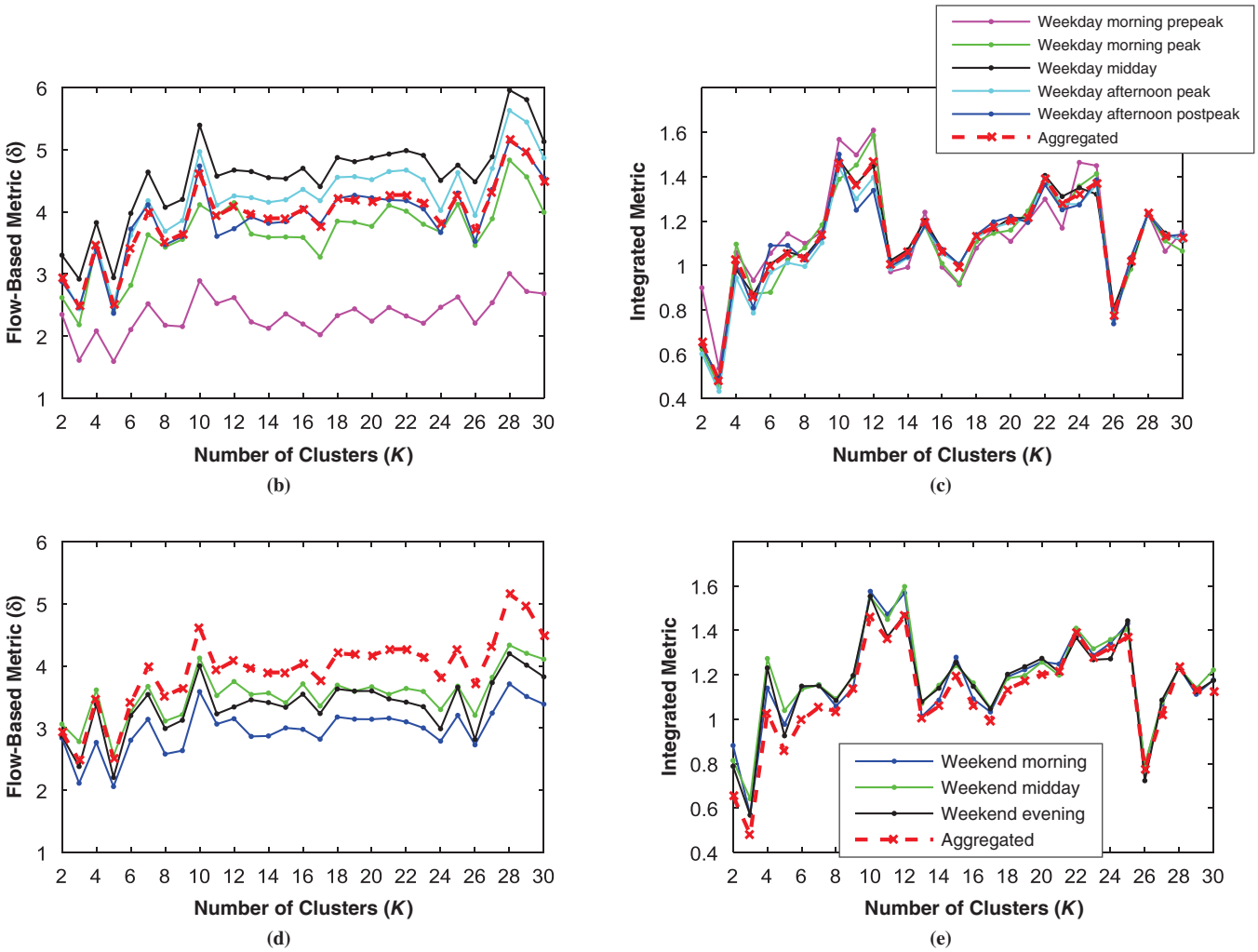


(e)

FIGURE 5 Illustrations of the optimal clustering, $K = 12$: (*a*) visualization of 12 clusters, (*b*) number of stations contained in each cluster, (*c*) illustrations of clusters' spatial variability, (*d*) constructed transit O-D matrices over the 31-day study period, and (*e*) visualization of the O-D passenger flow.



(a)



(b) (c) (d) (e)

FIGURE 6 Transit demand: (*a*) within-day and across-day temporal transit demand, (*b*) time-dependent flow-based metrics for different periods over weekdays, (*c*) integrated metrics for different periods over weekdays, (*d*) time-dependent flow-based metrics for different periods over weekend, and (*e*) integrated metrics for different periods over weekend.

the determination of the number of clusters, the proposed method was also implemented with multiple temporal passenger flow profiles from different periods. The following periods were investigated:

Weekdays
- Morning prepeak, before 7:30 a.m.;
- Morning peak, 7:30 to 10:00 a.m.;
- Midday, 10:00 a.m. to 3:00 p.m.;
- Afternoon peak, 3:00 to 7:30 p.m.;
- Afternoon postpeak, after 7:30 p.m.

Weekend
- Morning, before 10:00 a.m.;
- Midday, 10:00 to 6:00 p.m.;
- Evening, after 6:00 p.m.

Results for weekdays are shown in Figure 6*b* and Figure 6*c,* while results for weekends are in Figure 6*d* and Figure 6*e.* The red dashed line plotted in all these figures represents the result of aggregated passenger flow over the entire study period and can be used as a benchmark. Overall, the temporal flow variance was found not to have a significant influence on the final determination of number of clusters. The best choices still remain in the neighborhood of $K11$, although $K12$ in some cases turns out to be optimal while $K10$ is optimal in others. The general pattern remains stable.

One particular finding is that, during weekdays, the flow-based metric of midday is remarkably higher than the rest, while the flow-based metric of the morning prepeak hour always remains the lowest. This implies that more long-distance inter-cluster journeys are generated when people are going to their workplaces early in the morning. At midday, on the contrary, the intra-cluster flow is stronger than the inter-cluster one because these traveling activities are less related to commuting. However, the metrics of afternoon peak and afternoon postpeak suggest a more mixed composition of journey purposes, such as shopping, recreational, or household-related activities in the city after work. During the weekend, however, the flow-based metrics of all three periods stay at a low level, which implies that more inter-cluster journeys are performed than intra-cluster ones compared with weekdays. This can be explained by the fact that people normally go to the city for shopping or other leisure activities during the weekend.

CONCLUDING REMARKS

Accurate estimation of transit demand is crucial for both transit planning and operating processes. This paper proposes a *k*-means-based station aggregation method that can quantitatively determine the clustering by considering both flow and spatial distance information. Differing from the traditional way of grouping stops on the basis of traffic analysis zones, the proposed data-driven method offers another effective and efficient solution to those applications involving transit demand aggregation on the basis of directly observed flows rather than their proxies. The method was specified and applied to a case study of Haaglanden, Netherlands, by using the criterion that the ratio of average intra-cluster flow to average inter-cluster flow should be maximized while maintaining the spatial compactness of all clusters. This type of aggregation is particularly suited to urban areas characterized by a high density of transit stations, such as the case study area, Haaglanden. Travelers in such environs can choose different origin and destination stations and services.

The proposed method consists of four steps. First, the best clustering of each K is constructed by running the *k*-means algorithms a number of times with different initial centers and selecting the one that results with the minimum SSE, a measure for the variance of clusters. Then two metrics based on distance and passenger flow are computed, considering both intra-cluster and inter-cluster components. Finally, the two metrics are combined to determine the optimal number of cluster following the criterion adopted. This analysis process can be applied using other spatial and flow metrics of interest, depending on the application, case study characteristics, and data availability. The temporal variability analysis shows that the variance in passenger flow over time does not have a significant influence on the final determination of number of clusters when using the proposed method, which implies that this method is robust and can be potentially adopted for both short-term and long-term transit-related research.

One direction for future research is performing short-term transit demand prediction on the cluster level, which can be especially practical in the context of real-time transit operation but has not yet been well studied. Both temporal and spatial aggregation of transit demand are necessary to develop short-term predictions. In addition, transit O-D matrices are essential inputs to transit assignment models. Future research may examine the impact of different demand clustering on transit assignment performance (e.g., intra-cluster travel demand).

ACKNOWLEDGMENTS

This study was performed as part of the SETA project funded by the European Union's Horizon 2020 Research and Innovation program. The authors thank Menno Yap from the Department of Transport and Planning at Delft University of Technology for sharing his knowledge of the case study network and providing additional data for validation and visualization.

REFERENCES

1. Ceder, A. *Public Transit Planning and Operation: Modeling, Practice and Behavior.* CRC Press, 2015.
2. Pelletier, M.-P., M. Trépanier, and C. Morency. Smart Card Data Use in Public Transit: A Literature Review. *Transportation Research Part C: Emerging Technologies,* Vol. 19, No. 4, 2011, pp. 557–568. https://doi.org/10.1016/j.trc.2010.12.003.
3. Chu, K. K. A., and R. Chapleau. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2063, 2008, pp. 63–72. https://dx.doi.org/10.3141/2063-08.
4. Barry, J., R. Newhouser, A. Rahbee, and S. Sayeda. Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board,* No. 1817, 2002, pp. 183–187.
5. Gordon, J., H. Koutsopoulos, N. Wilson, and J. Attanucci. Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2343, 2013, pp. 17–24. https://dx.doi.org/10.3141/2343-03.
6. Ma, X., Y.-J. Wu, Y. Wang, F. Chen, and J. Liu. Mining Smart Card Data for Transit Riders' Travel Patterns. *Transportation Research Part C: Emerging Technologies,* Vol. 36, 2013, pp. 1–12. https://doi.org/10.1016/j.trc.2013.07.010.
7. Munizaga, M. A., and C. Palma. Estimation of a Disaggregate Multimodal Public Transport Origin–Destination Matrix from Passive Smartcard Data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies,* Vol. 24, 2012, pp. 9–18. https://doi.org/10.1016/j.trc.2012.01.007.

8. Nassir, N., A. Khani, S. Lee, H. Noh, and M. Hickman. Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2263, 2011, pp. 140–150. https://dx.doi.org/10.3141/2263-16.
9. Trépanier, M., N. Tranchant, and R. Chapleau. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations,* Vol. 11, No. 1, 2007, pp. 1–14. https://doi.org/10.1080/15472450601122256.
10. Wang, W., J. P. Attanucci, and N. H. Wilson. Bus Passenger Origin–Destination Estimation and Related Analyses Using Automated Data Collection Systems. *Journal of Public Transportation,* Vol. 14, No. 4, 2011, pp. 131–150. https://doi.org/10.5038/2375-0901.14.4.7.
11. Alsger, A., B. Assemi, M. Mesbah, and L. Ferreira. Validating and Improving Public Transport Origin–Destination Estimation Algorithm Using Smart Card Fare Data. *Transportation Research Part C: Emerging Technologies,* Vol. 68, 2016, pp. 490–506. https://doi.org/10.1016/j.trc.2016.05.004.
12. Alsger, A. A., M. Mesbah, L. Ferreira, and H. Safi. Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2535, 2015, pp. 88–96. https://dx.doi.org/10.3141/2535-10.
13. Barry, J., R. Freimer, and H. Slavin. Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2112, 2009, pp. 53–61. https://dx.doi.org/10.3141/2112-07.
14. Devillaine, F., M. Munizaga, and M. Trépanier. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2276, 2012, pp. 48–55. https://dx.doi.org/10.3141/2276-06.
15. Farzin, J. Constructing an Automated Bus Origin–Destination Matrix Using Farecard and Global Positioning System Data in Sao Paulo, Brazil. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2072, 2008, pp. 30–37. https://dx.doi.org/10.3141/2072-04.
16. Munizaga, M., F. Devillaine, C. Navarrete, and D. Silva. Validating Travel Behavior Estimated from Smartcard Data. *Transportation Research Part C: Emerging Technologies,* Vol. 44, 2014, pp. 70–79. https://doi.org/10.1016/j.trc.2014.03.008.
17. Lee, S., M. Hickman, and D. Tong. Development of a Temporal and Spatial Linkage Between Transit Demand and Land-Use Patterns. *Journal of Transport and Land Use,* Vol. 6, No. 2, 2013, pp. 33–46. https://doi.org/10.5198/jtlu.v6i2.268.
18. McCord, M., R. Mishalani, and X. Hu. Grouping of Bus Stops for Aggregation of Route-Level Passenger Origin-Destination Flow Matrices. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2277, 2012, pp. 38–48. https://dx.doi.org/10.3141/2277-05.
19. Hassan, M. N., T. H. Rashidi, S. T. Waller, N. Nassir, and M. Hickman. Modeling Transit Users Stop Choice Behavior: Do Travelers Strategize? *Journal of Public Transportation,* Vol. 19, No. 3, 2016, pp. 98–116. https://doi.org/10.5038/2375-0901.19.3.6.
20. Nassir, N., M. Hickman, A. Malekzadeh, and E. Irannezhad. Modeling Transit Passenger Choices of Access Stop. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2493, 2015, pp. 70–77. https://dx.doi.org/10.3141/2493-08.
21. Chu, K., and R. Chapleau. Augmenting Transit Trip Characterization and Travel Behavior Comprehension: Multiday Location-Stamped Smart Card Transactions. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2183, 2010, pp. 29–40. https://dx.doi.org/10.3141/2183-04.
22. Lee, S., M. Hickman, and D. Tong. Stop Aggregation Model: Development and Application. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2276, 2012, pp. 38–47. https://dx.doi.org/10.3141/2276-05.
23. Lee, S., and M. Hickman. Are Transit Trips Symmetrical in Time and Space? Evidence from the Twin Cities. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2382, 2013, pp. 173–180. https://dx.doi.org/10.3141/2382-19.
24. Tamblay, S., P. Galilea, P. Iglesias, S. Raveau, and J. C. Muñoz. A Zonal Inference Model Based on Observed Smart-Card Transactions for Santiago de Chile. *Transportation Research Part A: Policy and Practice,* Vol. 84, 2016, pp. 44–54. https://doi.org/10.1016/j.tra.2015.10.007.
25. Kieu, L.-M., A. Bhaskar, and E. Chung. A Modified Density-Based Scanning Algorithm with Noise for Spatial Travel Pattern Analysis from Smart Card AFC Data. *Transportation Research Part C: Emerging Technologies,* Vol. 58, 2015, pp. 193–207. https://doi.org/10.1016/j.trc.2015.03.033.
26. Kieu, L. M., A. Bhaskar, and E. Chung. Passenger Segmentation Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 16, No. 3, 2015, pp. 1537–1548. https://doi.org/10.1109/TITS.2014.2368998.
27. van Oort, N., T. Brands, and E. de Romph. Short-Term Prediction of Ridership on Public Transport with Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2535, 2015, pp. 105–111. https://dx.doi.org/10.3141/2535-12.
28. Bagherian, M., O. Cats, N. van Oort, and M. Hickman. Measuring Passenger Travel Time Reliability Using Smart Card Data. Presented at TRISTAN IX: Triennial Symposium on Transportation Analysis, Oranjestad, Aruba, June 12–17, 2016.
29. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* 1967, pp. 281–297.
30. Tou, J. T., and R. C. Gonzalez. *Pattern Recognition Principles.* Addison-Wesley, Massachusetts, 1974.
31. *MATLAB.* Version 8.6.0 (R2015b). MathWorks Inc., Natick, Massachusetts, 2015.
32. Ray, S., and R. H. Turi. Determination of Number of Clusters in *K*-Means Clustering and Application in Color Image Segmentation. *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques,* 1999, pp. 137–143.

The Public Transportation Group peer-reviewed this paper.