A multi-objective approach for station clustering in bike sharing systems

Selin Ataç
Nikola Obrenović
Michel Bierlaire

Transport and Mobility Laboratory (EPFL)

September 2021

STRC
21th Swiss Transport Research Conference
Monte Verità / Ascona, September 12 – 14, 2021

Transport and Mobility Laboratory (EPFL)

A multi-objective approach for station clustering in bike sharing systems

Selin Ataç, Michel Bierlaire
Transport and Mobility Laboratory
École Polytechnique Fédérale de Lausanne
TRANSP-OR, Station 18, CH-1015 Lausanne
phone: +41-21-693 81 00
fax: +41-21-693 80 60
{selin.atac,michel.bierlaire}@epfl.ch

Nikola Obrenović
BioSense Institute
University of Novi Sad
Dr Zorana Đinđića 1, 21000 Novi Sad, Serbia
phone: +41-AA-BBB CC DD
fax: +41-AA-BBB CC DE
nikola.obrenovic@biosense.rs

September 2021

Abstract

Increasing environmental concerns direct people to more sustainable solutions in all fields. In transportation, one of those solutions is vehicle sharing systems. Although these systems are convenient for the users, it creates many operational challenges, such as imbalance of the vehicles throughout the service area. Usually, staff-based rebalancing operations are conducted to maintain the balance, thus to provide higher level of service. These operations become difficult to solve with the increasing number of stations. Therefore, some heuristic approaches such as clustering are used to split the problem into smaller sub problems. This paper focuses on bike sharing systems with static rebalancing operations. Two multi-objective mathematical models are specifically crafted for the rebalancing-oriented clustering problem. These models and two agglomerative hierarchical clustering approaches are compared with respect to resulting cost of rebalancing operations.

Keywords

Bike sharing systems; Rebalancing operations; Optimization; Clustering

1 Introduction

The increasing environmental concerns triggered the sectors to seek more environmentally friendly options. According to Pachauri *et al.* (2014), the transportation sector is responsible for the 14% of the global greenhouse gas emissions. Therefore, we see that the experts have directed towards more sustainable solutions such as carbon neutral fuel and electric cars. Ride-sharing and vehicle sharing systems (VSSs) are also other solutions due to reduced car ownership.

A VSS offers short-term vehicle rental to its users. The users can unlock the vehicles using an RFID card or a mobile application. The trip length and duration are the main factors that affect the price of the trip. Additional to the environmental benefits that VSSs bring, it also offers financial advantages and convenience since the maintenance and insurance costs are covered by the VSS operator.

The first example of VSSs, that was initiated in Zurich, date back to 1940s. A car sharing system (CSS), Selbstfahrergemeinschaft, was used starting from 1948 (Shaheen *et al.,* 1998). In 1965, in the Netherlands, an organization named Provo introduced the first bike sharing system (BSS). However, since this initiative was not profit oriented and the bikes were not locked, the system was abused and the bikes were stolen (Shaheen *et al.,* 2010). Thanks to the improvements in technology and also the opportunity of making business, the number of bike sharing systems considerably increased from 13 in 2004, to 855 in 2014 (Fishman, 2016). An analysis on the five different stages of the evolution of bike sharing systems can be found in Midgley (2011).

On the other hand, a VSS operator faces some challenges to provide satisfactory service to its users. These include imbalance of vehicles, management of vehicles and parking spots, demand forecasting and pricing strategies. The better these are addressed, the better pricing and service the users get from the system. This also increases the revenue and level of service.

Although the vehicle types used in VSSs may differ, the VSS configurations can be analyzed in the same way. The trip type can be either return trip or one-way. The former requires that the vehicle is dropped to the same station that it was picked up. The latter is more flexible and allows a user to park the vehicle anywhere designed in the city, regardless of the pick up station. This brings up the problem of imbalance of vehicles throughout the service area. To overcome it and provide a higher level of service, the operators use rebalancing operations to relocate vehicles from overcrowded stations to those with a lack of vehicles. These operations can be user-based, static staff-based, or dynamic staff-based. The static staff-based rebalancing takes place when the system is closed or low in service, generally during the night, every day. The dynamic

rebalancing is continuously conducted during the system operation. The pricing can be done in two different ways: fixed and dynamic. In fixed pricing, the trip duration and/or length are the factors determining the price. In dynamic pricing, other factors such as origin, destination, and time of the day, affect the price. Dynamic pricing is generally used for user-based rebalancing operations by encouraging them to do unpopular trips. Finally, the parking organization in VSSs can be analyzed under two: station-based and free-floating. In station-based systems the users are required to park at some designated parking areas determined by the operator. The free-floating systems are more flexible since the pick-up and drop-off locations can be any parking spot in the service area.

In our previous work (Ataç *et al.,* 2020), we deal with a one-way station-based BSS. In this system static staff-based rebalancing operations are performed and fixed pricing is assumed. We compare two extreme scenarios, one with known demand and the other with unknown, to find the value of demand forecasting. The former assumes the Origin-Destination (OD) information is perfectly known, whilst the latter does not have any information on OD-trips. We improve an existing mathematical model for the rebalancing optimization from the literature and introduce a discrete event simulator to assess the value of demand forecasting.

As the introduced mathematical model in Ataç *et al.* (2020) is intractable for large instances, the analysis of the value of demand cannot be conducted in bigger systems. To overcome this issue, we decide to disaggregate the system into smaller sub systems to enable solution of the rebalancing optimization problem. Although this problem is studied in the literature using traditional clustering methods, we could not find any works that investigate clustering approaches with multiple objectives in this context. Therefore, in this paper, we develop two multi-objective mathematical models for clustering BSS stations, and we utilize two clustering methods from the literature, to compare and select the most proper clustering method.

The rest of the paper is organized as follows: Section 2 presents the literature review on rebalancing operations and clustering approaches in BSSs. Afterwards, we introduce the mathematical models for both rebalancing operations and the clustering in Section 3. The experimental results done on two BSSs, namely nextbike Sarajevo, Bosnia and Herzergovina and nextbike Berlin, Germany, are discussed in Section 4. Finally, in Section 5, we conclude the paper and suggest possible future research directions.

2 State of the art

In this section, we briefly discuss the literature on rebalancing operations and station clustering. We focus our review on the station-based BSSs. The reader may refer to Ataç *et al.* (2021) and Laporte *et al.* (2018) for more thorough literature surveys in the VSSs.

The rebalancing operations in VSSs are proposed as a solution to balance the number of vehicles throughout the service area according to the demand structure. The staff-based rebalancing operations can be conducted at night or when the demand is low, i.e., static rebalancing (Raviv *et al.,* 2013*,* Dell'Amico *et al.,* 2014*,* Schuijbroek *et al.,* 2017*,* Pal and Zhang*,* 2017*,* Erdogan *et al.,* 2014), or during the system operation, i.e., dynamic rebalancing, (Pfrommer *et al.,* 2014*,* Boyaci *et al.,* 2017). Some hybrid approaches also exist. For example, Nair and Miller-Hooks (2011) present a methodology for static rebalancing, but they apply it four times a day to obtain a closer application to dynamic rebalancing.

Raviv *et al.* (2013) consider a BSS with capacitated stations. The two mixed integer linear programming (MILP) formulations, that solve the rebalancing optimization, are based on one-commodity pick-up and delivery traveling salesman problem (1-PDTSP). Dell'Amico *et al.* (2014) work on the same problem and propose four MILPs. Both works minimize the operating costs and use valid inequalities to reduce the computation time. Raviv *et al.* (2013) also take the user satisfaction, loading and unloading times into account. Ho and Szeto (2017) utilize the arc-indexed formulation of Raviv *et al.* (2013) and revise it by adding station characteristic constraints. Erdogan *et al.* (2014) introduce flexibility in the model by introducing demand intervals. This way, they observe less computational complexity than 1-PDTSP. Shu *et al.* (2013) use a network flow model for the BSSs with proportionality constraints to obtain the number of trips supported, the initial allocation of bicycles at each station, the flow of bicycles, and the bicycle utilization rate at each time period. This framework allows them to test the value of rebalancing operations in the case study adapted from Singapore.

Heuristics are used for large size instances. These include tailor-made branch and cut algorithms (Dell'Amico *et al.,* 2014*,* Erdogan *et al.,* 2014*,* Chemla *et al.,* 2013b), Benders decomposition (Erdogan *et al.,* 2014), neighborhood search (Ho and Szeto*,* 2017*,* Cruz *et al.,* 2017), and clustering based approaches (Schuijbroek *et al.,* 2017*,* Liu *et al.,* 2016*,* Boyaci *et al.,* 2017*,* Feng *et al.,* 2017*,* Ma *et al.,* 2019*,* Lahoorpoor *et al.,* 2019). For instance, Cruz *et al.* (2017) propose a hybrid iterated local search based heuristic combined with randomized variable neighborhood descent. They show that their approach outperforms the previous methods by Chemla *et al.* (2013b) and Erdogan *et al.* (2014).

Schuijbroek *et al.* (2017) develop a mixed integer programming (MIP) and a constraint programming (CP) approach for the whole BSS to solve the rebalancing optimization. Then, they propose another MIP to cluster the stations and a heuristic that solves the rebalancing MIP for each cluster. The proposed MIP for clustering aims at minimizing the makespan of rebalancing operations. By using a Maximum Spanning Star approximation, they can solve the clustering problem in real time. Among the three methodologies, they see that the clustering based heuristic outperforms the other two. Liu *et al.* (2016) cluster the stations by selecting K stations as cluster centers and assign each station to the closest center. Then, the solution is modified by taking into account the rebalancing vehicle capacity. The routing is then solved for each cluster by the proposed mixed integer non linear programming (MINLP) model. Boyaci *et al.* (2017) introduces a clustering algorithm that is similar to k-medoid since they aim to minimize the dissimilarities within clusters.

In Feng *et al.* (2017), the authors examine both k-means and hierarchical clustering in order to analyze the BSS stations of Vélib' system in Paris. They conclude that there exists four main station clusters, which they name as employment, residential, starving, and overfed. Ma *et al.* (2019) and Lahoorpoor *et al.* (2019) cluster the BSS stations by their spatial and temporal characteristics. Among hierarchical clustering, expectation maximization clustering, and K-means clustering, Ma *et al.* (2019) observe that the latter shows the best performance for the case of Ningbo, China. They claim that the resulting seven clusters show different characteristics in terms of people's travel habits and land use around the stations. Lahoorpoor *et al.* (2019) build a similarity matrix based on the number of trips between each station. This information helps to identify the correlated stations by agglomerative hierarchical clustering using Ward linkage.

All in all, optimization models are a strong tool to solve the rebalancing problems. Regarding the large size instances, we see that decomposition methods, heuristics, and clustering based approaches are used.

3 Methodology

In previous work, we propose a methodological framework to assess the value of demand forecasting (Ataç *et al.,* 2020). The two main modules, namely simulation and optimization, help us to compare two extreme scenarios, the optimization of rebalancing operations with and without the knowledge of the future demand. However, with the increasing problem size, this rebalancing optimization becomes intractable in real time. Clustering of VSS stations is one of the methods used in the literature to reduce the computational complexity by splitting the main

Table 1: Notation for the model given by ($F1_M$)

Sets and indices	
N	number of stations (i, $j \in \{1,...,N\}$)
V	set of stations from 0 to N, where 0 is the depot
Parameters	
m	the number of relocation vehicles available
Q	the capacity of each relocation vehicle
c_{ij}	the cost of traveling from i to j, where $i,j \in$N
q_i	the demand at each station, where $i \in$V
$qCount$	number of stations involved in the optimization problem
Decision variables	
x_{ij}	1 if arc (i, j) is used by a relocation vehicle, 0 otherwise, where $i, j \in V$
θ_j	the load of a vehicle after it leaves node j, where $j \in V$
u_i	auxiliary decision variable for the MTZ constraints

problem into smaller sub problems. Subsequently, the rebalancing problem is solved separately in each station cluster. Therefore, this work addresses two agglomerative hierarchical methods and two multi objective clustering approaches, that cluster the BSS stations.

This section presents the two sub modules that we use in this paper, i.e., rebalancing optimization (Section 3.1) and clustering (Section 3.2). The rebalancing optimization module determines the routing for the rebalancing vehicles whilst the clustering module deals with station clustering. These two modules are used together to compare the results of different clustering approaches.

3.1 Rebalancing optimization

As discussed in Section 2, the optimization programs are widely used to find the routing for rebalancing operations in BSSs. As this paper considers a one-way station-based BSS that adopts static rebalancing, we decide to use and improve one of the MILP formulations from Dell'Amico *et al.* (2014). The original model ($F1$) can be found in Appendix A.

We give the modified model as ($F1_M$) and the related notation in Table 1. The c_{ij} corresponds to the length of the shortest path between station i and station j. The cost from the depot to and from any station is assumed to be zero. q_i can take any integer value. Since the model is

solved for the subset of stations that show non zero demand as in Liu *et al.* (2016), we introduce another parameter $qCount$, that gives the number of stations with non zero demand.

As $(F1)$ includes exponential number of subtour elimination constraints (SECs), we propose a modification to this formulation by replacing the classical SECs by Miller-Tucker-Zemlin (MTZ, Miller *et al.* (1960)) constraints, i.e., (6) and (7) in Ataç *et al.* (2020). Constraints (8) are also added to prevent the subtours to the same station. Also the valid inequalities ((13) and (14)) proposed by Dell'Amico *et al.* (2014) are added to the model. These inequalities cut the solutions where a rebalancing truck going through three nodes which have a total supply/demand larger than the capacity of the vehicle, where

$$S(i,j) = \{h \in V \setminus \{0\}, h \neq i, h \neq j : |q_i + q_j + q_h| > Q\}.$$

Given these, $(F1_M)$ finds the routing plan for the relocation vehicles.

$$(F1_M)\ \min \quad \sum_{i\in V}\sum_{j\in V} c_{ij}x_{ij} \tag{1}$$

$$\text{s.to} \quad \sum_{i\in V} x_{ij} = 1 \qquad \forall j \in V \setminus \{0\} \tag{2}$$

$$\sum_{i\in V} x_{ji} = 1 \qquad \forall j \in V \setminus \{0\} \tag{3}$$

$$\sum_{j\in V} x_{0j} \leq m \tag{4}$$

$$\sum_{j\in V\setminus\{0\}} x_{0j} - \sum_{j\in V\setminus\{0\}} x_{j0} = 0 \tag{5}$$

$$u_i - u_j + N * x_{ij} \leq N - 1 \qquad \forall i, j \in V \setminus \{0\} \tag{6}$$

$$1 \leq u_i \leq N - qCount \qquad \forall i \in V \tag{7}$$

$$x_{ii} = 0 \qquad \forall i \in V \tag{8}$$

$$\theta_j \geq \max\{0, q_j\} \qquad \forall j \in V \tag{9}$$

$$\theta_j \leq \min\{Q, Q + q_j\} \qquad \forall j \in V \tag{10}$$

$$\theta_j - \theta_i + M(1 - x_{ij}) \geq q_j \qquad \forall i \in V, j \in V \setminus \{0\} \tag{11}$$

$$\theta_i - \theta_j + M(1 - x_{ij}) \geq q_j \qquad \forall i \in V \setminus \{0\}, j \in V \tag{12}$$

$$x_{ij} + \sum_{h\in S(i,j)} x_{jh} \leq 1 \qquad \forall i, j \in V \setminus \{0\}, h \in S(i,j) \tag{13}$$

$$\sum_{h\in S(i,j)} x_{hi} + x_{ij} \leq 1 \qquad \forall i, j \in V \setminus \{0\}, h \in S(i,j) \tag{14}$$

$$\theta_0 = 0 \tag{15}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in V \tag{16}$$

3.2 Clustering

Clustering based approaches are used to split a problem into smaller sub problems to reduce the computational complexity. We consider and compare different clustering approaches, i.e., agglomerative hierarchical clustering (AHC) with Ward linkage and proximity of stations as a similarity matrix, AHC with Ward linkage and number of trips between stations as a similarity matrix adapted from Lahoorpoor *et al.* (2019) (Section 3.2.1), and two multi-objective mathematical models (Section 3.2.2).

To select the best clustering approach, we set our performance measures, which will later define the objective function components for the mathematical models, as follows:

- (P1) the total in-cluster Manhattan distance, that shows the compactness of the cluster,
- (P2) the deviation of the total in-cluster demand from zero, that shows whether the clusters are self-sufficient, and
- (P3) the deviation of number of stations per cluster from the average number of stations per cluster, that shows whether the number of stations visited by a rebalancing vehicle is balanced among clusters.

Next, we present these four clustering methods.

3.2.1 Agglomerative hierarchical clustering (AHC) with Ward linkage

In AHC, each element is treated as a singleton cluster at the beginning of the algorithm. This bottom-up approach connects a pair of clusters that are the most similar to produce a bigger cluster. The algorithm halts as soon as all the elements are in one cluster.

The data is used to compute the similarity (dissimilarity) matrix between each pair of elements in the data set. According to a linkage function, the closest elements (or clusters) are grouped together at one higher level in the hierarchy, which forms the dendrogram. Then, the decision maker determines a convenient level to cut the dendrogram, which also corresponds to the number of clusters.

There are several linkage functions introduced in the literature such as single, complete, group average, and Ward. This paper considers Ward linkage, which aims to minimize total within-cluster variance. This linkage is chosen since different similarity measures can be used.

Proximity as a similarity matrix This method uses the proximity of two stations, which corresponds to the physical distance between a pair of stations, as a similarity matrix. The advantage of this method is that it produces geographically convenient clusters, helping to improve (P1). On the other hand, it does not pay regard to the performance measures (P2) and (P3).

Number of trips as a similarity matrix Different similarity matrices may also be used. Lahoorpoor *et al.* (2019) introduce a methodology for clustering BSS stations using the number of trips from one station to the other as a similarity matrix. This matrix is created using the origin-destination trip information. This way, they claim that they can discover the groups of stations which interact the most. In other words, the constructed clusters are more likely to be self-sufficient, which implies better performance in (P2).

3.2.2 Multi-objective mathematical model approach

Although AHC is convenient in terms of computation time, it does not take multiple objectives into account. Therefore, we develop two multi-objective mathematical models, that consider all the three performance measures, i.e., (P1), (P2), and (P3).

MINLP The first mathematical model, given by ($C3N$), is a mixed integer non linear model. The corresponding notation is given in Table 2. The objective function components, i.e., (17), (18), and (19), consider all the three performance measures, i.e., (P1), (P2), and (P3), respectively.

The objective function (17) minimizes the in-cluster distance, which is a sum of all the L_1 distances between each station. The second objective (18) aims to minimize the positive and negative total deviation from the zero total demand within a cluster. Lastly, the third objective (19) minimizes the deviation of number of stations across the clusters. The objective weights, i.e., α, β, and γ, help to obtain a single objective. One should note that the problem can also be solved using lexicographic approach, without the use of weights.

Eq. (20) enforces that all stations, that have non zero demand, are assigned to one and only one cluster. Eq. (21) ensures that the distance between each pair of stations in a cluster is determined as in-cluster distance. Eq. (22) determines the positive and negative deviation from the zero total demand within a cluster, whichever is applicable. Similarly, Eq. (23) detects the positive

Table 2: Notation for the model given by ($C3N$)

Parameters	
N	number of stations ($i,\ j \in \{1, ..., N\}$)
C	number of clusters ($c \in \{1, ..., C\}$)
lon_i, lat_i	the longitude and latitude of station i, $i \in N$, respectively
d_{ij}	the distance from station i to station j, $i,j \in N$
q_i	the demand at each station, $i \in N$
α, β, γ	weight of 1^{st}, 2^{nd} and 3^{rd} objective function, respectively
Decision variable	
s_{ic}	1 if station i is assigned to cluster c, 0 otherwise, $i \in N, c \in C$
Auxiliary decision variables	
$inClusterDist_c$	the total Manhattan distance between each pair of stations in cluster c, $c \in C$
m_{ijc}	1 if both i and j are in cluster c, 0 otherwise, $i, j \in N, c \in C$

and negative deviation of number of stations across the clusters. Finally, (24)-(27) ensure that the domain constraints of the decision variables are satisfied.

$$(C3N)\ \min \quad \sum_{c\in C} \alpha{\cdot}inClusterDist_c \tag{17}$$

$$+ \sum_{c\in C} \beta{\cdot}(devD_c^+ + devD_c^-) \tag{18}$$

$$+ \sum_{c\in C} \gamma{\cdot}(devSN_c^+ + devSN_c^-) \tag{19}$$

$$\text{s.to} \quad \sum_{c\in C: q_i\neq 0} s_{ic} = 1 \qquad \forall i \in N \tag{20}$$

$$\sum_{i,j\in N: j\geq i} s_{ic}\cdot s_{jc}\cdot d_{ij} = inClusterDist_c \qquad \forall i,j \in N, \forall c \in C \tag{21}$$

$$\sum_{i\in N} s_{ic}\cdot q_i = devD_c^+ - devD_c^- \qquad \forall c \in C \tag{22}$$

$$\sum_{i\in N} s_{ic} = \frac{N}{C} + devSN_c^+ - devSN_c^- \qquad \forall c \in C \tag{23}$$

$$s_{ic} \in \{0,1\} \qquad \forall i \in N, c \in C \tag{24}$$

$$devSN_c^+, devSN_c^- \geq 0 \qquad \forall c \in C \tag{25}$$

$$devD_c^+, devD_c^- \geq 0 \qquad \forall c \in C \tag{26}$$

$$inClusterDist_c \geq 0 \qquad \forall c \in C \tag{27}$$

This non linear model is linearized by introducing an auxiliary variable, m_{ijc}, that represents the multiplication of s_{ic} and s_{jc}. With this, (21) is replaced by (28)-(32), which are given below:

$$m_{ijc} \leq s_{ic} \qquad \forall i,j \in N, \forall c \in C \tag{28}$$

$$m_{ijc} \leq s_{jc} \qquad \forall i,j \in N, \forall c \in C \tag{29}$$

$$m_{ijc} \geq s_{ic} + s_{jc} - 1 \qquad \forall i,j \in N, \forall c \in C \tag{30}$$

$$\sum_{i,j\in N: j\geq i} m_{ijc}\cdot d_{ij} = inClusterDist_c \qquad \forall i,j \in N, \forall c \in C \tag{31}$$

$$m_{ijc} \in \{0,1\} \qquad \forall i,j \in N, \forall c \in C \tag{32}$$

We call the linearized model as $(C3)$. The complexity of $(C3)$ is given by $O(N^2 \cdot C)$. As the number of clusters indirectly depend on the number of stations, i.e., a function of number of stations, the complexity can be identified as $O(N^3)$.

MILP Given the complexity of the $(C3)$, we develop another model to solve the clustering problem. In addition to the ones in $(C3)$, we introduce two other decision variables. These determine the cluster center and the in-cluster distance is calculated by summing the distance to

Table 3: Additional notation for the model given by ($C4$)

Parameters	
M	big-M value
Decision variables	
$lonC_c$, $lonC_c$	the longitude and latitude of cluster c, $c \in C$, respectively
Auxiliary decision variables	
$devSN_c^+$, $devSN_c^-$	the positive and negative deviation of number of stations in cluster c from the average number of stations per cluster, $c \in C$, respectively
$devD_c^+$, $devD_c^-$	the positive and negative deviation of total demand from 0 in cluster c, $c \in C$, respectively
$diffLon_{ic}$	the distance in longitude between station i and cluster c, $i \in N, c \in C$
$diffLat_{ic}$	the distance in latitude between station i and cluster c, $i \in N, c \in C$
md_{ic}	the Manhattan distance between station i and cluster c, $i \in N, c \in C$

these centers from all the elements in that cluster. Some additional decision variables help us to calculate the in-cluster distance. The additional notation is given in Table 3 and the developed model is presented as ($C4$).

$$
\begin{aligned}
(C4)\ \min \quad & (17) + (18) + (19) & & \\
\text{s.to} \quad & (20) & & \\
& lon_i - lonC_c \leq diffLon_{ic} & \forall i \in N, \forall c \in C & (33)\\
& lonC_c - lon_i \leq diffLon_{ic} & \forall i \in N, \forall c \in C & (34)\\
& lat_i - latC_c \leq diffLat_{ic} & \forall i \in N, \forall c \in C & (35)\\
& latC_c - lat_i \leq diffLat_{ic} & \forall i \in N, \forall c \in C & (36)\\
& diffLon_{ic} + diffLat_{ic} \leq md_{ic} + M \cdot (1 - s_{ic}) & \forall i \in N, \forall c \in C & (37)\\
& \sum_{i \in N} md_{ic} \leq inClusterDist_c & \forall c \in C & (38)\\
& (22), (23), (24) & & \\
& diffLon_{ic}, diffLat_{ic}, md_{ic} \geq 0 & \forall i \in N, c \in C & (39)\\
& lonC_c, latC_c \geq 0 & \forall c \in C & (40)\\
& (25), (26), (27) & &
\end{aligned}
$$

The constraints (33)-(36) calculate the absolute distance between each cluster center and station. Then, these are used to calculate the manhattan distance (Eq. (37)), which helps us to determine the total in cluster distance (Eq. (38)). Note that, this value is different than the value obtained in ($C3$). The Big-M value is set to the maximum possible distance between two stations, i.e., $(\max lon_i - \min lon_i) + (\max lat_i - \min lat_i)$. Eqs. (39) and (40) ensure the domain constraints are satisfied.

The complexity of this model is given by $O(N \cdot C)$. Although it does not express exponential complexity at a first glance, the number of clusters being a function of number of stations makes the complexity of the ($C4$) $O(N^2)$. Therefore, we can conclude that ($C4$) is expected to perform better than ($C3$).

4 Computational experiments

To evaluate the methodology two case studies are selected, i.e., nextbike Sarajevo, Bosnia and Herzergovina and nextbike Berlin, Germany. The main motivation of choosing these two systems is to evaluate demand forecast in different scenarios represented by system size, geo location, landscape, economic strength, etc.

4.1 Data and analysis

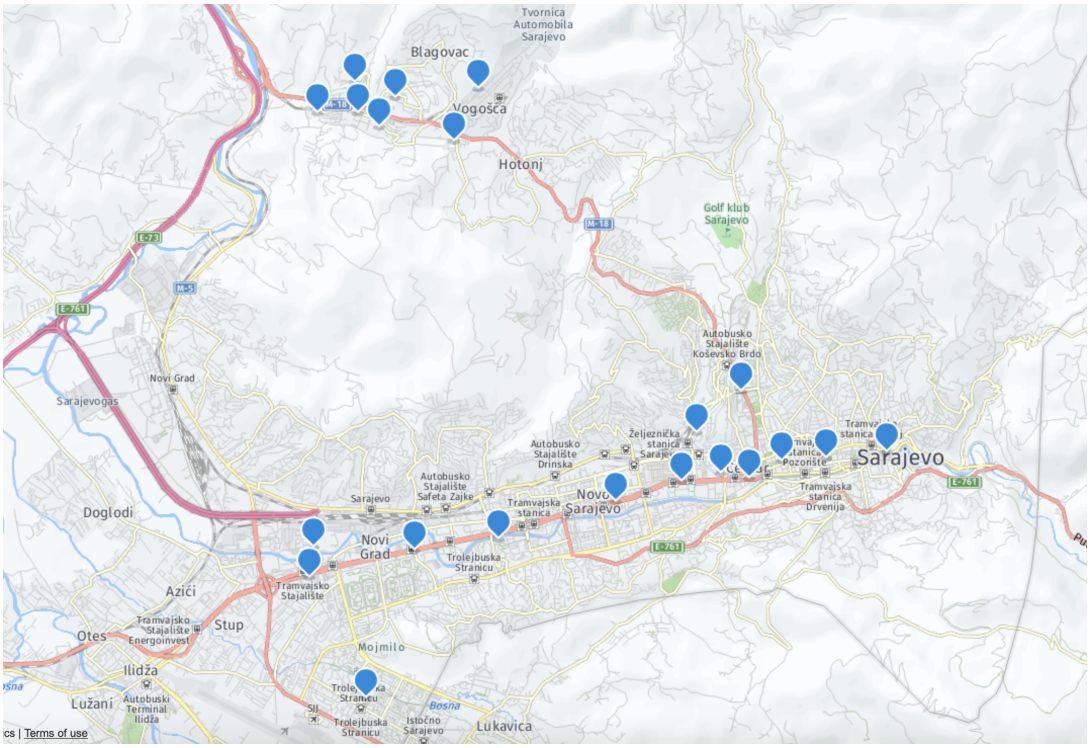
The network of the nextbike Sarajevo and nextbike Berlin BSS stations can be seen in Figure 1. nextbike Sarajevo BSS has 21 stations with approximately 120 bikes operating, whilst nextbike Berlin BSS operates with around 3000 bikes in 298 stations.

In Sarajevo, we see that the network is divided into two main station clusters. This is due to the existence of a small hill in between the two sub networks. On the other hand, in Berlin we do not see any accummulation of stations at some specific parts of the city. This is expected since the city is mostly flat, which makes it easy to access any part of it by bike.

In Figure 3, we see the time series plot of number of pick-ups (blue straight line) and drop-offs (red dashed line), where each row corresponds to a weekday. The data for the presented graphs date from April 5, 2021 Monday to April 11, 2021 Sunday. The first row shows the plot for Monday, the second for Tuesday, so on and so forth. As it is conducted in the literature, the trips that have unreasonable trip length and duration are cleaned from the data sets.

Figure 1: nextbike BSS stations

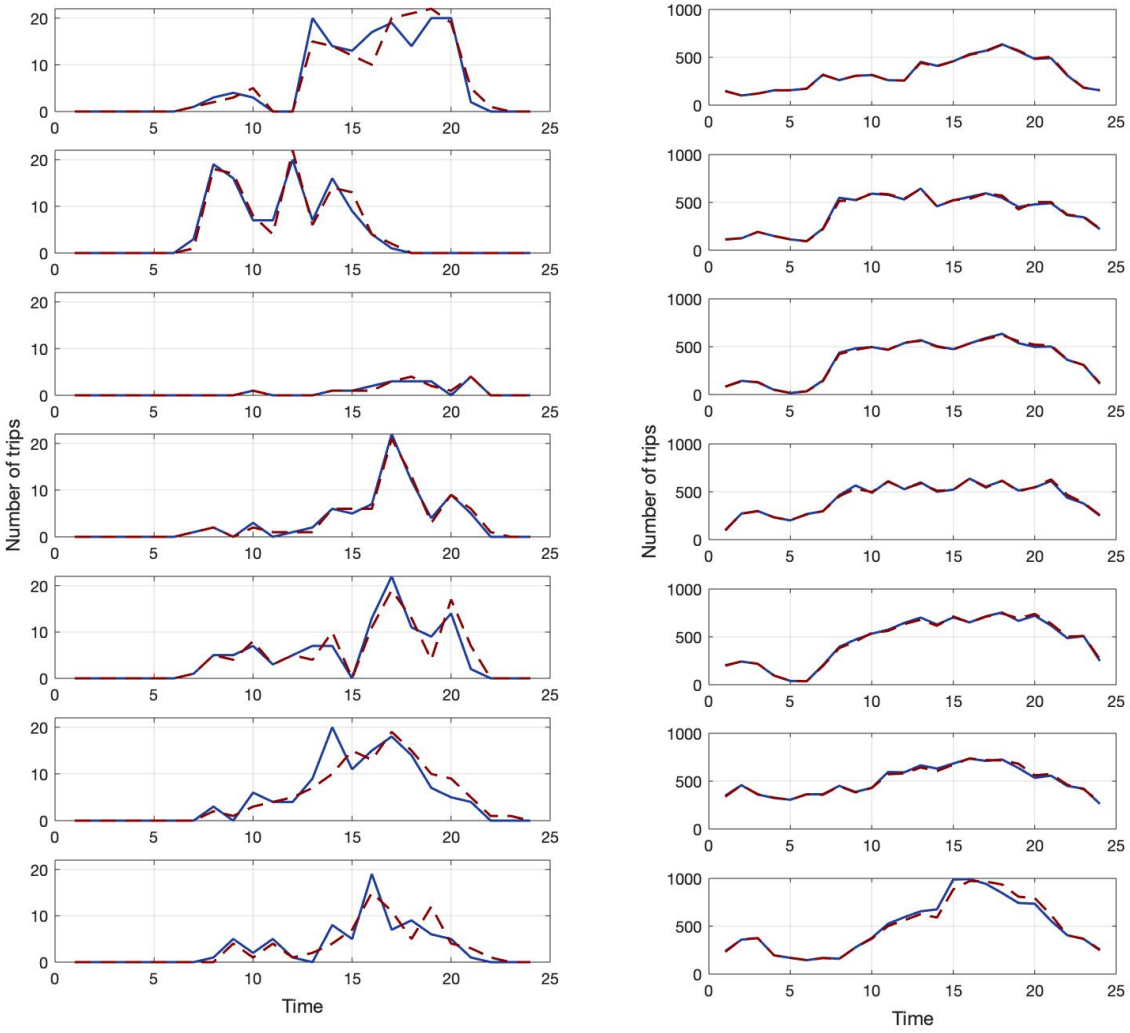
(a) Sarajevo



(b) Berlin

Figure 3: Number of pick-ups and drop-offs over a week

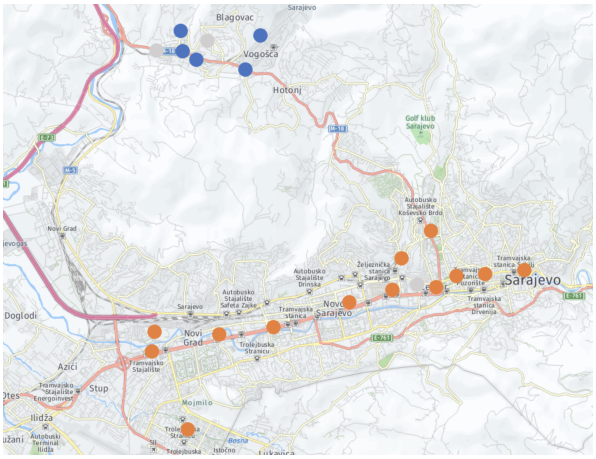
(a) Sarajevo (b) Berlin



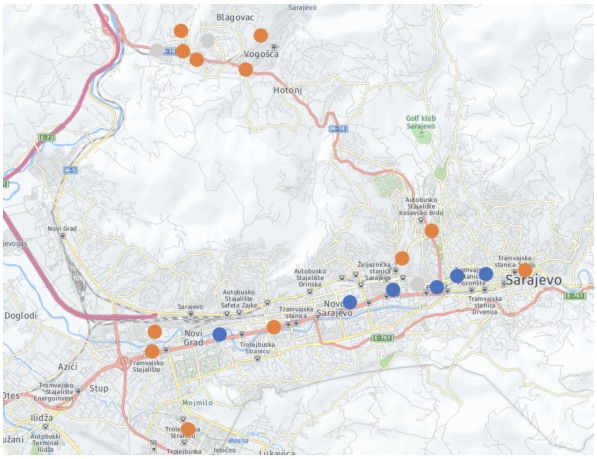
The unexpected behavior in Sarajevo on April 7, 2021, Wednesday can be explained by the snowfall that occured during the previous night. In Berlin, we see that the demand increases on average on the weekends. Furthermore, the trips are longer on the weekends whereas people prefer shorter trips on the weekdays. This can be explained by the fact that people tend to use the BSS more for leisure during the weekend compared to the weekdays. The same applies to Sarajevo case study as well.

Figure 5: Sarajevo - 2 Clusters

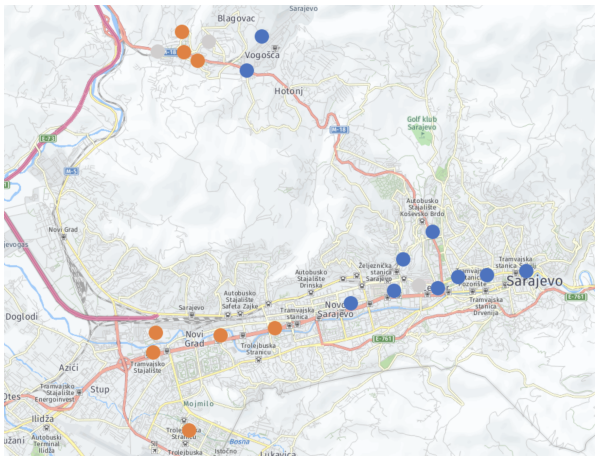
(a) Clustering with ($C1$)



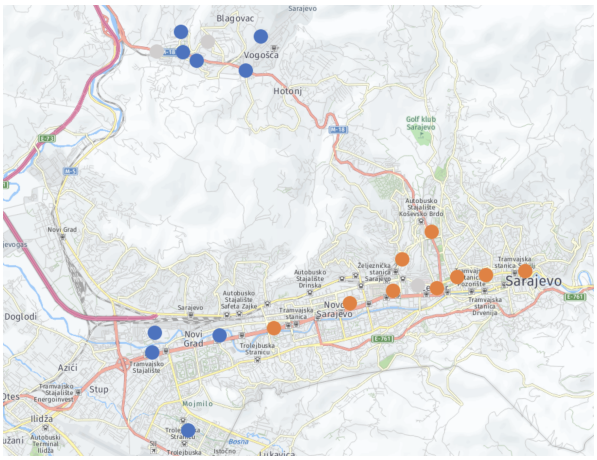
(b) Clustering with ($C2$)



(c) Clustering with ($C4_{DD}$)



(d) Clustering with ($C4_{ICD}$)



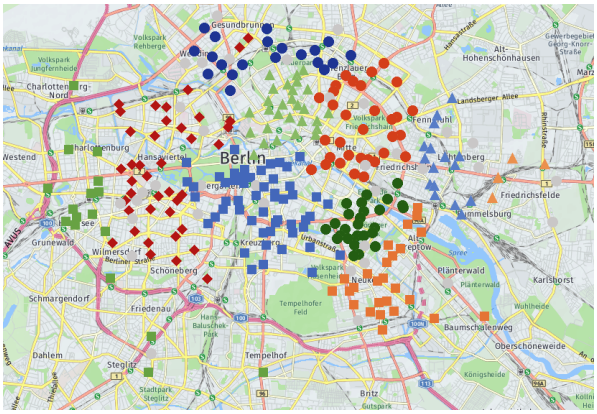
4.2 Results and discussion

The optimization models are implemented on a computer with 8 GB RAM and 2.3 GHz Intel Core i5 processor in *python* and *python* API for CPLEX 12.10.

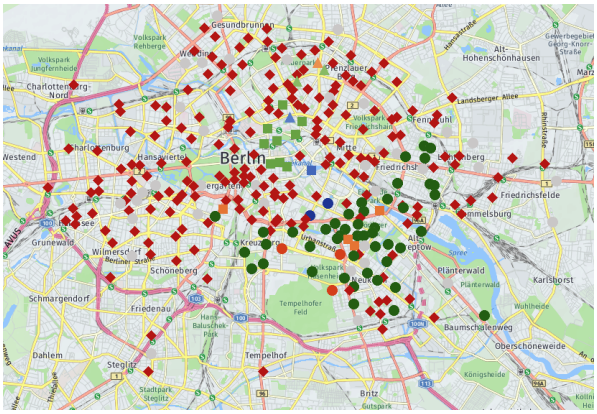
We give the resulting clusters with the four different clustering methods for Sarajevo and Berlin, in Figure 5 and Figure 7, respectively. Here, ($C1$) and ($C2$) correspond to AHC with similarity matrix as proximity matrix and OD-trips matrix, respectively. ($C4_{DD}$) solves ($C4$) where the minimizing the deviation of demand is the most important objective, i.e., $\beta > \alpha > \gamma$, whilst for ($C4_{ICD}$) solves ($C4$) where the objective function coefficient favor minimizing the in-cluster distance, , i.e., $\alpha > \beta > \gamma$. Note that ($C3$) is not included here because it is intractable to solve it in real time.

Figure 7: Berlin - 10 Clusters

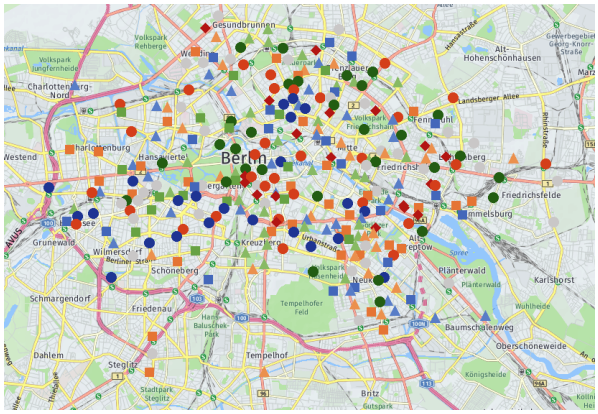
(a) Clustering with ($C1$)



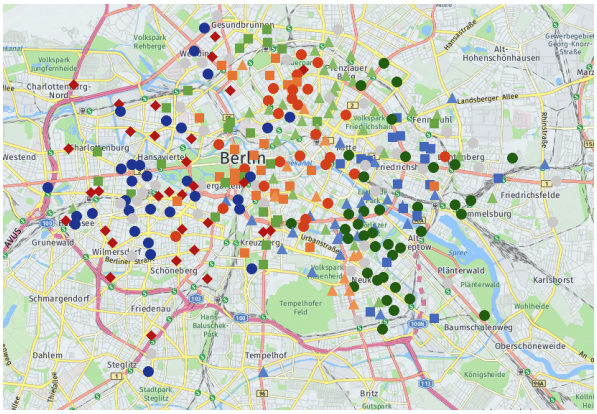
(b) Clustering with ($C2$)



(c) Clustering with ($C4_{DD}$)



(d) Clustering with ($C4_{ICD}$)



Given that the units of the three objective functions are different than each other, we first try lexicographic method. However, this method is not able to produce any solutions in real time. Therefore, we assign extreme values as the objective weights, to approximate the lexicographic method.

As expected, ($C1$) creates geographically convenient clusters. Although for Berlin, the number of stations per cluster does not differ among different clusters, this is not true for Sarajevo case study. Here, we see the influence of city structure. The Sarajevo case study results confirm that the number of stations per cluster is not considered by ($C1$).

We observe that the clusters that ($C2$) has resulted in are unbalanced in terms of number of stations for the Berlin data set. Specifically, among ten clusters, three of them have 9, 52, and 225 stations, whilst the remaining seven clusters have at most three stations.

With ($C4_{DD}$), we see that the most of the clusters span the whole service area. This is due to the

Table 4: Rebalancing cost for all the clustering methods

Dataset	# of clusters	$(C1)$	$(C2)$	$(C4_{DD})$	$(C4_{ICD})$
Sarajevo	2	9.728	15.591	12.709	12.627
Berlin	10	75.351	372.332	139.880	126.983
	15	83.393	103.923	163.261	173.630
	20	90.471	120.289	159.271	197.483

fact that the second objective, i.e., minimizing the deviation from zero total demand, is the most important objective. On the other hand, the results of $(C4_{ICD})$ show more collective clusters compared to $(C4_{DD})$. This is expected since the first objective, i.e., minimizing the in-cluster distance, is more important than the other two objectives. However, it should be noted that the results of both $(C4_{DD})$ and $(C4_{ICD})$ are found with a large optimality gap. This indicates that the found solutions might be far from the optimal.

The costs resulted from application of $(C2)$ increase as well compared to $(C1)$. The accummulation in few stations would explain this increase. The resulting solutions of the rebalancing optimization with different clustering methods are given in Table 4. Note that these values represent the total kilometers driven by the trucks, which linearly determines the rebalancing cost. Compared to $(C1)$, both $(C4_{DD})$ and $(C4_{ICD})$ produce more scattered clusters. The increase in kilometers traveled by the trucks can be explained by this fact. We see that, the additional demand-based objective does not work well, i.e., results in increase in rebalancing costs.

5 Conclusion and future work

The rebalancing operations in VSSs are one of the mostly used approaches to keep the balance of the vehicles according to the trip demand to increase the level of service. When large systems are in consideration, the rebalancing optimization problem becomes hard to solve. Therefore, heuristic approaches such as branch-and-bound and station clustering, are used to divide the problem into sub problems. In this study, we utilize several clustering methods to test their effects on the rebalancing optimization. Among four clustering approaches considered, two of them are standard hierarchical clustering algorithms, while the remaining are specifically crafted for the original problem at stake, i.e., to take into account the rebalancing-related objectives.

We perform computational experiments on two BSSs, nextbike Sarajevo, Bosnia and Herzegovina, and nextbike Berlin, Germany. The results show that with AHC using proximity as a similarity measure, we obtain geographically collective clusters. When the similarity measure is changed to number of trips between stations, some clusters become extra big whereas the rest remains very small, i.e., at most three stations per cluster. With the proposed mathematical models, this problem is overcome. However, the clusters tend to spread throughout the service area especially when the zero total demand is considered as the most important objective. In other words, the areas spanned by each cluster tend to overlap some other clusters. Consequently, larger distances are driven by rebalancing trucks within clusters. This results in higher rebalancing costs although the areas are self-sufficient, that means no inter-cluster rebalancing is required.

Since we were able to solve the mathematical models with a large optimallity gap, the future work includes a heuristic approach for multi-objective station clustering. Other data sets might be considered to derive conclusions about the relation between the city and demand structure, and the clustering and rebalancing optimization results.

6 References

Ataç, S., N. Obrenovic and M. Bierlaire (2020) Vehicle sharing systems: Does demand forecasting yield a better service?, paper presented at the *20th Swiss Transport Research Conference*.

Ataç, S., N. Obrenović and M. Bierlaire (2021) Vehicle sharing systems: A review and a holistic management framework, *EURO Journal on Transportation and Logistics*, **10**, 100033, ISSN 2192-4376.

Boyaci, B., K. G. Zografos and N. Geroliminis (2017) An integrated optimization-simulation framework for vehicle and personnel relocations of electric carsharing systems with reservations, *Transportation Research Part B: Methodological*, **95**, 214 – 237, ISSN 0191-2615.

Chemla, D., F. Meunier and R. Wolfler Calvo (2013b) Bike sharing systems: Solving the static rebalancing problem, *Discrete Optimization*, **10** (2) 120 – 146, ISSN 1572-5286.

Cruz, F., A. Subramanian, B. P. Bruck and M. Iori (2017) A heuristic algorithm for a single vehicle static bike sharing rebalancing problem, *Computers & Operations Research*, **79**, 19 – 33, ISSN 0305-0548.

Dell'Amico, M., E. Hadjicostantinou, M. Iori and S. Novellani (2014) The bike sharing rebalancing problem: Mathematical formulations and benchmark instances, *Omega*, **45**, 7–19, ISSN 0305-0483.

Erdogan, G., G. Laporte and R. Wolfler Calvo (2014) The static bicycle relocation problem with demand intervals, *European Journal of Operational Research*, **238** (2) 451–457, ISSN 0377-2217.

Feng, Y., R. C. Affonso and M. Zolghadri (2017) Analysis of bike sharing system by clustering: the velib' case, *IFAC-PapersOnLine*, **50** (1) 12422–12427, ISSN 2405-8963. 20th IFAC World Congress.

Fishman, E. (2016) Bikeshare: A review of recent literature, *Transport Reviews*, **36** (1) 92–113.

Ho, S. C. and W. Szeto (2017) A hybrid large neighborhood search for the static multi-vehicle bike-repositioning problem, *Transportation Research Part B: Methodological*, **95**, 340 – 363, ISSN 0191-2615.

Lahoorpoor, B., H. Faroqi, A. Sadeghi-Niaraki and S.-M. Choi (2019) Spatial cluster-based model for static rebalancing bike sharing problem, *Sustainability*, **11** (11), ISSN 2071-1050.

Laporte, G., F. Meunier and R. Calvo (2018) Shared mobility systems: an updated survey, *Annals of Operations Research*, **271**, 12 2018.

Liu, J., L. Sun, W. Chen and H. Xiong (2016) Rebalancing bike sharing systems: A multi-source data smart optimization, paper presented at the *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1005–1014, 08 2016.

Ma, X., R. Cao and Y. Jin (2019) Spatiotemporal clustering analysis of bicycle sharing system with data mining approach, *Information*, **10** (5), ISSN 2078-2489.

Midgley, P. (2011) Bicycle-sharing schemes: enhancing sustainable mobility in urban areas, *United Nations, Department of Economic and Social Affairs*, **8**, 1–12.

Miller, C. E., A. W. Tucker and R. A. Zemlin (1960) Integer programming formulation of traveling salesman problems, *J. ACM*, **7** (4) 326–329, October 1960, ISSN 0004-5411.

Nair, R. and E. Miller-Hooks (2011) Fleet management for vehicle sharing operations, *Transportation Science*, **45** (4) 524–540, ISSN 00411655, 15265447.

Pachauri, R. K., M. R. Allen, V. R. Barros, J. Broome, W. Cramer, R. Christ, J. A. Church, L. Clarke, Q. Dahe, P. Dasgupta *et al.* (2014) *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*, Ipcc.

Pal, A. and Y. Zhang (2017) Free-floating bike sharing: Solving real-life large-scale static rebalancing problems, *Transportation Research Part C: Emerging Technologies*, **80**, 92–116, ISSN 0968-090X.

Pfrommer, J., J. Warrington, G. Schildbach and M. Morari (2014) Dynamic vehicle redistribution and online price incentives in shared mobility systems, *IEEE Transactions on Intelligent Transportation Systems*, **15** (4) 1567–1578, Aug 2014, ISSN 1558-0016.

Raviv, T., M. Tzur and I. A. Forma (2013) Static repositioning in a bike-sharing system: models and solution approaches, *EURO Journal on Transportation and Logistics*, **2**, 187–229.

Schuijbroek, J., R. Hampshire and W.-J. van Hoeve (2017) Inventory rebalancing and vehicle routing in bike sharing systems, *European Journal of Operational Research*, **257** (3) 992 – 1004, ISSN 0377-2217.

Shaheen, S., S. Guzman and H. Zhang (2010) Bikesharing in europe, the americas, and asia: Past, present, and future, *Institute of Transportation Studies, UC Davis, Institute of Transportation Studies, Working Paper Series*, **2143**, 01 2010.

Shaheen, S., D. Sperling and C. Wagner (1998) Carsharing in europe and north america: Past, present, and future, *Transportation Quarterly*, **52**, 06 1998.

Shu, J., M. Chou, Q. Liu, C. Teo and I.-L. Wang (2013) Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems, *Operations Research*, **61**, 1346–1359, 11 2013.

A Formulation (F1) from Dell'Amico *et al.* (2014)

$$(F1)\ \min \quad \sum_{i\in V}\sum_{j\in V} c_{ij}x_{ij} \tag{41}$$

$$\text{s.to} \quad \sum_{i\in V} x_{ij} = 1 \qquad \forall j \in V\setminus\{0\} \tag{42}$$

$$\sum_{i\in V} x_{ji} = 1 \qquad \forall j \in V\setminus\{0\} \tag{43}$$

$$\sum_{j\in V} x_{0j} \le m \tag{44}$$

$$\sum_{j\in V\setminus\{0\}} x_{0j} - \sum_{j\in V\setminus\{0\}} x_{j0} = 0 \tag{45}$$

$$\sum_{i\in S}\sum_{j\in S} x_{ij} \le |S| - 1 \qquad \forall S \subseteq V\setminus\{0\}, S \neq \emptyset \tag{46}$$

$$\theta_j \ge \max\{0, q_j\} \qquad \forall j \in V \tag{47}$$

$$\theta_j \le \min\{Q, Q + q_j\} \qquad \forall j \in V \tag{48}$$

$$\theta_j - \theta_i + M(1 - x_{ij}) \ge q_j \qquad \forall i \in V, j \in V\setminus\{0\} \tag{49}$$

$$\theta_i - \theta_j + M(1 - x_{ij}) \ge q_j \qquad \forall i \in V\setminus\{0\}, j \in V \tag{50}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in V \tag{51}$$

The objective function (41) minimizes the cost. (42) and (43) make sure that every node except the depot is served exactly once. Following two sets of constraints, (44) and (45), assure that no more than m vehicles are used and all used vehicles return to the depot at the end of their route. The constraint set (46) is a typical cutset constraint which is used for subtour elimination. (47) and (48) defines the upper bound on the load of a vehicle. The flow conservation is achieved by (49) and (50). Last constraint set (51) imposes binary restrictions on decision variables x_{ij}'s.