ESTIMATING ORIGIN-DESTINATION MATRICES USING AN EFFICIENT MOTH FLAME-BASED SPATIAL CLUSTERING APPROACH

A. A. Heidari *, A. Moayedi , R. Ali Abbaspour

School of Surveying and Geospatial Eng., College of Eng., University of Tehran, North Kargar Ave., Tehran, Iran
Email: (as_heidari, alimoayedi2013, abaspour)@ut.ac.ir

KEYWORDS: Automated Fare Collection, Origin-Destination Matrices, Clustering, Grey Wolf Optimization, Genetic Algorithm, Moth Flame Optimizer

ABSTRACT:
Automated fare collection (AFC) systems are regarded as valuable resources for public transport planners. In this paper, the AFC data are utilized to analysis and extract mobility patterns in a public transportation system. For this purpose, the smart card data are inserted into a proposed metaheuristic-based aggregation model and then converted to O-D matrix between stops, since the size of O-D matrices makes it difficult to reproduce the measured passenger flows precisely. The proposed strategy is applied to a case study from Haaglanden, Netherlands. In this research, moth-flame optimizer (MFO) is utilized and evaluated for the first time as a new metaheuristic algorithm (MA) in estimating transit origin-destination matrices. The MFO is a novel, efficient swarm-based MA inspired from the celestial navigation of moth insects in nature. To investigate the capabilities of the proposed MFO-based approach, it is compared to methods that utilize the K-means algorithm, gray wolf optimization algorithm (GWO) and genetic algorithm (GA). The sum of the intra-cluster distances and computational time of operations are considered as the evaluation criteria to assess the efficacy of the optimizers. The optimality of solutions of different algorithms is measured in detail. The traveler's behavior is analyzed to achieve to a smooth and optimized transport system. The results reveal that the proposed MFO-based aggregation strategy can outperform other evaluated approaches in terms of convergence tendency and optimality of the results. The results show that it can be utilized as an efficient approach to estimating the transit O-D matrices

1. INTRODUCTION

One of the imperative concerns through the evolution of urbanization in developing countries is mitigating the impact of traffic congestion (Ceder, 2007). The transit-oriented transport systems are attracting widespread interest because they can alleviate traffic congestion. The objective of these systems is to approximate and assess travelers demands according to models by evaluating spatiotemporal information relevant to the transportation requirements (Alfred Chu and Chapleau, 2008). In accordance with necessities of transportation systems, several studies focused on gathering transit demand data to realize an efficient approach (Alsger et al., 2015).

Transit origin–destination (O-D) estimation is one of the remarkable up shots of handing out smart card fare data (SCFD). The SCFD can be utilized for O-D estimation as a precious source of info, which will assist practitioners and researchers to dissect the capriciousness of users' travel patterns/activities. Hence, automated fare collection (AFC) systems emerged. The AFC systems can record travelers' boarding and/or alight info (Pelletier et al., 2011). These systems are extensively employed for demand estimation and transit planning in preference to old-fashioned transit surveys. The transit demand is typically organized in the form of origin-destination matrices, in which each element specifies the travel volume between the origin and destination in the networks (Alsger et al., 2015). Traditionally surveys such as home interviews, roadside interviews, flagging techniques, and aerial photography are some alternative approach to estimate the O-D

matrices (Cascetta, 1984). More recently, researchers have also studied the estimation of O-D matrices by means of SCFD (Farzin, 2008; Nassir et al., 2011).

In reality, users' activity may not originate from or be destined to a single stopping point, which it is dependent on their reasonable walking distance (Furth and Navick, 1992). Consequently, regarding travel demand, different choices in transit services can be made. According to the level of aggregation, aggregated stop groups can be symbolized as a single node (Lee et al., 2013). Stop groups that serve common or similar land use patterns and activities can be represented by a single node, which is advantageous to reduce the complexity of the transit networks. Nassir et al. evolved a transit stop choice model. They supposed that travelers can select their path by choosing a stop from desirable sets. Hassan et al. found that stop selection behavior is influenced by path and stop attributes. Furthermore, they declared that users' socioeconomic characteristic along with trip timing has an imperative role in choosing transit stops. Many studies demonstrated the significance of stop or station aggregation in analyzing AFC data. Interested readers can refer to (Lee et al., 2012; Lee and Hickman, 2014) for more info.

To pore over the spatial travel patterns and demand, of late, unsupervised learning approaches have been utilized by dint of their putative merits in tackling clustering tasks. The clustering is an unsupervised classification method, which groups the data into different subclasses (Jain et al., 1999). Instinctively, data within the same cluster are more similar to each other

---

* Corresponding author

than they are grouped into different clusters with regard to nature and heterogeneity (Hartigan, 1975). The well-known k-means clustering technique, which was developed by Mac Queen in 1967, has been addressed in several works (Jain, 2010). For instance, Ding Luo et al. proposed a four-step k-means-based technique to determine the optimal clusters. Their strategy applies spatial distance and traveler flow metrics to attain the optimal number of clusters. The effectiveness of the K-means highly depends on the initial choice of centroids. Hence, during the process of optimization, it may be stuck in local optima (LO) (Heidari et al., 2017a, 2017b, 2017c). The informed choice of initial centroids has an extremely influence on the efficacy of this algorithm and to mitigate the chance of stagnation to LO. However, many metaheuristic algorithms (MA) such as genetic algorithm (GA) and gray wolf optimizer (GWO) have been utilized to deal with clustering tasks (Mirjalili et al., 2014). The GA is very well-known and widely used MA, but still, it is computationally expensive. Furthermore, it might be sensitive to initial parameters and time taken for convergence. Therefore, several nature-inspired MA have been proposed for tackling optimization problems (Heidari et al., 2015a, 2015b; Heidari and Delavar, 2016). One of the recent MA is moth-flame optimization (MFO) (Mirjalili, 2015). The MFO was proposed in 2015 and it tries to inspire the idealized spiral motions of moths nearby the flames to develop a population-based MA. It has enriched exploration and exploitation operators with dynamic behaviors.

The main contribution of this paper is to investigate the performance of MFO on estimating transit O-D matrices. Several papers utilized traditional clustering techniques, but here a new MFO-based clustering approach is proposed and utilized to realize the O-D matrix construction task. According to previous works, it was observed that the performance of MFO isn't still assessed on estimating O-D matrix. To alleviate the deficiencies of traditional techniques, this research proposes a new MFO-based station aggregation algorithm (MFOSGA) that can determine the clustering aggregation based on the spatial detachment info. Then, the aggregation data are used to estimate and construct the O-D matrices. To confirm the efficacy of MFOSGA, its effectiveness is compared to the classical strategies including the GA and k-means techniques. Additionally, the efficacy of the GWO is investigated here.

2. PROBLEM DEFINITION

2.1 Estimating O-D matrices

This paper proposes a new transit passenger description by passenger segmentation using MFO-based aggregation. With the purpose of O-D matrices construction, the aggregated stations should be spatially compact. The passengers are segmented into recognizable groups, which it can assist in the mining of the daily changes in the travel pattern of any shape and size. Because of its random nature, the outcomes of the *MFO* algorithm can be different in each run .In order to obtain improved and more reliable results, the MFO algorithm was independently run 30 times. In the traffic planning, four sequential steps can be implemented: trip generation, trip distribution, modal splitting and trip Assignment. However, some steps might not be applicable depending on the situation.(Papacostas and Prevedouros, 2001)

2.2 Trip generation

In the first step, a number of trips that can be started or ended at each zone of the network should be determined. The size of these zones can be found by a number of clusters and it depends on the purpose of the model and reasonable precision. A centroid node is assigned to each of these zones. As a result, a single point represents a large area in the network.

2.3 Trip distribution

Trip distribution can establish the OD-flows $f = \{f_{od}\}$, $(o, d) \in$ $(O \times D)$ with connecting sums of trips between each centroid. According to the gravity model for trip distribution, the OD-flows from point (o) to point (d) can be expressed as:

$$g_{od} = k \cdot p_0 a_d f(\pi_{od}) \tag{1}$$

where k is a weighting factor, p_o counts the trip ends at the origin, a_d counts the trip ends at the destination, and $f(\pi_{od})$ is a weighting friction function.

2.4 Modal split

In the modal split, the flow for each OD-pair is partitioned into different travel modes such as walk, private car, public transit, etc. To determine how the OD-matrix can be disaggregated into different travel modes, the utility of each model must be calculated. This utility is employed to determine the chance of a certain traveler who can use a particular transportation mode. The following relation can be utilized to obtain the utility for travel mode $k \in K$.

$$u_k = \sum_{t \in T} \alpha_t x_{tk} + \varepsilon_k \tag{2}$$

where α_t is a weighting parameter, x_{tk} is a value of attribute t for travel mode k, and ε_k is a random error.

2.5 Trip assignment

The final step is to determine how the trips should be distributed on the transportation network.

2.6 Spatial clustering

Suppose that a data set $X = \{x_1, \ldots, x_N\}$, $x_n \in R_d$ and an integer k is given, which k represents the number of clusters. The goal of clustering is to determine the partition $S = \{s_1, s_2, \ldots, s_N\}$ of X into k $(k \leq n)$ non-empty subsets, in such a way that every element be included in one and only one of the subsets. The above-mentioned portioning must maximize the homogeneity of points within a group and, concurrently the heterogeneity of points between groups. In order to achieve this goal, generally the well-known and classical function, total mean square quantization error (MSE), can be used. The objective function can be defined as follows:

$$F(X, S) = \sum_{i=1}^{k} \sum_{x_i \in s_l} d(x_i, c_l)^2 \tag{3}$$

where $d(x_i, c_l)$ shows the distance between the object in each cluster and it's centroid. For this aim, several distance functions can be utilized but Euclidean distance can be considered as the most practical one, Euclidean distance between two objects x_i and x_j is defined as:

$$d(x_i, x_j) = \left(\sum_{p=1}^{d} (x_i^p - x_j^p)^2\right)^{0.5} \tag{4}$$

3. THE MOTH FLAME OPTIMIZER (MFO)

The MFO algorithm is a novel, efficient swarm-based MA inspired from the celestial navigation of moth insects in nature. The key inspiration behind the MFO is the transverse bearings

in moths. In MFO, The moths can play the featuring role of search agents that can explore and exploit far and wide over the solution space, whereas flames are considered the as best location of agents that seen so far. Flames can be described as flags that are installed by moths during the course of iterations. In MFO, the position of moths is updated based on Eq. (5):

$$M_i = S(M_i, F_j) \tag{5}$$

where M_i shows the i-th moth and F_j indicates the j-th flame.
In MFO, moths update their situation according to a spiral pattern, which is formulated as follows:

$$S(M_i, F_j) = D_i e^{bt} \cos(2\pi t) + F_j \tag{6}$$

where D_i is the distance between moths and flames, b is a constant, and t is a random value inside [-1, 1]. D_i can be obtained based on Eq. (7):

$$D_i = |F_j - M_i| \tag{7}$$

According to spiral model in the conventional MFO, a moth can be capable of circumnavigating a flag; consequently, the exploration and exploitation propensies of MFO can be befallen. To highlight exploitation inclination, t can be rationalized as a random value inside $[r, 1]$, where r should be linearly updated from -1 to -2 during the process. To improve exploitation, the number of flags should be decreased based on Eq. (8):

$$FN = round((N - l) \times (N - l)/T) \tag{8}$$

where l is the current iteration, N indicates the number of flames, and T specifies the upper bound of iterations.
The slow but sure decrement in the number of flames can assist MFO for making a steady equilibrium among exploration and exploitation capabilities. The computational complexity of MFO is as follows:

$$O(MFO) = O(t(n^2 + n \times d)) = O(tn^2 + tnd) \tag{9}$$

where n shows the number of search agents, t denotes the upper limit of iterations, and d indicates the number of variables.
The interested readers can refer to (Mirjalili, 2015) to see more details about the steps of MFO and its performance on other optimization tasks. The structure of the proposed MFO-based aggregation technique (MFOSGA) for estimating transit origin-destination matrices can be seen in Figure 1.

4. EXPERIMENTAL RESULTS

In this section, the efficacy of the proposed MFO-based estimation technique is investigated in depth. For these simulations, each algorithm is coded and implemented using MATLAB® software. For these tests, a T6400@4 GHz Intel Core (TM) PC with 2 GB of RAM was used to evaluate the algorithms. For each trial, the MFO-based strategy is substantiated throughout 30 independent runs. Note that the GWO and MFO algorithms have no initial parameters. The GWO has just an internal default parameter that is set to 2 in these experiments. It is worth noting that r parameter in MFO is linearly decreased from -1 to -2 during the aggregation process.
For GA algorithm, the best initial setting was found after several trial and error tests. Therefore, the crossover percentage was set to 0.8; mutation percentage was set to 0.3, mutation rate was set to 0.02, selection pressure was set to 8, and gamma was set to 0.05. The proposed strategy is applied to a case study from Haaglanden, Netherlands.

The performance of the MFO-based algorithm is compared to some well-established algorithms including K-means, GWO, and GA algorithms. The sum of the intra-cluster distances and time of operation is considered as the evaluation criteria.
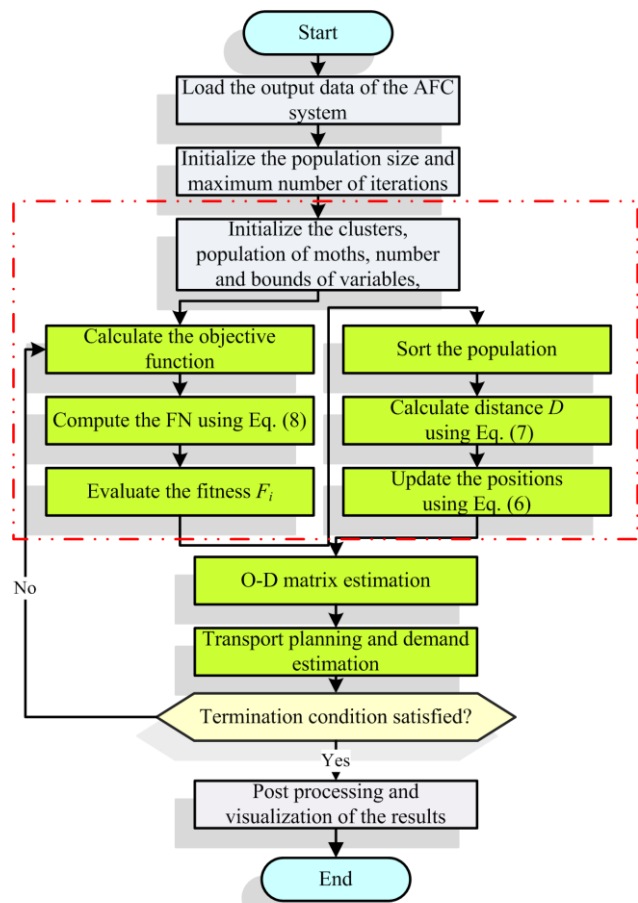

Figure 1. Flowchart of the proposed MFOSGA approach

In Table 1, the best, worst, average, median and the standard deviation (STD) of solutions for each algorithm is reported in detail.

Metric	K-means	GA	GWO	MFO
Best	2.5119	2.4812	6.9036	2.4964
Average	2.6717	2.5871	9.4578	2.6682
Worst	2.9106	2.5871	11.7018	2.9681
STD	0.1006	0.0678	1.6251	0.1373
Median	2.671	2.582	8.8574	2.6388

Table 1. Performance of different techniques

From the results in Table 1, it can be seen that the MFO can demonstrate a superior efficacy compared to the K-means and the GWO technique during 30 independent runs. The MFO is capable of providing the minimum value for the sum of intra-cluster distances compared to K-means and the GWO optimizer. Moreover, the STD values of MFO-based solutions are lower than those of K-means technique and the GWO in all runs;
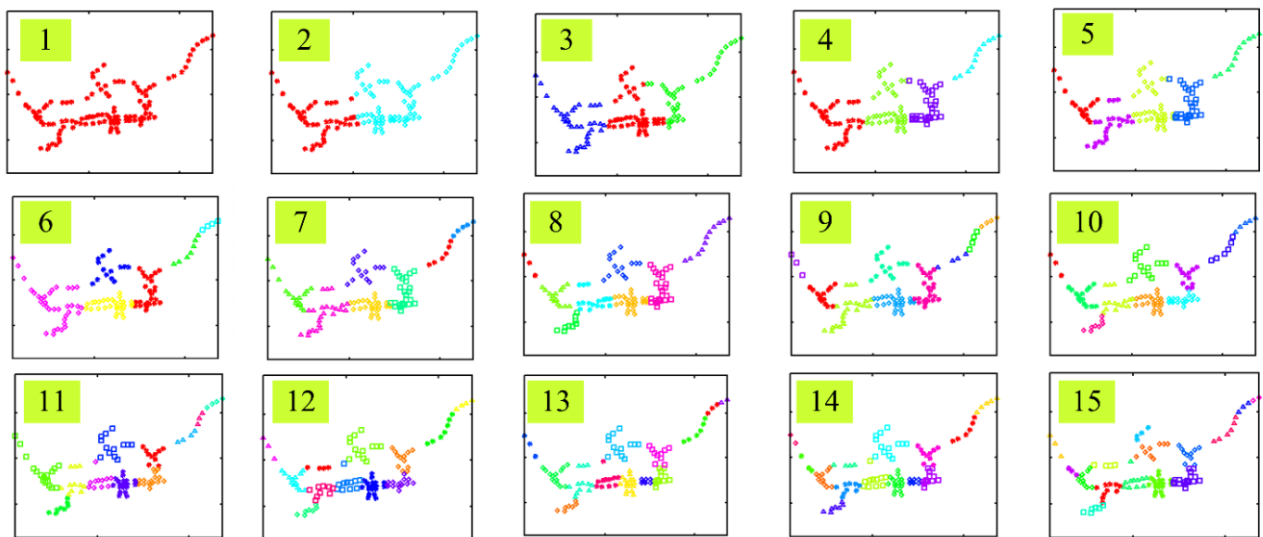
Figure 10. Clustering results for K inside interval [1, 15]

This fact shows that the MFO can reveal better results as a truthful and viable technique. For this reason, it can reach to the optimal solutions in all runs, while K-means may trap in LO. The main reason for this outperformance is that the MFO can perform a smooth transition from exploration to exploitation tendency. Meanwhile, the MFO can utilize a swarm of search agents; therefore, the LO escaping capacity of this method is high. Because each moth is assigned to a flame and the sequence of flames is updated during iterations, the MFO can demonstrate an extensive exploration tendency compared to the GWO and GA algorithms. This feature can increase the chance of the MFO to escape from stagnation to the LO. In addition, adaptive nature of FN value can make a fine tradeoff between the exploration and exploitation inclinations. The results are also graphically compared in Figure 2.
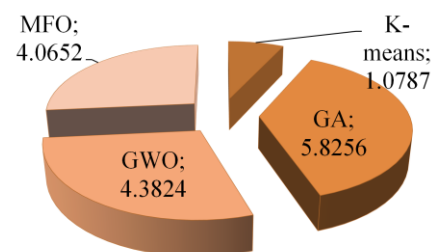

Figure 3. Comparison of computational time of MFO, GWO, GA and K-means methods

As it can be seen from Figure 3, the K-means can perform as the fastest algorithm. However, when only metaheuristics are considered, the MFO can be regarded as the fastest technique as well. The number of stations in each cluster is shown in Figure 4. It can be inferred that the cluster 10 contained more than 20 stations. On the other hand, the cluster 1 located in isolated parts of the network and includes just one station.
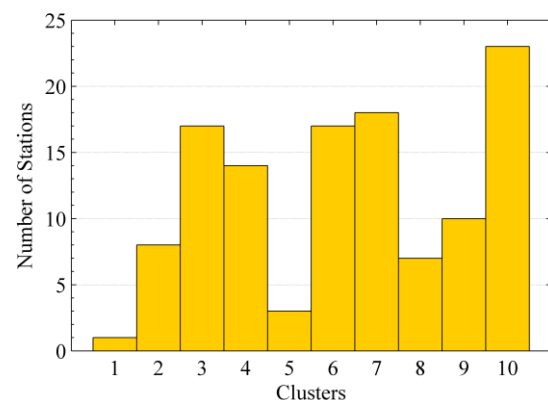

Figure 2. Performance of different techniques

Figure 4. Number of stations in each cluster

Furthermore, the performance of the proposed MFO-based technique is compared to other considered techniques in terms of computational time. The results can be seen in Figure 3.

Variation in the sum of intra-cluster distances is exposed in Figure 5. As it can be seen in Figure 5, when the number of clusters increases, the cost function also decreases. Initially, the graph shows a dramatic decline, but after approaching 10 clusters, the acceleration of reduction is moderately decreased.
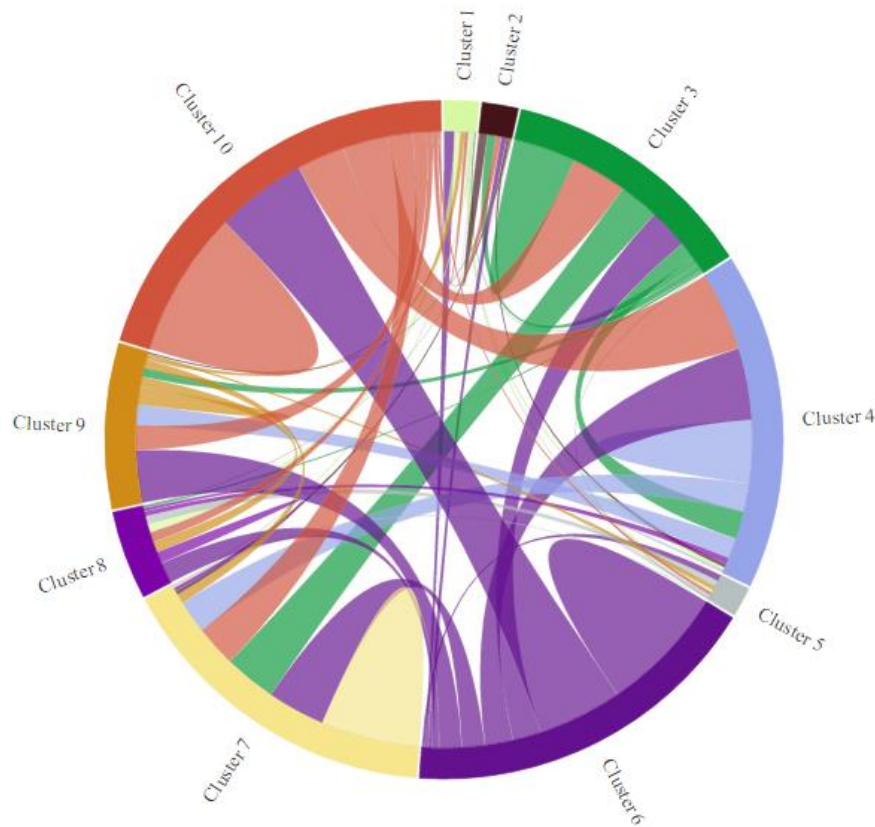
Figure 6. Visualization of the O-D matrix

This also shows that the MFO-based strategy can effectively exploit the vicinity of the fruitful areas of the search space.
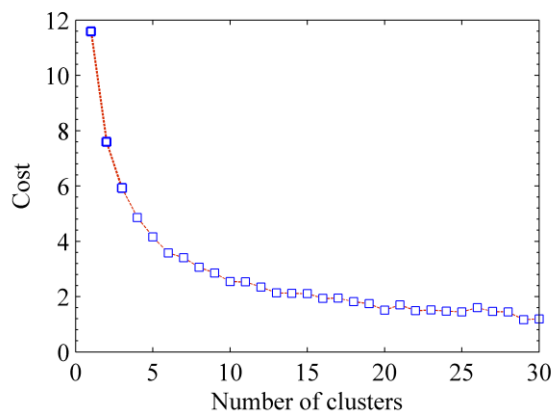

Figure 5. Decreasing behavior of cost function

Figure 7. Clusters' spatial variability

In Figure 6, the aggregated passenger demand at the cluster level is shown using a chord chart. According to this chord chart, most demands related to clusters 4 and 10 since they might contain crowded stations, because of the commercial complexes or other city centers organizations.

In Figure 7, the clusters' spatial variability is demonstrated. According to Figure 7, cluster 7 covers a wide area and stations are far from each others. On the other hand clusters 5 and 9 have the minimum clusters' spatial variability and cover a small area. Note that cluster 1 has covered no station.
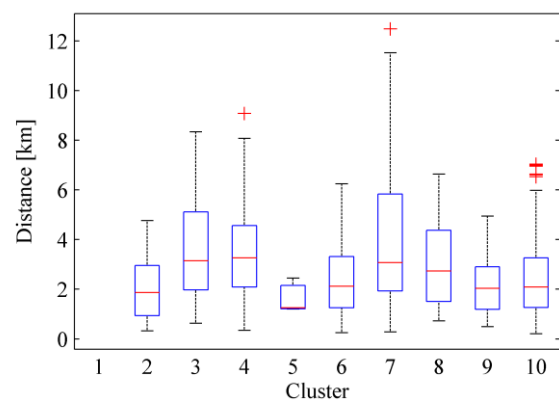
The convergence behaviors of MFO algorithm is compared with the GA and GWO techniques in Figure 8. It can be seen that the convergence of the MFO and GA algorithms has demonstrated a very competitive convergence trend. It was observed that the MFO can outperform GWO in terms of convergence acceleration. The MFO can lastly converge to desirable results compared to the GA and GWO. The reason is that the used adaptive constant in MFO can stimulate the optimizer to show accelerated convergence nearby the flames during aggregation procedure.
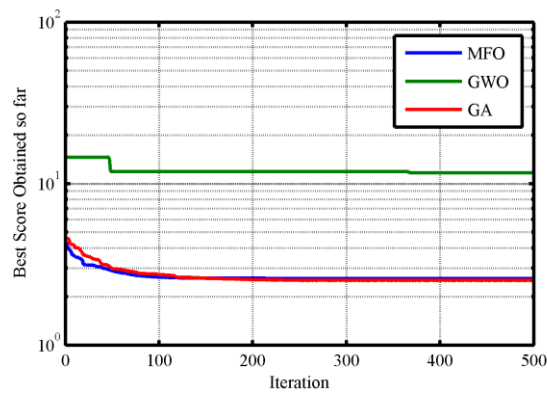
Figure 8. Convergence trends of different optimizers

The clustering results of MFO-based aggregation strategy are exposed in Figure 9. Note that these results are related to a specific case that we have 10 aggregated station groups (ASG).
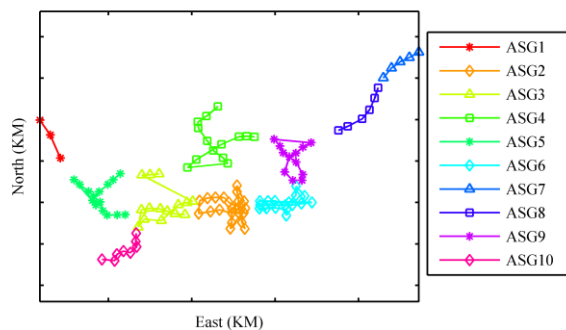

Figure 9. The clustering results for *K*=10.

In Figures 10-13, the results of MFO-based aggregation strategy for k from 1 to 15 are exposed.
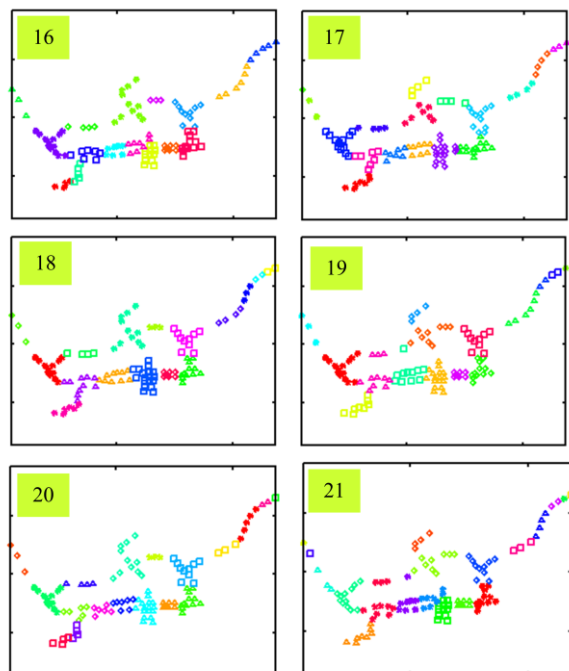

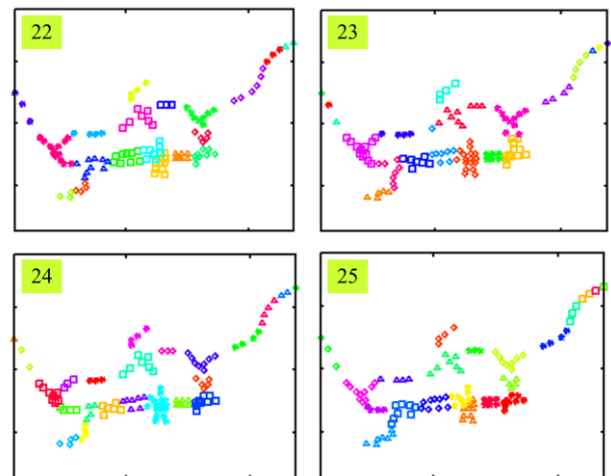Figure 11. Clustering results for *K* inside interval [16, 21]

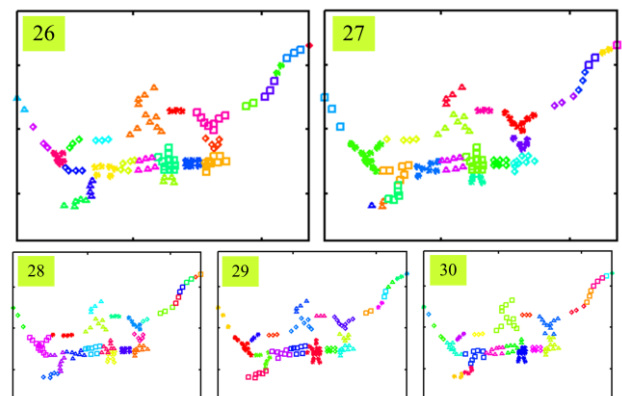Figure 12. Clustering results for *K* inside interval [22, 25]

Figure 13. Clustering results for *K* inside interval [26, 30]

According to the results, it can be concluded that the proposed MFO-based aggregation strategy can outperform other GA and GWO in terms of convergence tendency, time, and optimality of the results. As it can be seen in different figures, the results show that it can be utilized as an efficient approach to estimating the transit O-D matrices.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The AFC data can be used as a great source of info for studying the mobility patterns of individuals in public transit networks. In this research, the AFC data and the MFO optimizer were utilized to estimate the transit origin-destination matrices in transit systems. The MFO is a new population-based MA that tries to inspire the celestial navigation of moths. To scrutinize the efficacy of the proposed MFO-based approach, it was compared to K-means, GWO, and GA-based strategies. The optimality of solutions of different methods was measured. It can be concluded from the results that the proposed MFO-based aggregation strategy can provide better results than other evaluated approaches in terms of convergence tendency and optimality of the results. The results reveal that this method can be utilized as an efficient approach to estimating the transit O-D matrices. For future works, the performance of other new algorithms can be evaluated in dealing with this problem.

REFERENCES

Alfred Chu, K., Chapleau, R., 2008. Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board*, 2063, pp. 63-72.

Alsger, A. A., Mesbah, M., Ferreira, L., Safi, H., 2015. Use of smart card fare data to estimate public transport origin–destination matrix. *Journal of the Transportation Research Board*, 2535, pp. 88-96.

Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. *Transportation Research Part B: Methodological*, *18*(4), pp. 289-299.

Ceder, A., 2007. Public Transit Planning and Operation: Theory, Modeling and Practice. Elsevier, Butterworth-Heinemann.

Farzin, J., 2008. 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*, 2072, pp. 30-37.

Furth, P. G., & Navick, D. S., 1992. Bus route OD matrix generation: Relationship between biproportional and recursive methods. *Transportation Research Record*, 1338, pp. 14-21.

Hartigan, J. A., & Hartigan, J. A., 1975. *Clustering algorithms* (Vol. 209). New York: Wiley.

Heidari, A. A., Abbaspour, R. A. and Jordehi, A. R., 2017a. An efficient chaotic water cycle algorithm for optimization tasks. Neural Computing and Applications, 28, pp. 57–85.

Heidari, A. A., Abbaspour, R. A., Jordehi, A. R., 2017b. Gaussian bare-bones water cycle algorithm for optimal reactive power dispatch in electrical power systems. *Applied Soft Computing*, 57, pp. 657–671.

Heidari, A. A., Pahlavani, P., 2017c. An Efficient Modified Grey Wolf Optimizer with Lévy Flight for Optimization Tasks. *Applied Soft Computing*, 60, pp. 115–134.

Heidari, A. A. and Delavar, M. R.: A Modified Genetic Algorithm for Finding Fuzzy Shortest Paths in Uncertain Networks, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLI-B2, 299-304, doi:10.5194/isprs-archives-XLI-B2-299-2016, 2016.

Heidari, A. A., Kazemizade, O., and Abbaspour, R. A.: An Effective Harmony Search Algorithm with Opposition based Chaos-Enhanced Initialization for Solving Uncapacitated Facility Location Problems, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XL-1/W5, 307-311, doi:10.5194/isprsarchives-XL-1-W5-307-2015, 2015.

Heidari, A. A., Mirvahabi, S. S., and Homayouni, S.: An effective hybrid support vector regression with chaos-embedded biogeography-based optimization strategy for prediction of earthquake-triggered slope deformations. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XL-1/W5, 301-305, doi:10.5194/isprsarchives-XL-1-W5-301-2015, 2015.

Jain, A. K., 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), pp. 651-666.

Jain, A. K., Murty, M. N., Flynn, P. J., 1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), pp. 264-323.

Lee, S., Hickman, M., Tong, D., 2012. Stop Aggregation Model: Development and Application. *Transportation Research Record: Journal of the Transportation Research Board*, 2276, pp. 38-47.

Lee, S. G., Hickman, M., 2014. Trip purpose inference using automated fare collection data. *Public Transport*, 6(1-2), pp. 1-20.

Mirjalili, S., 2015. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowledge-Based Systems*, 89, pp. 228-249.

Nassir, N., Khani, A., Lee, S., Noh, H., Hickman, M., 2011. 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*, 2263, pp. 140-150.

Papacostas, C.S., Prevedouros, P.D., 2001. Transportation Engineering and Planning. Prentice Hall.

Pelletier, M. P., Trépanier, M., & Morency, C., 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4), pp. 557-568.