Student thesis series INES nr 562

Optimizing the locations of bike-sharing stations using GPS-based trip data: A Spatio-temporal demand coverage approach

Birkan Caliskan

2021
Department of
Physical Geography and Ecosystem Science
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden

Birkan Caliskan (2021).
Optimizing the locations of bike-sharing stations using GPS-based trip data: A Spatio-temporal demand coverage approach

Master degree thesis, 30 credits in *Geomatics*
Department of Physical Geography and Ecosystem Science, Lund University

Level: Master of Science (MSc)

Course duration: *January* 2021 until *September* 2021

Optimizing the locations of bike-sharing stations using GPS-based trip data: A Spatio-temporal demand coverage approach

---

Birkan Caliskan

Master thesis, 30 credits, in *Geomatics*

Supervisors:

Pengxiang Zhao
Dep. of Physical Geography and Ecosystem Science, Lund University

Ali Mansourian
Dep. of Physical Geography and Ecosystem Science, Lund University

Exam committee:

Micael Runnström
Dep. of Physical Geography and Ecosystem Science, Lund University

Olive Niyomubyeyi
Dep. of Physical Geography and Ecosystem Science, Lund University

Acknowledgements

I would like to express my appreciation to Pengxiang Zhao and Ali Mansourian, my supervisors, for their support and useful suggestions on this project. I would also like to thank Micael Runnström, Olive Niyomubyeyi and Peter Nezval for their constructive feedback. Lastly, my thanks go to my colleagues from the Department of Physical Geography and Ecosystems Science and fellow students from Lund University.

ABSTRACT

Birkan Caliskan: Optimizing the locations of bike-sharing stations using GPS-based trip data: A Spatio-temporal demand coverage approach

Bicycle sharing systems are increasingly embraced by cities around the world in recent years. Their benefits over other transportation modes in certain frames encourage traffic/urban planners, and public authorities to establish such systems in urban areas. However, planning the placement, size and operation of such beneficial micro-mobility mode is a vital issue that concerns designers. This master thesis study aims to develop an optimization framework for an optimal design of bicycle sharing locations and capacities. Accordingly, this study proposed a two-stage optimization framework to design suitable placement of bike-sharing stations using user demand derived from a big bicycle GPS dataset in Shanghai, China. The proposed framework aims to achieve optimized placement and size by maximizing spatiotemporal demand coverage of bicycle trips. The implementation in the real-world scenario demonstrates that it can finely optimize the placement and the capacity of the bicycle sharing stations. The assessment part also proved that the suggested model performs better coverage of user demands concerning coverage scores of the proposed model and two other optimization methods.

Keywords: Geomatics, Bike-sharing, Genetic Algorithm, Location-allocation, Site selection, Optimization

Table of Contents

List of Figures

List of Tables

1 Introduction

Bicycle sharing systems are recently widely used in many countries as a sustainable alternative to motorized vehicles. Cities are implementing bicycle sharing systems including pedal bikes and e-bikes due to their substantial benefits. Today, there are 2007 bicycle sharing systems with about 9,440,776 bicycles in operation around the world, and 299 new systems are being planned or constructed (Meddin et al., 2020).

A typical bike-sharing system is established, known as a docked bicycle system allowing the users to rent bicycles among stations, on a service. The users borrow the bicycle from a docked station by the help of the smartphone application or embedded computer at the station. After the trip, the rented bicycle is returned and locked to the same or any other point in the system.

Previous studies on the impact of the presence of bicycle sharing schemes proved environmental, economic and social benefits with the substitution from private motorized vehicles. Kou et al., (2020) quantified the environmental advantages for unit distance travelled considering in-depth trip information such as trip distance, purpose and time of the day for the trip. Results indicated that the reduction of greenhouse gas pollution ranges from 283 to 581 g CO2-eq. Otero et al., (2018) assessed the health benefits of substitution from private car trips to bicycle sharing stations trips including e-bikes in 12 cities in Europe. Their assessment indicates that health benefits outweighed the health risk with a ratio of 19:1 and 0.01-0.13 deaths per 1000 bikes avoided, with lower benefits for e-bikes. Hamilton and Wihman (2018) estimate that the presence of bicycle sharing stations results in a roughly 4% reduction in traffic congestion with a huge economic benefit in private and public extent.

Bicycle sharing systems rule out the obligation of permanent possession of bicycles, which prevents potential theft crimes as well as the storage problem of bicycles (Shaheen et al. 2010). Yet, although theft crimes are an unsolved issue for bicycle sharing systems, GPS tracking allows substantial after crime monitoring. Additionally, another burden for e-bikes owners is battery charging.

In some cities where traffic jams are the substantial problems due to the huge private cars population, especially during peak hours, replacing transportation mode with bicycles does not have the velocity drawback, in fact, the mobility of bicycle during traffic jam propose a quicker travel experience (Jensen et al. 2010). Moreover, difficulty in finding a parking slot during peak times, and a long walking distance from the parking slot to the destination discourage users to drive private cars. Farhih-Imani et al. (2017) found that during weekdays trips with less than 3km, bicycle sharing systems are an alternative either faster or competitive to taxi trips in dense urban areas. They examined the taxi

trips in a radius of 250m of stations of origin and destination since such taxi patterns can be substituted by bicycle sharing system trips. Therefore, bike share schemes are faster and convenient than walking and driving in heavily congested areas. However, substitution from walking to cycling rules out environmental benefits.

1.1 Research objectives and questions

This study primarily aims to develop a placement model that maximizes the spatio-temporal demand coverage of shared bicycle stations. Accordingly, following research objectives are devised to achieve this.

- Investigating spatio-temporal rider patterns and behaviour on bicycle trips from a 2-week real GPS trajectory dataset.
- Determining thresholds for statistical indicators of bike use to clean out data from insignificant trips in order to obtain major origin-destination flow.
- Develop and formulate a mathematical location-allocation model in respect to certain requirements and constraints, and accordingly create a script that can be applied to design optimal site placement.
- Apply the proposed model to a real-world scenario in order to assess the performance of the framework against traditional optimization methods such as POI-based and population density optimization.

These objectives are achieved by conducting an intensive literature review and attempts with different units and parameters to develop an accurate model to solve the location-allocation problem.
This study proposes a genetic algorithm-based method to optimize the location of bicycle stations and the number of lockers. The research questions to be addressed are as follows:

- How can locations and sizes of bicycle sharing stations be optimally established based on the maximal spatio-temporal coverage model?
- What are the best suited genetic parameters (population size, mutation and crossover rate) for use with the GPS data?
- To what extent the optimized bicycle sharing stations can cover existing free-floating bike demand?
- Can the proposed optimization method successfully optimize the location and capacity of bicycle sharing stations against other proposed location-allocation models in the literature?

2 Background

In this section, existing studies on facility location problems, algorithms to optimize location-related problems, and factors that impact the design of the bike share stations are reviewed. Literature review was carried out to identify previous optimization models and frameworks on site selection problems.

2.1 Facility location problem

Facility location can be described as determining the optimal location of a set of facilities while minimizing the cost and considering some constraints (Hale and Moberg, 2003). The search space of facilities is mostly divided into three categories: discrete, continuous and network spaces.

- Discrete space: a set of certain locations is used as an input for the optimal placement.
- Continues space: every point in the search space is a potential optimal location for the placement.
- Network space: particular lines and arcs are described as a search space (Hale and Moberg, 2003).

Due to the fact that facility location is a common problem for several disciplines, civil, industrial and electrical engineers, transportation/urban planners, and geographers are involved in to study. Thus, there is now a wide range of literature published. A few of studies to mention are: emergency humanitarian logistical problems to minimize the accessibility cost (Boonmee et al. 2017), optimal location and sizing of renewable distributed generations to minimize the electrical losses and enhance the voltage profile (Ali et al. 2017), finding ideal drone delivery launch location to achieve cost-efficient and faster delivery operations while minimizing either the total cost or delivery completion time (Salama et al. 2020). Although the facility location is the common problem for the aforementioned literature, they differ in the number and size of facilities, cost and objective functions to minimize and maximize. Therefore, the facility location model developed on those metrics derived from the objectives are unique formulations.

2.2 Algorithms for optimization problems

Optimization in this study can be defined as finding the optimal set of values in respect to given problems while satisfying the desired objectives and constraints. Travel salesman problem, a phenomenon that seeks the shortest path among a set of discrete points (station, cities) is an example of an optimization problem. A set of paths between target points in order is an expected optimal solution for this problem. Besides location science, applications of optimization are commonly used

in other fields such as medicine, business and engineering (Ghaheri et al. 2015; Zelenkov et al. 2017; Bouktif et al. 2018).

Brute-force algorithms are simple to implement due to their plain structure that performs selection of a possible set of candidates based on enumerated process from 1 to n in the search space. However, as the search spaces increase, the number of solutions for candidates increases. Additionally, it is difficult to implement objectives and constraints for candidate solutions into a simple enumerating method. Li et al. (2009) defined and formulated the optimization problems of implementing brute-force methods for vast search space, which cause them to be avoided.

Since brute-force methods are not capable to solve complex facility location problems that have huge data to handle due to the huge exponentially increasing solution space proportionately to the dataset, quite a few algorithms were developed to facilitate the optimization procedure. Those heuristic algorithms produce near-best solutions easily, with less computation effort. The most commonly used algorithms are genetic algorithms (Holland 1975), ant colony algorithms (Dorigo et al. 1991), simulated annealing (Kirkpatrick 1984), and swarm intelligence (Kennedy and Eberhart 1995).

Genetic algorithm mimics the theory of evaluation and the law of Mendelian inheritance to simulate the process of natural evaluation (Goldberg 1988). It aims to find the most optimized solution among high number of proposals by their fitness score derived from the function constructed by objectives and constraints of the problem. Zhang et al. (2016) proposed genetic algorithm-based multi-objective optimization approach to optimize the location of health care facilities in Hong Kong. Dong et al. (2014) examined a genetic algorithm that was applied to optimize locations for electric vehicle charging stations based on charging behaviour and the budget constraint in the real-world driving context.

Ant colony optimization (Dorigo et al., 1991) is a heuristic computation algorithm consisting of many cooperating simple agents with low-level interactions among them that mimic the natural behaviour of ant colonies. In the current literature, the ant colony optimization is mostly applied to the problem of balancing a bicycle sharing system that aims to adjust empty and filled lockers among stations. Gaspero et al. (2013) proposed a hybrid metaheuristic method combining an ant colony optimization and constraint programming which ultimately minimizes the cost for working effort (absolute deviation among target bikes and total travel time) of balancing bike-sharing systems. The results derived from the framework validated in the real-world instance from the Vienna Citybike system outperform the constraint programming itself. Li et al. (2019) developed a framework with an

advanced ant colony optimization algorithm to solve green vehicle routing problems in multiple depots by maximizing revenue and minimizing costs, travelling time and carbon emission.

Simulated annealing derived from the name of the phenomenon application of the thermal annealing process of materials in metallurgy, which aims to achieve optimal materials formed by augmented crystals by heating and slow cooling. Vincent and Lin (2015) proposed simulated annealing algorithm to solve an open location routing problem in which logistics vehicles do not come back to centres after servicing customers, while their algorithm minimizes the total cost including facility operation costs, vehicle fixed costs, and travel costs. Yu et al. (2017) proposed simulated annealing algorithm with a restart strategy to solve hybrid vehicle routing problems by applying a mathematical model that minimizes the total cost of travel driven by the vehicles.

Particle swarm optimization algorithms find the optimal solution by iterating the placement of particles in the problem space evaluated in respect to the objective function in each iteration. Although analogous to genetic algorithms, particle swarm optimization does not use crossover and mutation operators. Xu et al. (2013) established the particle swarm optimization to identify optimal placement of electric vehicle charging stations with the objective to minimize the total travelled distance by EV. Zhang et al. (2018) proposed integer linear programming model with reduced variable neighbourhood search to consider the problem of collecting bicycles in need of repair by applying A hybrid discrete particle swarm optimization algorithm.

2.3 Factors influencing the demand of bicycle sharing system

In the development of the bike share system and the later stage of the built bike shares, understanding the influencing factors of rider's behaviours and decisions is a crucial investigation. In the current literature, the impact of those factors is examined from several perspectives through user surveys, correlation and regression analysis. Some factors such as the presence of the bicycle infrastructure, vicinity to public transit, proximity to bike stations from user demand and topography are considerable factors for the placement of bike-sharing stations, while other factors provide insightful meaning about user behaviours that can be used to increase operational efficiency for the bike share system.

Distance to the nearest docking station from the initial rider pick-up demand is investigated by Bachand-Marleau et al. (2012) and Gu et al. (2019) to find out the maximum acceptable distance that a rider is willing to walk. Accordingly, 500 metres radius around each sharing station is considered a

standardized operational area for initial rider demand. Design of the bicycle sharing system such as station capacity, station density and number of stations have been found to significantly affect the use of bicycle sharing systems. Without a doubt, higher station density and greater station capacity are positively correlated with bicycle usage for every station (Tran et al. 2015).

Bicycle sharing demand is significantly influenced by seasonal variation and weather conditions. Seasonal variations and weather condition includes humidity, temperature, wind and rains. Campel et al. (2016) employed a stated preference survey to determine the factors influencing rider demand under heavy weather conditions. The study revealed a strong negative correlation by heavy rain, hot and cold temperatures. Faghih-Imani et al. (2014) revealed a positive correlation that demonstrates that people are not willing to ride a bicycle under rainy and high humid conditions. Another study revealed the optimal perceived temperature for the highest ridership activity between 20 and 30 degrees Celsius as well as a negative correlation with precipitation, snow on the ground and humidity since those undesirable weather conditions restrain comfortable ride experience and cause higher risk of accident and injury (El-Assi et al. 2017).

Some have investigated the impact of topography on bike-sharing demands since the hilliness of the urban area generates certain riding patterns between pick-up and return stations. The regression analysis by Mateo-Babiano et al. (2016) revealed that hilltops stations are avoided by riders to return bicycles. Additionally, between the two stations, there are 1.9 times more trips in the downhill direction than uphill.

The presence of cycling infrastructure is another influence on the bikeshare activity due to the safety concerns of riders. Particularly, new bikeshare users are more sensitive to the lack of bicycle infrastructure than experienced users as a result of unsafe conditions caused by the mixed traffic (Fishman et al. 2015). In agreement with these findings, the impact of bike facilities (bike paths, lines etc.) is also an essential factor that encourages riders for bicycle trips (Faghih-Imani et al. 2014). More specifically, the presence of bicycle paths along the trip and less number of intersections with car traffic positively correlated with bikeshare demand considering the safety thoughts of users (El-Assi et al. 2017).

2.4 Related works on optimizing facility placement

There is now rich literature on facility location optimization from a computational point of view, utilizing various data and methods. Romero et al. (2012) combined a bi-level mathematical model

with a genetic algorithm to create an optimal design of the bike-sharing system in the city of Madrid while considering constraints that minimize the cost and maximize user coverage. Likewise, García-Palomares et al. (2012) used two location-allocation model approaches: minimizing impedance and maximizing coverage based on the spatial pattern of the potential demand. Conrow et al. (2018) conducted an analysis to provide an equitable spatial distribution based on the method constraining the data with no station more than half-mile, and no more than one mile between stations, also the accessibility to the designed bike path network. Tu et al. (2016) investigated genetic optimization of the placement of electric taxis, using big GPS trajectory data. Likewise, Zhao et al. (2020) proposed a constrained optimization model for the placement of 'taxi canteens'. They also utilized a support vector machine to identify spatiotemporal patterns of cabdrivers' dining.

3 Methodology

The optimization model and the framework of the bicycle sharing station design were developed and the flowchart of this study is presented in figure 1. Firstly, the obtained dataset was pre-processed to extract significant trips and clipped within the study area. Subsequently, an optimization model that aims to optimize the location and capacity of bicycle sharing stations was developed based on objectives and constraints. Secondly, a genetic algorithm was created to solve site selection problems. Following the sensitivity analysis of the genetic algorithm parameters, an algorithm was applied to our dataset to examine the functionality of the model. Lastly, point of interest (POI) based and population density-based optimization was performed to assess our results.

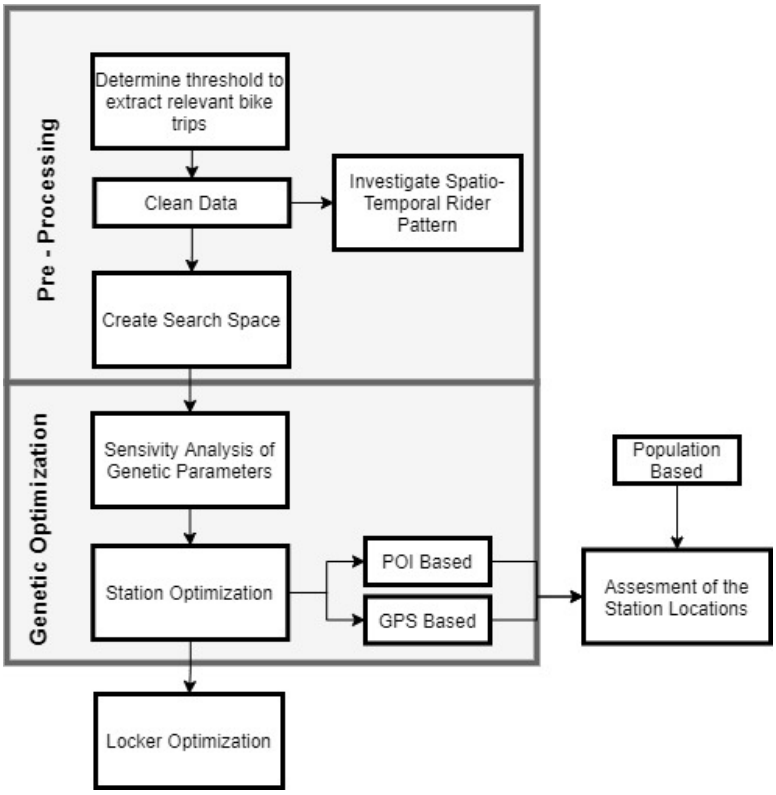

Figure 1. Framework of the methodology.

3.1 Study area and data

3.1.1 Study area

Shanghai, the study area of this research is the largest city in China with a population of approximately 26 million. It covers a land area of 6800 km² located on the central-eastern coast of China. After a year of the bicycle sharing system establishment in Shanghai, by July 2017, the scheme reached 1.5 million bicycles with 13 million registered users (Jia et al. 2019). The case study of this research is implemented within the inner road of Shanghai, which is equal to 140km² (Figure 2).

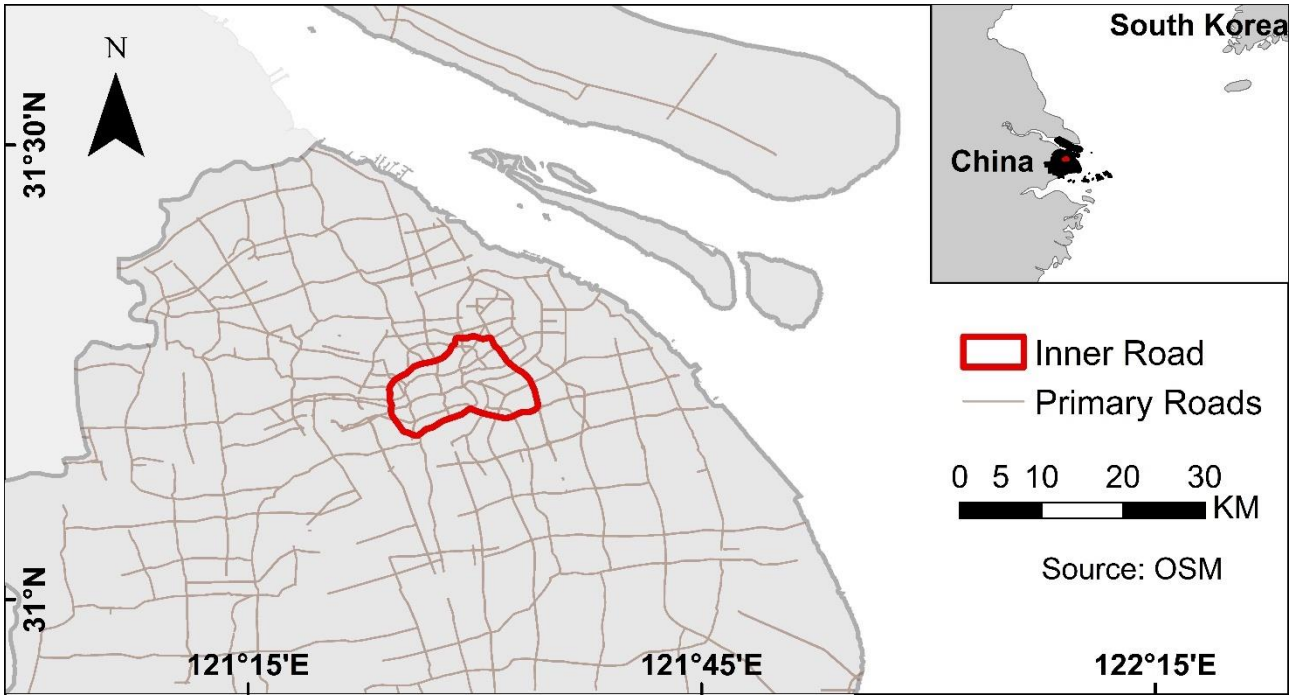

Figure 2. Overview map of Shanghai with the inset map of east China. The Red line represents the inner road of Shanghai studied in this thesis. Layers Data Source: OpenStreetMap

3.1.2 Data

In this study, the raw bicycle GPS trajectory dataset was obtained from the Shanghai bicycle sharing company (Mobike) on nearly 27 million individual trips made by around 651 thousand bicycles over a two week period from August 28, 2018 to September 6, 2018. Research of the data is conducted by Aoyong et al. (2020) to measure utilization efficiency within each sub-region by calculating time of booking (ToB) for each bike, and to explore influences of the built environment and social-demographic characteristics on bike-sharing utilization. Due to the size of the dataset, the inner road of Shanghai was selected as a case study implementation due to higher bicycle utilization rate (Aoyong et al. 2020).

This dataset includes two 'DateTime' stamps when a trip is initiated from the origin location and when the trip is completed at the destination location. The dataset also contains GCJ-02 coordinate pairs of origin and destination locations with 10 decimal places as well as the bike ID of each trip. Although this degree of precision grants sufficient proximity to the location, real-time GPS accuracy of receivers ranges between 4.9m for GPS-enabled smartphones and centimetre level accuracy for advanced receivers of high-end users (GPS Accuracy 2021).

The data is provided as 66 individual data frames, which are randomly divided. Due to the magnitude of the data frame, Google Colab working environment is created. Google Colab allows to create and run Python codes through the user's browser on the remote computers in Google servers where a Linux operating system is based. The Dataset is uploaded on Google Drive due to the convenience of the transmission of data between storage and execution servers. Therefore, the working environment is set for Python scripting and visualization of results. System specification of Google Colab includes 25.51 GB RAM, which allows users to process big data.

3.1.3 Anomaly / Outlier removal - data cleansing

The origin or destination location of demand points exceeding the study area is removed, and only points corresponding to the inner roads of Shanghai were stored. Latitude and longitude coordinates are projected into the UTM51N projection system.

Prior to calculations on raw GPS trajectory data, outliers/anomalies in the dataset should be filtered out since noisy data has a negative influence on the processing. Identification of anomalies is performed based on the benchmarking of the characteristics of the dataset such as discrepancies among timestamps and locations.

First Python scripting is created to filter out the demand points (see Appendix A). The cleaning operation is performed to remove points from certain trips with low quality. To detect those trips, the characteristics of each trip should be calculated. Therefore, durations, speeds and distances of trips are calculated based on the coordinates and timestamp of origin and destination. Certain thresholds are decided to rule out the trips with poor quality after performing an analysis on the dataset.

Trips that have 0 duration are removed prior to speed calculation since those trips calculate infinitive speed value and corrupt the statistical measurements (mean, median and standard deviation). For the detection of outlying points and to understand the extent of the significant variables, bivariate analysis of duration and distance is performed (figure 3). The fact that duration and distance are positively related, outliers can easily be observed from the graph. It is obvious that most of the points are clustered below 240 minutes and 15 kilometres.

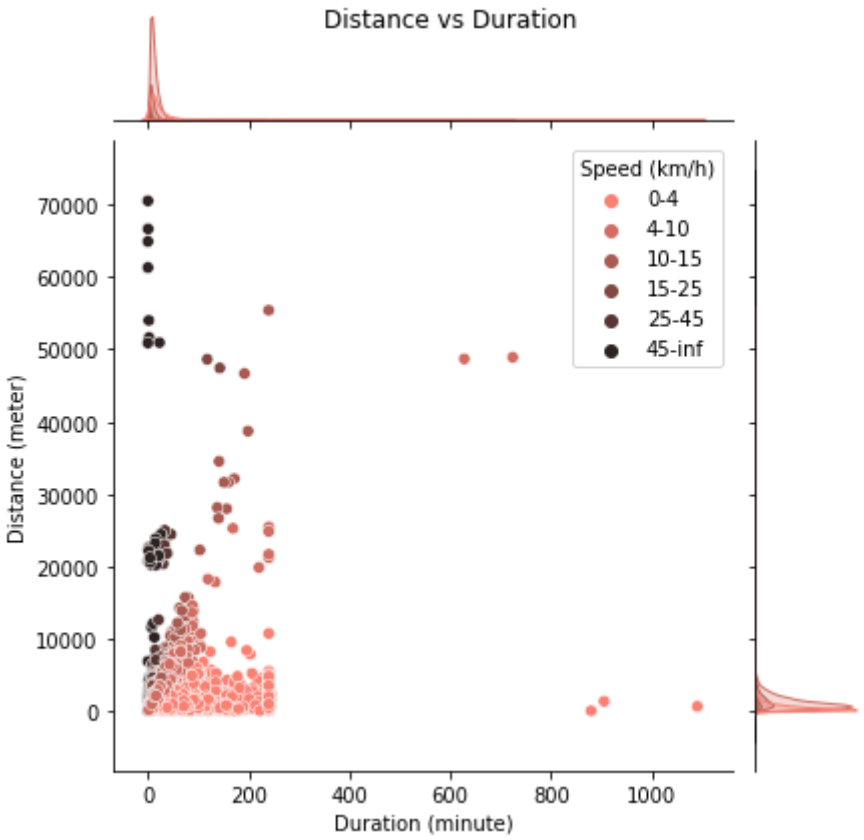

Figure 3. Scatter plot of duration and distance categorized in different shades of red by speed values. Trips were analysed between duration and distance to understand to what extent the data vary and determine the outlier trips.

Anomaly identification of trips can be divided into three parts: distance, speed and duration. Assumptions on anomalies and thresholds are as follows:

- Distance: Trips longer than a certain threshold are not significant to be used as an input for a location-allocation problem. It is assumed that those trips could be performed for leisure travelling. For example, users might want to ride a bike for a sportive or touristic activity. Those trips tend to be stopped for a long time and continue to ride. A threshold for extracting meaningful distance is set to above 200 metres and below 5000 metres. The maximum threshold is set by considering the fact that influence of trip distance on an individual's decision for bicycle commuting propensity. Therefore, since people are highly likely unwilling to commute from home to work/school for above 5 kilometres, the maximum threshold is set accordingly (Van Wee et al. 2006).
- Duration: Those anomalies occur when the user returns the bicycle in a brief time. A threshold for the minimum renting duration is set to extract trips above 0.5 minutes and below 40 minutes. Likewise, the assumptions for the distance, trips that are longer and shorter than those thresholds are associated with sportive or touristic activity. Thus, those trips should be removed since they are not origin-destination related.

- Speed: Slow trips are associated with the purpose of renting that is not performed for commuting to the destination location. For example, users might rent bicycles for touristic/leisure purposes. Therefore, a rider might stop in some location, while the renting activity is not completed. A threshold for those outliers is decided while considering the slow speed due to the possible traffic congestion and an average walking speed of an adult person (Advani and Tiwari 2006). Therefore, a threshold is set to 4km/h and 25 km/h. Trips above 40km/h cannot be associated with a cycling activity. Those fast trips could occur due to the rider might travel with a bicycle onboard of bus or other faster transportation mode.

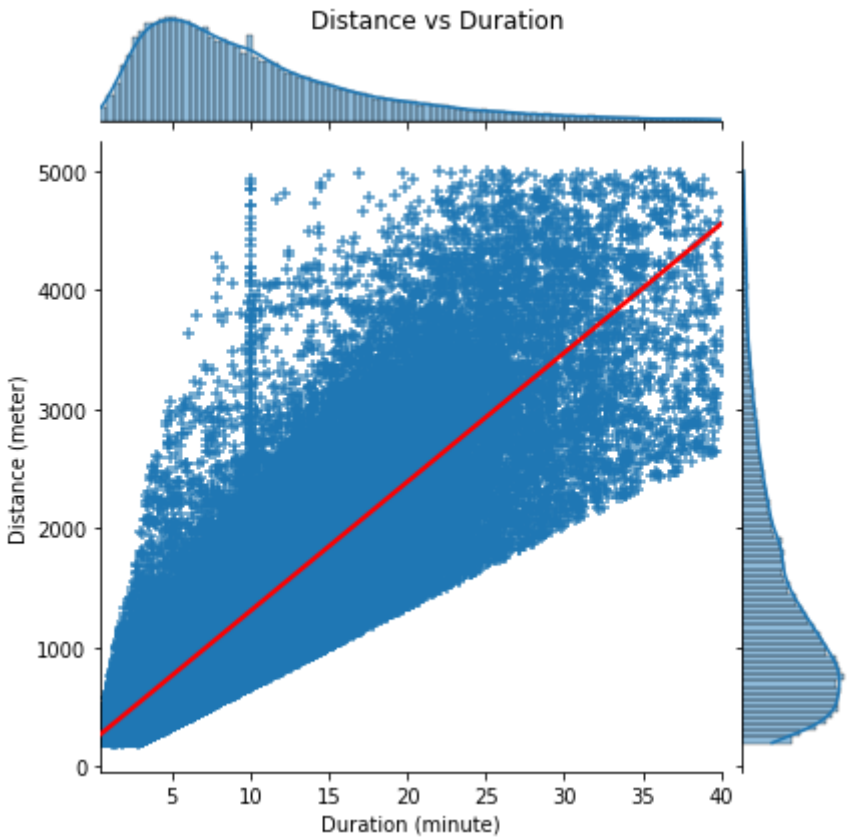

Figure 4. Scatter plot showing the relationship between duration and speed of bicycle trips with the solid red regression line. A positive relationship was achieved after removing the outliers in the data.

Figure 4 indicates the positive linear relationship between duration and distance after the data cleansing. The empty triangle on the left-bottom of the figure occurs due to the trips below 4 km/h were cleaned.

Overall, the minimum and the maximum thresholds of the distance, duration and speed were selected and applied to the dataset to remove insignificant trips. Accordingly, in the study area, over 5 million bike trips are preserved. Additionally, prior to the optimization step, trips preserved within the inner

road of Shanghai were randomly sampled to 500,000 trips. The sampling process further reduced the number of bicycles to over 150,000.

3.2 Optimization model

Church and Revelle (1974) developed a model called the maximal covering location problem to achieve maximal coverage of demand points by a given number of service points. By the light of it, the following preliminary objectives will be defined and formulated in the optimization model:

The first objective (1) maximizes the spatio-temporal demand coverage of each candidate station. Constraint (2) indicates that the total number of candidates sharing stations to be located is equal to the maximum number of stations given. Constraint (3) specifies that the Euclidean distance between the demand point and the candidate station cannot be greater than the maximum desirable walking distance. Objective (4) maximizes the number of lockers at stations concerning pick-up and drop-off demands. Constraint (5) requires that the number of lockers at the stations cannot be larger than the maximum value specified.

$$Max = \sum_{j=1}^{m} \sum_{i=1}^{m} q_j \, x_{ji} \quad \forall q \in Q \tag{1}$$

$$\sum_{p \in P} p = M \quad \forall p \in P \tag{2}$$

$$D(q_j p_i) \leq S \quad \forall q \in Q, \forall p \in P \tag{3}$$

$$Max = \sum_{t \in T} \sum_{z \in Z} \sum_{j=1}^{n} q_{jtz} \, O_{ijt} + \sum_{t \in T} \sum_{z \in Z} \sum_{j=1}^{n} a_{jtz} \, o_{ijt} \quad \forall q \in Q, \forall a \in A \tag{4}$$

$$r_i \leq c_{max} \tag{5}$$

Notations of the model

Parameters	Description
P	Set of candidates sharing stations
Q	Set of demand points
$D(q_j p_i)$	Euclidean distance between the demand point j and the candidate stations i
S	A threshold for maximum coverage radius of candidate sharing stations for demand points. In other words, the maximum desirable walking distance from demand points to the nearest station. (500m).
M	Maximum number of sharing stations
B	Set of drop-off demand points
T	A threshold for the shortest distance between stations. This threshold also maintains the maximum service distance to a demand point, meaning that no one is most likely willing to walk this distance to access the nearest station. Greater travel distance to the closest stations discourages users to walk.
c_{max}	A maximum number of bicycle lockers at each station.
r_i	Number of lockers of sharing station
x_{ji}	A binary value if demand point j is assigned to candidate stations i
o_{ijt}	If the set of demand points allocated to the candidate station i is the maximum in a period of time among other time periods.

3.3 Algorithms

This section introduces the algorithm implementations: input and output variables, and a comparison between the brute-force algorithm and genetic algorithm.

Input and output variables of the scripts and overall procedure

Inputs:

Geographic coordinates of demand points (x_j, y_j)

The total number of candidate stations (M)

A threshold for maximum coverage radius of candidate sharing stations for demand points (500m) (S)

A threshold for the shortest distance between stations. (T)

Output:

Geographic coordinates and size of candidate sharing stations. Z (x_i, y_i).

Procedure:

- Initiate the algorithm with a randomly selected candidate station (x_{i0}, y_{i0}).

  Iterate:

- Calculate the Euclidean distance from each demand point to candidate stations. Since the shortest distance between the candidate station and the demand point indicates the nearest station, allocate all demand points to the corresponding candidate. Also, consider the constraint on maximum coverage radius(S).
- Sum up the number of demand points allocated to the corresponding candidate station in order to determine the station covering the maximum station.

3.3.1 Brute-force algorithm

Since calculating precise locations of candidate stations increases the computation excessively, dividing the study area into grids will reduce the computation process. Considering the average stations' size, it is decided that the 50 metres grid size is appropriate to divide the study area. Therefore, huge computation resources and time will be saved. The placement of candidate locations and the distance calculation will be performed based on the centre of the grid. The assumption is that users are willing to walk up to a 500 metres distance to access picking up a bicycle from the nearest sharing station so that maximal coverage of the stations is constrained within a 500 metres radius. Similarly, the continued activity of the scheme is divided into time periods, and T denotes the set of periods. As all demand needs to be covered by the corresponding sharing station, the number of available lockers at the station should be maximized.

Pseudocode of the brute-force algorithm

```
p=[0]
m = maximum station
totaldistance = 0
For x:
        For y:
                For q in demandpoints:
                        distance(q, pxy)
                        If distance <= 500m:
                                totaldistance += distance
                p.append = x + y:points

h=[]
Appendmax = []
h.append = p.pop(max(p))
For m in xrange(m-1):
        for s in h:
                if distance(max(p),s) >= 800 metres:
                        Append max.append(True)
                else:
                        Append max.append(False)
        If all(a) == True
                p.append(max(p))
        else:
                p.pop(max(p))
```

The pseudocode of a simple optimization algorithm for the location of the stations is presented above. Starting from the initialization of the data frame in which the total number of demand points covered from each candidate station to demand points within the 500 metres radius is to be stored. Firstly, 2 iterations will go through all grids from left-upper corner to right-bottom corner one by one, and nested loops will store how many demand points are covered by candidate stations, with 500 metres desirable walking distance in mind. The second part of the algorithm determines the grids that have maximum coverage by considering the minimum distance among stations.

Since there are 3 nested loops in the first and 2 nested loops in the second part, the time complexity of the brute-force optimization algorithm is equal to O(n^3+ n^2), where n represents the input size. A heuristic approach to the optimization problems is essential as the number of input variables increases, the corresponding number of distances to calculate encumber the resources of the computer.

3.3.2 Genetic algorithm

Genetic algorithm in computer science is influenced by the theory of evaluation and the law of Mendelian inheritance to simulate the process of natural evaluation (Goldberg 1988). As in Darwin's concept of the survival of the fittest, the environment naturally selects the best-fit creature and others are eliminated. Genetic algorithm applies the same dynamics in computer science. Through iterations

for the creation of a better population, the one who survives last will reproduce the next generation so that characteristics of the survivor are transmitted. The fittest product results in the best solution to the problem. The effective process of genetic algorithm is greatly influenced by the parameters as well as the best operations and values of parameters for one type of genetic algorithm can be worse for another case (Vasconcelos et al. 2001).

Pseudocode of a genetic algorithm

```
Initialize m number of genes(x-y coordinates) for each individual
Initialize n number of individuals for the initial population
Repeat:
        Calculate the fitness score of each individual
        Select fittest individuals (elitism)
        Crossover among survivors (roulette wheel selection)
        Mutate individuals
        Replace population with offsprings
```

3.3.3 Population, individuals and genes

In regard to the objective, the structure of the population is defined accordingly. Since the objective is to find the optimal multi facilities, genes of the individuals are formed by the x and y pair of coordinates of the candidate station. Individual and chromosomes are used interchangeably in this study. The size of the individuals is equal to the desired number of candidate stations. On the other hand, the size of the population is the decision that should be made by considering the fact that the trade-off between computational time and the quality of the result. Each gene consists of the X-Y coordinates of the candidate location, which represents the centre of grids in M x N size of the study area.

$$\sum_{1}^{p}\sum_{1}^{i}\sum_{1}^{m*n} \mathrm{X}, Y \qquad (6)$$

At the beginning of the genetic algorithm, each zone has the same probability of being a candidate for the station placement. The equation above represents the random procedure of the initial population generation, where m and n are the numbers of rows and columns of the study area, i is the

size of the individuals, and p is the population size. The initialization step ends when the number of chromosomes has reached the size of the population defined.

$$P = \{I_1, I_2, I_3 \dots I_n\} \qquad I = \{x_1y_1, x_2y_2, x_3y_3 \dots x_ny_n\} \tag{7}$$

3.3.4 Elitism, crossover and mutation

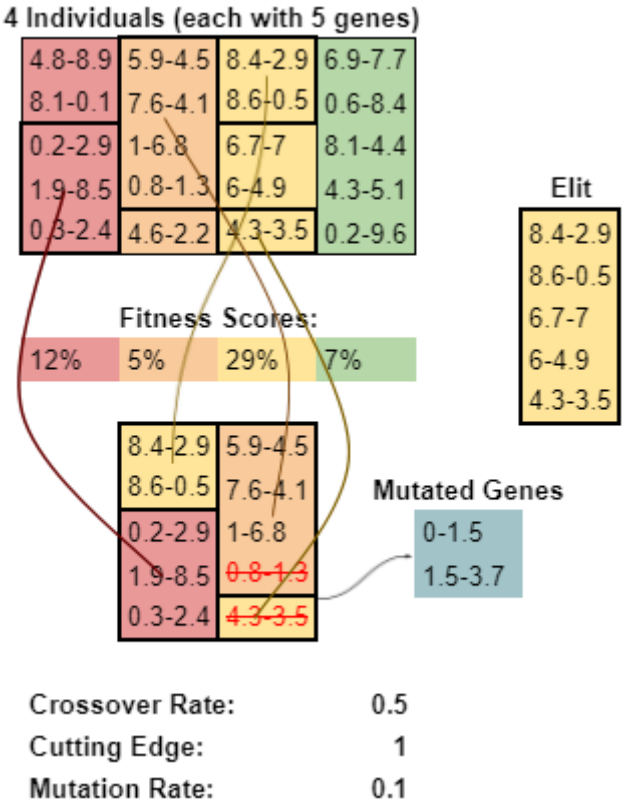

Figure 5. A simplified version of chromosomes, genes, selection, elitism rule, crossover and mutation functions in genetic optimization. The upper section shows 4 individuals formed by 5 genes with number pairs in decimal digits. The bottom table shows the selected genes in two individuals to be transferred new population. Also, mutated 2 genes (blue table), and selected elit genes (middle-right) are depicted.

Figure 5 depicts the crossover procedure, which is one of the main operations in the genetic algorithm. Population in this imaginary scenario contains 4 individuals (each with 5 genes). Genes are number pairs with one decimal digit. The fitness scores of each individual are presented below the population table ranging from 5% to 29%. The elitism rule always selects the fittest individual to be copied to the next generation. Without the elitism rule, there is always a probability to eliminate the individual with the highest fitness score from the previous generation. Since the yellow individual that has the highest fitness score is the elite of this population, the whole chromosome is transmitted to the next generation. Due to the fact that the crossover rate is 0.5, there will be two offspring elected by the

weighted roulette wheel function. The cutting edge for crossover operation is randomly selected. For example, if the chromosomes are formed by 30 genes, the algorithm randomly picks a pair of chromosomes and selects a random cutting edge in chromosomes ranging from 1 to 29. Subsequently, genes are compounded to create two offspring to be transmitted to the new generation. Those individuals having higher fitness have more chances to be elected. The weighted roulette wheel function in the imaginative case above chooses the yellow, orange, and red genes in a random manner and genes of these individuals are cut from the random edge and mate with a random set of genes from another individual. One should keep in mind that a raw crossover operation will cause premature convergence. In other words, after a while, genes in the optimum individual will be identical due to the fact that individuals will be dominated by the fittest gene resulting in a lack of genetic diversity (Pandey 2014). To avoid premature convergence, during the crossover operation, the identity check is conducted between chromosomes to be mated. Accordingly, if a gene from the parents is identical, the function keeps the first and regenerates a random gene to replace the second gene.

The mutation function modifies a random number of genes in a random individual with respect to the mutation rate and the extent. To find out the optimal procedure of the genetic algorithm, the parameters of the genetic algorithm are determined following an empirical approach to several different parameters in section 4.2. Therefore, analyses of the performance of each parameter combination will lead to the best set of values of parameters.

3.3.5 Fitness function

Likewise, natural selection, a fitness function in a genetic algorithm allows the evaluation of each individual with respect to a certain objective. The score of each individual is a key factor for the selection of the offspring in the new generation and termination of individuals of those having a low fitness score. Scoring individuals is a crucial procedure that leads to more optimal results in each generation. Subsequently, it is expected that the average fitness score and the best fitness score of the population in the new generation are greater than the old one.

End of the algorithm:

- If the maximum number of generations is set, genetic algorithm stops when this number is reached.
- If the fitness scores of individuals are close to each other. This can be observed by the ratio of the maximum fitness score and the mean of the fitness scores. If the ratio is close to 1, individuals are uniform, and the algorithm will stop.

3.4 Implementation

3.4.1 Python script

Python is a high-level, cross-platform, easy-to-read programming language that supports any kind of scientific computation, including engineering, data analytics and deep learning, with a large number of high-quality libraries supported by an immense community (Blank et al. 2020). A python script for the genetic algorithm is created (see Appendix B). Firstly, the unwanted fields are removed, since only coordinates of origin and destination of each trip for the distance calculation are needed. The extent of the study area is extracted, and the area is divided into 50m x 50m grids. Therefore, 311 grids in the x-direction and 241 grids in the y-direction are obtained. The centre of those grids is potential candidate stations. A random function in each iteration picks up a random x and y coordinates and a random cutting point for crossover. One should keep in mind that even though the grid size is set to 50 metres the rule for the minimum distance between stations was applied in the framework.

As it is explained in the previous part of this report, chromosomes are encoded as the combination of a set of genes that contains x and y coordinate pairs of the candidate station. Population size is set to 100 individuals which are randomly initiated by a random generator. The fitness function is created to evaluate each chromosome in order to decide how individuals are transmitted to the next generation. Accordingly, a fitness function is created to evaluate each individual based on their performance on coverage of demand points. Sensitivity analysis of different combinations of mutation and crossover parameters was tested to ascertain optimal numbers.

3.4.2 Lockers optimization

The size of the bicycle station can be optimized based on the temporal pattern of pick-up and drop-off demands. Sum off pick-up demands in a period of time should be satisfied by the available bicycles in the nearest candidate station. Likewise, a set of empty lockers at the destination sharing station should conform to drop-off demands in a period of time. Therefore, the second objective is set to maximize lockers at each station to serve pick-up and drop-off demands in a given period of time. Temporal demands were aggregated into 5 minutes units for the whole week. The total number of drop-off and pick-up demand points at the peak time of each station indicates the maximum number of lockers required.

The second objective is derived from the following assumptions:

- The number of available empty lockers in a candidate station i should be greater than the drop-off demand points in a time period t.
- The number of available bicycles in a candidate station i should be greater than pick-up demand points in a time period t.

The capacity of stations is calculated based on the origin and destination demands of bike GPS data. The locker optimization framework was developed based on two parameters. In the first step of the optimization, demand points falling within 500 metres radius of the candidate station are allocated to the corresponding candidate station. In the second step of the optimization, demands are grouped into 5 minutes intervals and the time periods that have the maximum and minimum demands are obtained to calculate the greatest demand estimation. Finally, in each time period origin and destination demand is subtracted to find out the state of supply of each station. Negative numbers indicate the state of being undersupplied, while positive numbers signify the state of being oversupplied. The process of locker optimization is covered in Appendix C.

4 Results

In the following section, spatial and temporal patterns of trips are summarized to understand the flow of the rider behaviour. Additionally, sensitivity analysis of the developed algorithm was carried out and optimal parameters were identified. Findings for the placement of stations and their capacities are presented. Furthermore, examination of the results was conducted by comparing findings with the results from the POI-based optimization approach.

4.1 Spatio-temporal patterns

Temporal variation of trips is presented in figure 6. per hour of the day. The pick-up demand is more than drop-off demand in the morning; drop-off demand is greater after 18 in a day. Both demands peak twice a day around 8 in the morning and around 18 in the afternoon. The distribution of trips fluctuates between 10 a.m. and 4 p.m., while the lowest number of trips occurs from midnight to 6 a.m. in the morning. Peak hours are associated with commuting purposes. Trip demands are relatively lower in the afternoon in comparison to the highest peak in the day. Drop-off demand is always greater than pickup demand in the evening. After 6 p.m. both demands are gradually decreasing and reach the lowest fluctuation pattern at night.

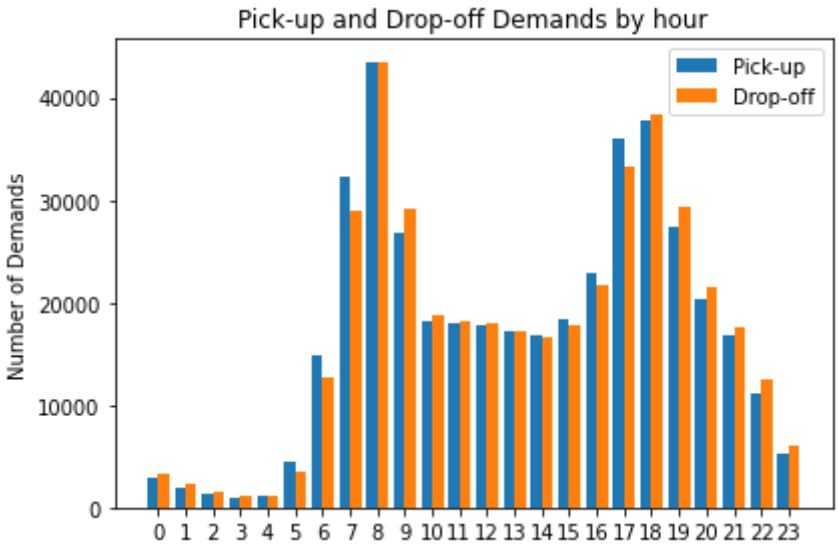

Figure 6. Temporal variation of the drop-off and pick-up demands in a day over Shanghai. Bar plot summarizes the total number of demands, which was counted per hour. Time period: August 28, 2018 to September 6, 2018.

Point clusters are identified when points are located near or close by in space to each other or when a set of points is characterized by low and high values (associated as cold spots and hot spots respectively). Kernel density estimation is a continuous presentation for an explanatory method to detect the location of clusters. Relative densities of events are calculated and presented in figure 7. Clustering analysis revealed that demands are not uniformly distributed over Shanghai. Most

demands can be observed in the north of the city and relatively less in the centre. A tendency analysis is conducted to reveal the centography of the point pattern. The mean centre with a red x mark on the graph above points out the centre of mass, while green dots indicate the median centre of the point pattern. The standard ellipse with a dashed red line points out the dispersion of the point pattern.

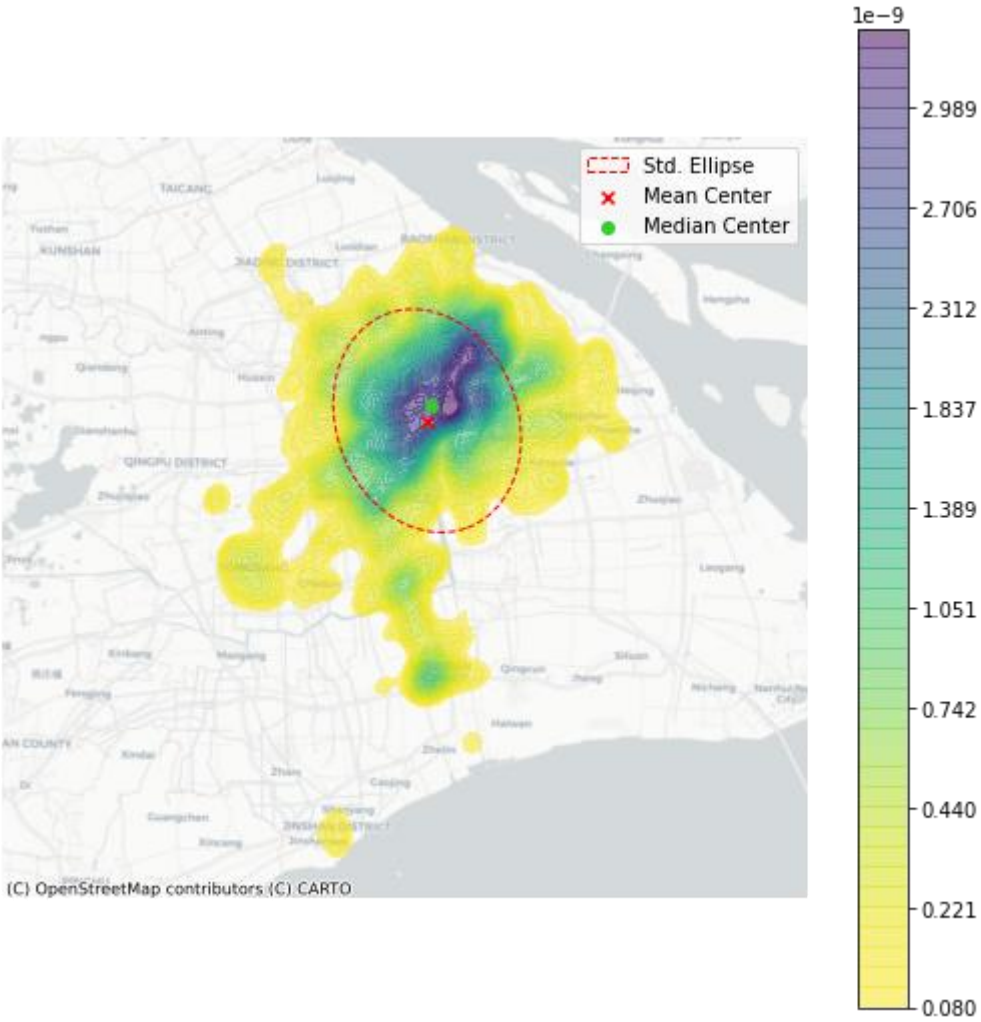

Figure 7. Kernel density estimation of pick-up demands across Shanghai derived from the period of 14 days, and standard deviation ellipse, mean centre and median centre of the trip data. Time period: August 28, 2018 to September 6, 2018

4.2 Impact of the genetic algorithm parameters

In this section, different parameters of the genetic algorithm, namely population size, crossover and mutation rate are compared and adjusted to find optimal parameters to solve the placement of bicycle sharing station problem. Several scenarios for each of these three parameters are configured to scrutinize their influence on the function of the genetic algorithm in order to determine optimal values. Throughout the sensitivity analysis of each parameter, the other remaining parameters were preserved unchanged. Also, perpetual remaining parameters were determined prior to each test in sequence: population size, crossover rate, and mutation rate. One that should be kept in mind is that since results

rely on random initiation of population and random selection of genes for crossover and mutation, another session of sensitivity analysis might indicate slightly different verifications. It should also be noted that since running the script requires substantial computation time, using minimal remaining parameters during each test executes sufficient understanding of comparisons of different parameters behaviour.

Line graphs in the following sections plot the coverage performance of the best chromosome in each generation. Similar patterns are observed in the genetic optimization process. There is a steep rise at the beginning of the algorithm, followed by a slight increase. Additionally, in most scenarios, demand coverage is saturated after the generation number 75 and reaches the maximum coverage.

4.2.1 Impact of the population size

Four different test scenarios are applied to assess the impact of the population size of the genetic algorithm (figure 8). Those four different population numbers involve 50, 100, 150, and 200, while other parameters are constant. Results from four different numbers indicate that fitness scores range from 1.9 million coverage to 2.1 million. Population size 100 derived the best result with 2,057,405 coverage. Scenario with a population size of 50 derived the lowest coverage outcome. It should be noted that a larger population requires exponential computation time, although the fitness function does not comply with similar improvements. Likewise, larger populations demand more computer memory use, which is likely to be a problem when it comes to bigger search space and larger demand points.

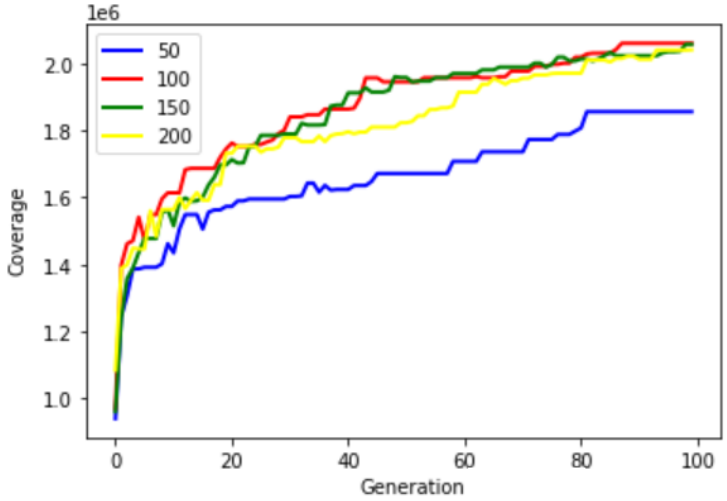

Figure 8. The impact of the population size on the fitness score. Four different scenarios were performed to determine ideal population size.

4.2.2 Impact of the crossover rate

Four different test scenarios are applied to determine the impact of the various crossover rates on the efficiency of the genetic algorithm investigated (figure 9). Those four different population numbers involve 50, 100, 150, and 200, while other parameters are constant. The coverage of scenarios fluctuates between 1,997,411 and 2,057,405. The maximum coverage is achieved with the crossover rate of 0.95%, whereas the lowest one is 0.8%. Although the best coverage is obtained with the crossover rate of 0.95%, it is observed that the crossover rate does not have a significant influence on the genetic algorithm.

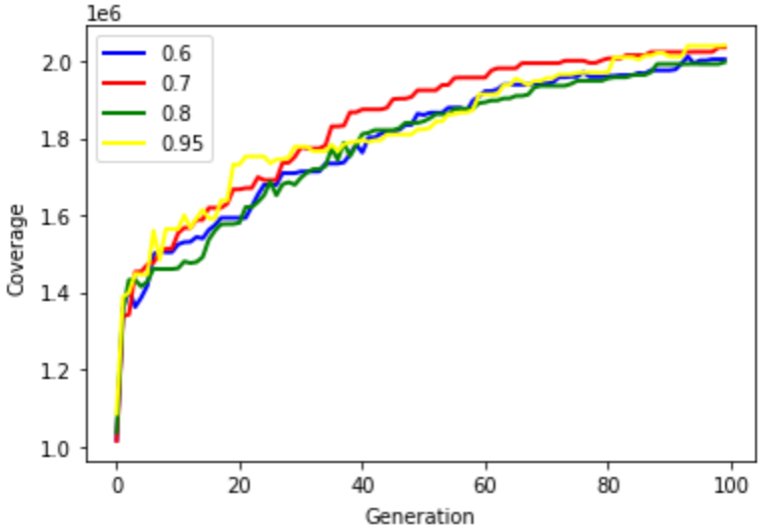

Figure 9. The impact of the crossover rate on the fitness score. Four different scenarios were performed to determine ideal crossover rate

4.2.3 Impact of the mutation parameter

Four different mutation rates are determined and applied to find out the effect on the performance of the genetic algorithm. The performance of coverage with each scenario is represented in figure 10. As the mutation rate increases, candidate stations divert towards different locations, which may cause missing of the candidate stations having higher fitness scores. The optimal solution was achieved with a mutation rate of 0.1. It is also observed that as the mutation rate increases, coverage of the stations decreases. Additionally, in the scenario with the 0.4 mutation rate, the effect of the mutation rate is apparent, since the plunges in the coverage are depicted stronger. A higher mutation rate mutates several chromosomes, such that, the next generation will be reproduced by random genes, instead of the fittest genes that come from the previous generation.

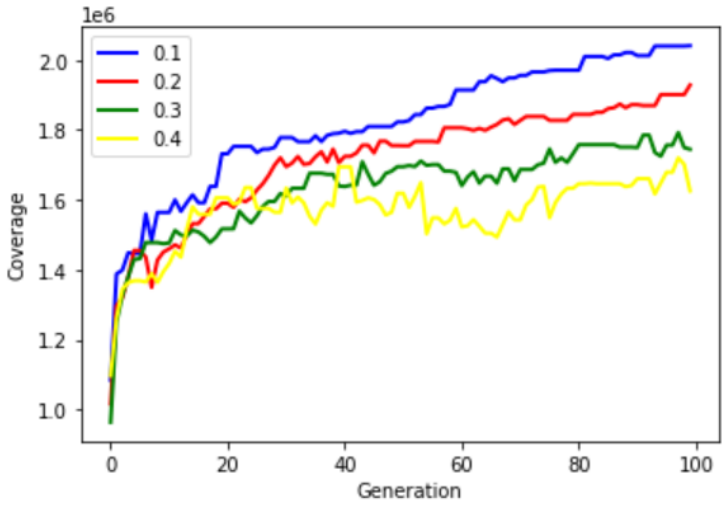

Figure 10. The impact of the mutation rate on the fitness score. Four different scenarios were performed to determine the ideal mutation rate.

4.3 Site selection of stations

By considering the trade-off between computing time and genetic parameters, results of four optimizing scenarios were obtained based on constant parameters: 100 generation number, 100 population size, 0,95 crossover rate and 0.1 mutation rate. The impact of the number of stations on the demand coverage is investigated through four scenarios by altering the station number from 25 to

100. Scenarios S1, S2, S3, and S4 are with 25, 50, 75, and 100 stations, respectively. Performances from each scenario including capacity, coverage and number of bicycles are presented in table 1.

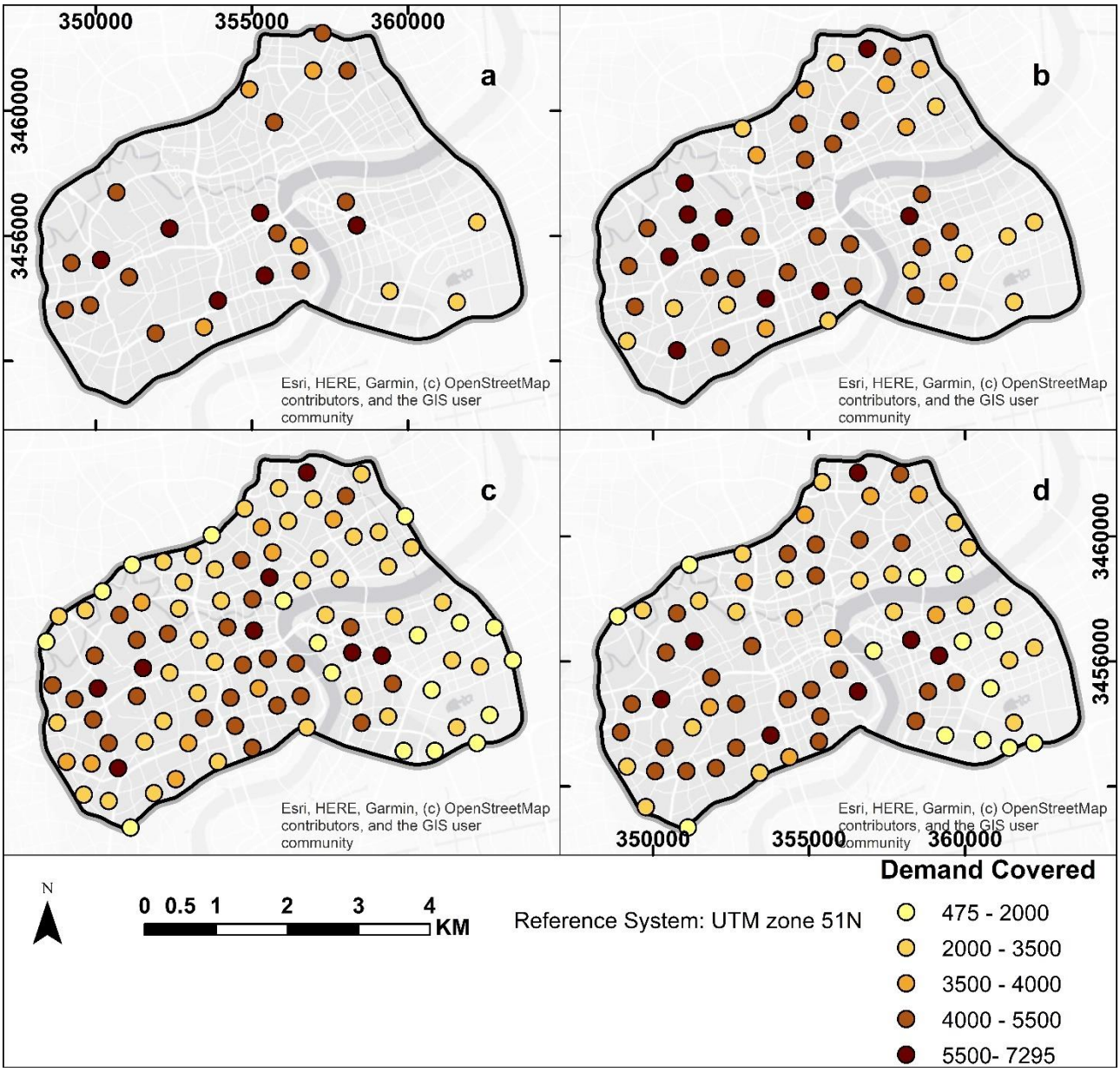

Figure 11. The optimized location of bicycle sharing stations for 25 stations (a); 50 stations (b); 75 stations (c); 100 stations (d) within the inner road of Shanghai. Selected stations are coloured according to their demand coverage. Base map source: Contextily Geo Tiles.

Figure 11 depicts the location of the candidate stations for 4 scenarios, where demand points and candidate stations are symbolized by blue and red colour schemes respectively. Spatial variation of the located candidate stations noticeably varies in different scenarios. In scenario 1 with 25 candidate stations, selected sites appear in four distinct areas. There are 5 stations located in the north of the study area, while other stations are selected in the south. As the number of stations is increasing from S1 to S4, selected sites appear to be more uniformly distributed and the distances between candidate

stations are shorter. Additionally, with the increasing station density, some candidate stations are likely to be selected in the outlying areas.

Results obtained from the proposed model are summarized in figure 11. As the number of stations increases from 25 to 100, total coverage of bicycle trips rises from 119,995 to 342,032, while the average distance from demand points to the nearest candidate station fluctuates between 328.37 m to 311.58 m. Meanwhile, since more dense station placement covered more trips, the demand coverage rate increased from 24% (S1) to 68.41% (S3).

4.4 Assessment of the location of stations

This section identifies the assessment results of the POI-based optimization in contrast to GPS-based and population density-based optimization described in chapter 4.3. Performance of the demand coverages and the benchmarking on the spatial distribution of the results are presented.

4.4.1 Point of interest-based evaluation

Over 259,000 POI are utilized to optimize the placement of bicycle stations in the inner road of Shanghai. For the assessment of our model, the station size of the POI optimization is set to 50 and genetic parameters are employed constantly in both models. It should be noted that although the traditional optimization process is performed based on the points of interest dataset, the demand coverage of the model is computed using the origin demand locations of the bike GPS dataset.

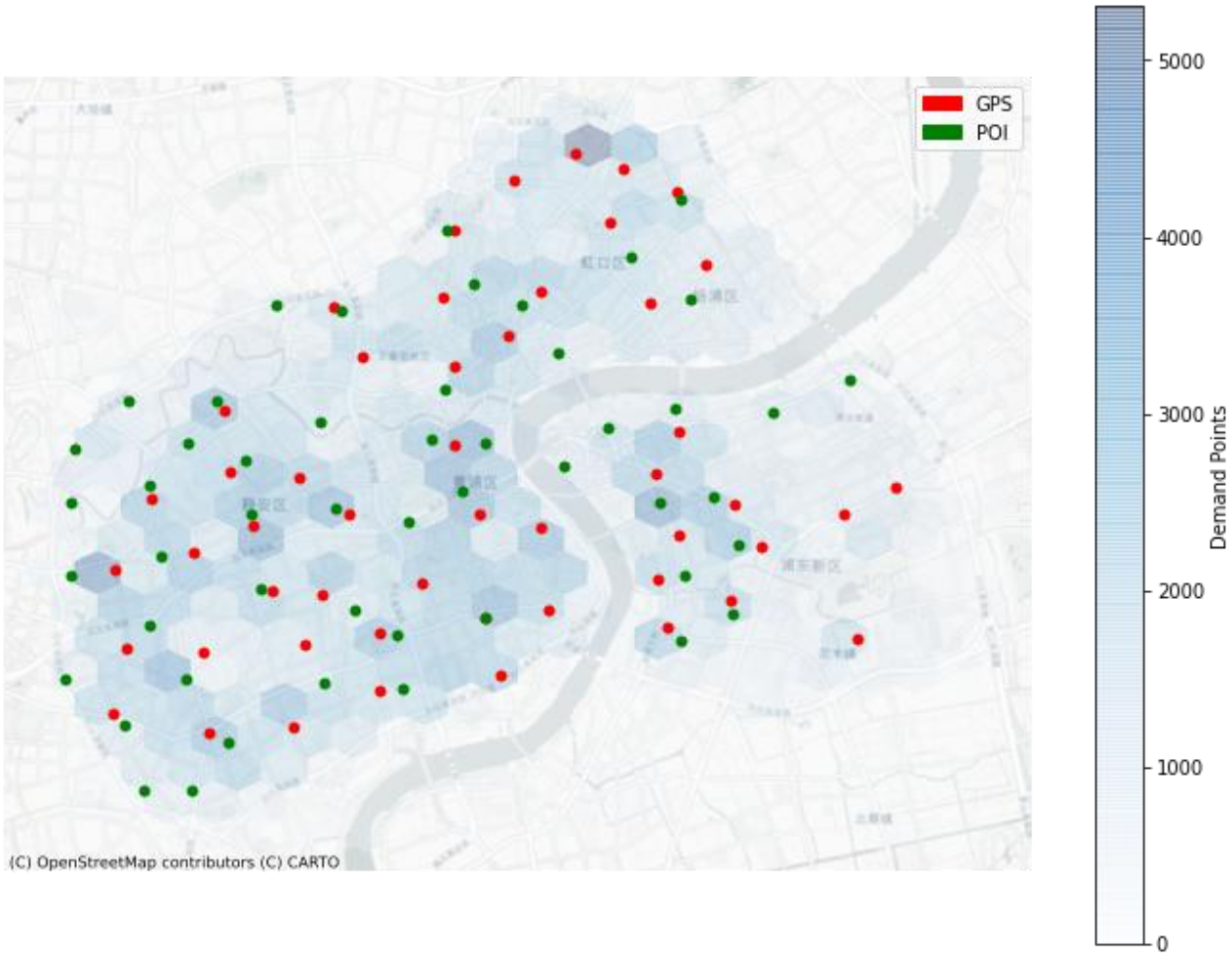

Figure 12. Candidate Stations optimized by two models within the inner road of Shanghai; GPS-based optimization model in red colour; POI-based optimization model in green colour. Demand points are aggregated into hexagons and coloured corresponding to the total number of demand points. Base map source: Contextily Geo Tiles.

The distribution of the candidate stations obtained using GPS and POI dataset is displayed in figure 13, where red and green points represent stations with the blue colour scale for demand points aggregated in grids. The spatial dispersion of candidate stations in both models substantially differs. Candidate stations derived from the POI-based optimization prominently diverge from the rider pattern of origin demand, whereas stations derived from GPS-based optimization are relatively placed in sites having high demand. Therefore, GPS-based optimization outperforms the traditional optimization approach derived from the POI dataset.

The demand coverage rate of the POI-based optimization attains 40.02%, while GPS-based optimization attains a 45.73% coverage rate (table 2). This demonstrates that GPS-based optimization increases the coverage capability to some degree, which is 5.71% equivalent to 28,550 demand points. Although differences in the spatial distribution of the stations exist to a great extent, coverage rates do not vary substantially. Nevertheless, since the inaccurately optimized placement of the bicycle

sharing system is not able to cover the existing demand of the riders, those systems are required to be rebalanced more and are subject to reap less profit compared to optimal systems.

4.4.2 Population density-based evaluation

In the population density-based optimization, the number of candidate stations is set to 50. The optimization results with 50 stations achieved from our study are used to compare the coverage efficiency with the model derived based on the population density data (figure 14). It should be noted that a comparison of models was obtained by evaluating the coverage performance on demand points. Optimal location of stations determined by population density-based model is calculated with the bruteforce algorithm. This primitive algorithm orders the population density grids in descending order and centres of the largest 50 grids are selected for the optimal location of the candidate station. Subsequently, the coverage of the candidate stations within 500 metres radius around them is calculated to compare with the results from demand-based optimization.

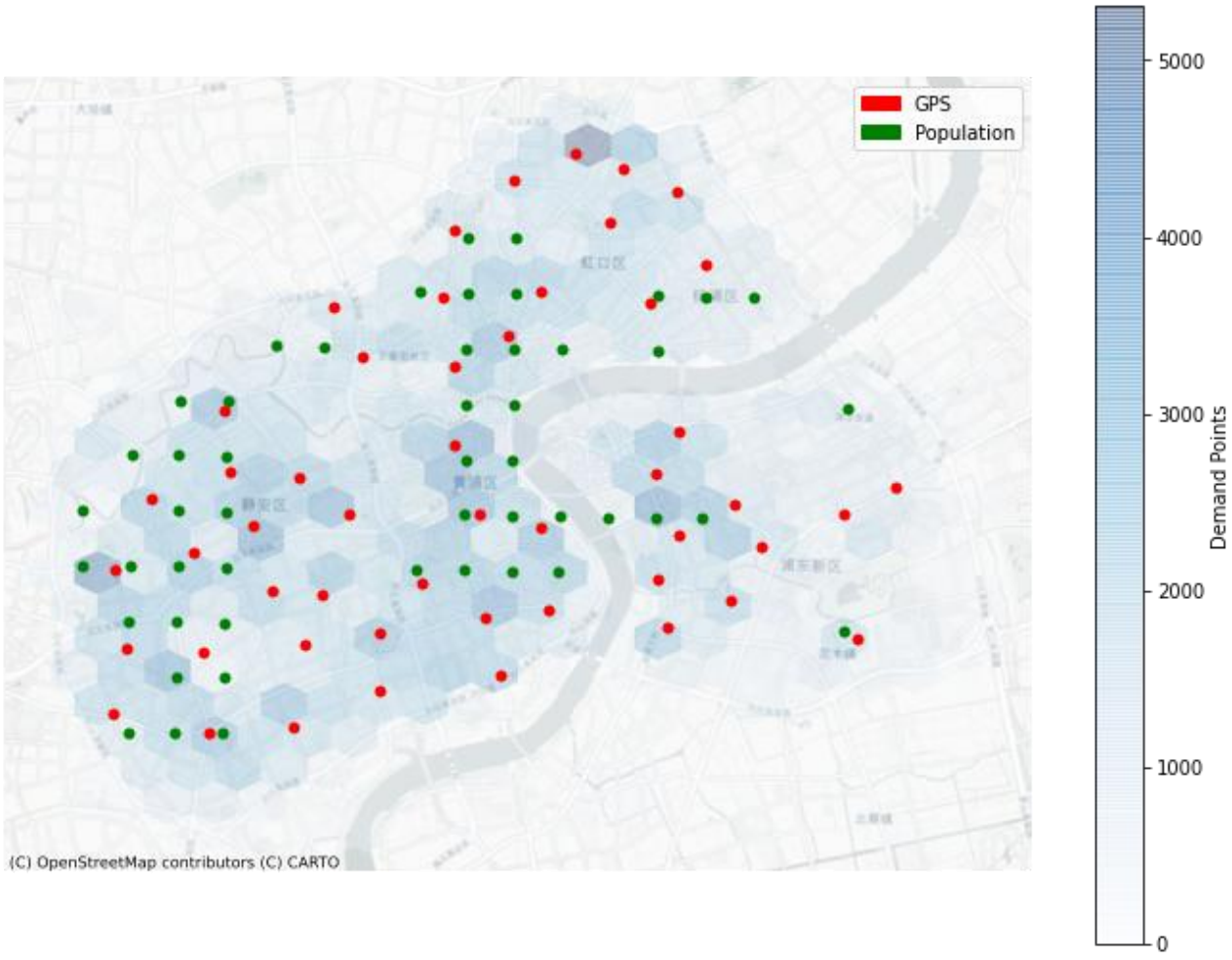

Figure 13. Candidate Stations optimized by two models; GPS-based optimization model in red colour; Population density-based optimization model in green colour. Demand points are aggregated into hexagons and coloured corresponding to the total number of demand points. Base map source: Contextily Geo Tiles.

The demand coverage rate of the population-based optimization attains 38.5%, while GPS-based optimization attains a 45.73% coverage rate (table 2). Therefore, GPS-based optimization increases the coverage capability by 7.23% equivalent to 36,150 trips.

Table 1. Demand coverage rate and total numbers of demand covered for 3 optimization approaches: GPS-based genetic optimization, POI-based and population-based optimization.

Method	Demand Coverage Rate	Demand Covered
GPS	45.73	228,627
POI	40.02	200,111
Population	38.5	192,508

Though coverage rate appears not to be significantly improved, considering the average demand coverage of the stations (4,572), the differences in the demand covered in fact indicate substantial improvement. For instance, the coverage improvement between GPS and POI approaches is equal to 28,550 trips, which in fact from a station perspective means over 6 stations, while approximately 8 stations for the population density approach.

4.5 Capacity results

Following the optimal candidate station placement, further optimization is performed on the capacity of the stations. The findings from the capacity allocation are illustrated in figure 12, which shows the results for S1 and S2 along with the location of stations marked in red circles and the number of lockers required for maximal coverage.

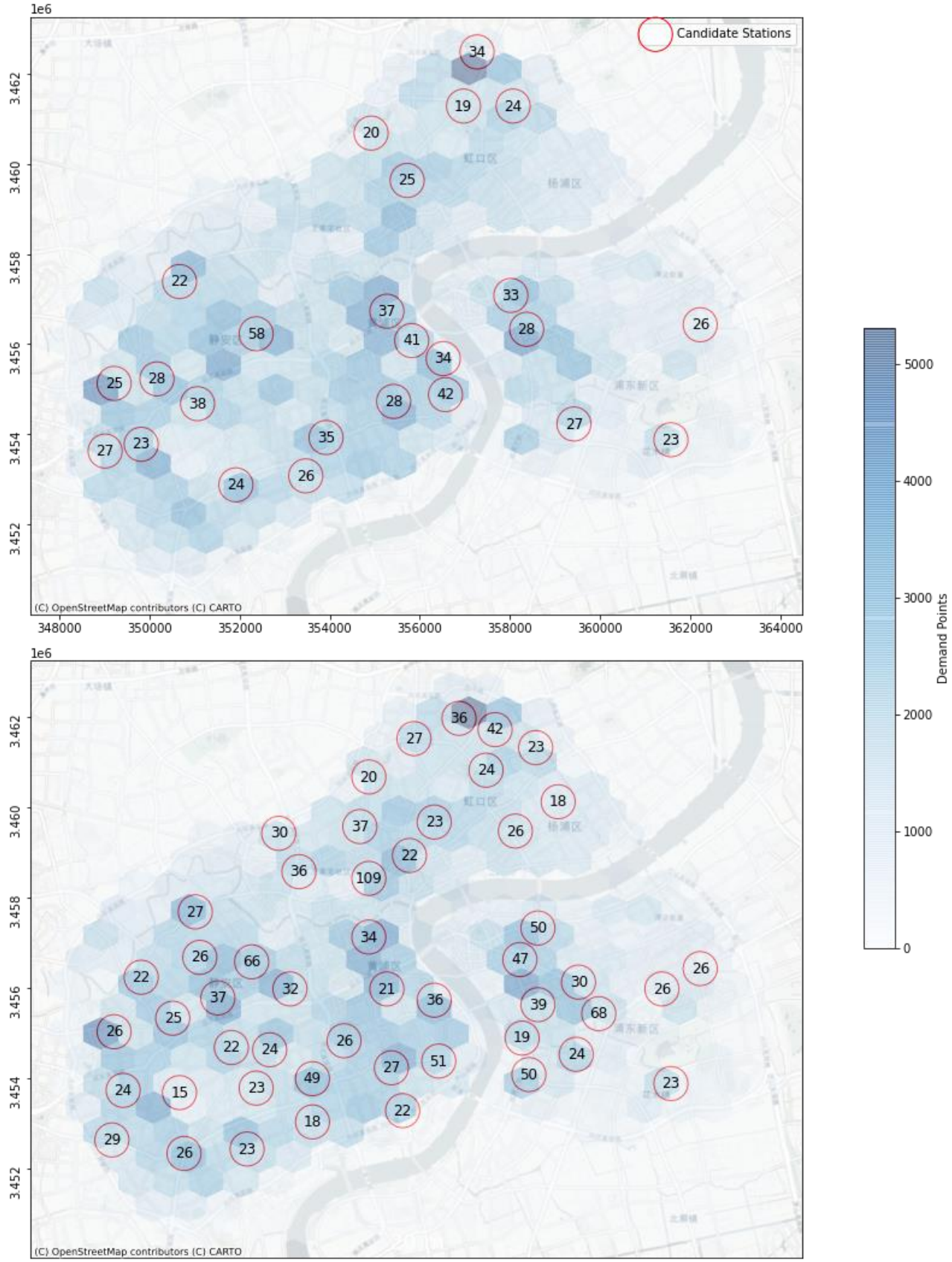

Figure 14. Station capacity results for scenario 1 (25 stations) and 2 (50 stations), while hexagons are an indication of demand density within the inner road of Shanghai. Location of selected stations and their locker sizes are depicted in red circles, Demand coverage of the stations and other scenarios can be found in figure 11. Base map source: Contextily Geo Tiles.

The capacity of stations was calculated based on the timestamp and the location of bike trips. Five minutes interval was used to group demands into a period of time and the time that reaches peak demand is considered as suitable to calculate station capacity. In S1, the station capacity of the design ranges from 19 to 58 while in S2, the station's capacity ranges from 15 to 109. Those capacities are the indicator of the desirable empty lockers or lockers with the available bicycles to pick-up depending on the situation of being oversupply and undersupply. The obtained results for four scenarios are summarized in table 1, where average demand distance from stations, total covered demand and the demand coverage rate as well as the minimum and the maximum number of lockers required to build stations.

Table 2. Average distances(m) from stations to bikes, demand coverage rate and the total demand covered as well as minimum lockers and maximum lockers required in stations for four scenarios.

Scenarios		Station Placement		Locker optimization	
	Average Distance	Demand Coverage Rate	Demand Covered	Min locker	Max Locker
25	328.37	24	119,995	19	58
50	332.38	45.73	228,627	15	109
75	311.58	54.94	274,688	9	68
100	315.69	68.41	342,032	10	116

5 Discussion

In this chapter, the results and other findings of this study are discussed as well as differences and similarities with other studies in the literature are explored. Additionally, proposals for potential future works are suggested.

The bike GPS dataset obtained in this study with over 27 million trips exceeds the size of the datasets that have been used for facility location problems in the current literature. Following the anomaly cleaning and clipping the trips into a more confined area, the dataset was reduced to over 5 million trips. Yet, the size of the dataset used in this study is still bigger than the published studies in Shenzhen (Tu et al. 2016), 53,092 trips in Changsha (Jie et al. 2017) and over 700,000 trips in Chengdu (Liu et al. 2019). On the other hand, the total area of the case study, the inner road of Shanghai, is smaller than Shenzhen (Tu et al. 2016) and greater than China Optics Valley (Yang et al. 2020). One of the limitations of the location-allocation problem is the magnitude of the search space and dataset. This issue severely affects the computation complexity of the optimization process; thus, the development of the algorithm is a crucial step to tackle such complex problems.

In this master thesis project, a big bicycle GPS trip dataset is utilized to propose an efficient bike-sharing system design, whereas several studies developed their framework on other type of datasets such as POI, cellular network, taxi trajectory data, population density. Although Liu et al. (2016) developed a bike-sharing system by considering multiple influential factors including bicycle and taxi trajectories, the bicycle trajectory dataset comprises an already built docked bicycle system. Conrow et al. (2018) proposed a covering model to site bicycle sharing stations using population density along with the consideration of equitable station distribution. On the other hand, Yang et al. (2020) utilized GPS trajectory data of a free-floating bike-sharing service to solve the spatial configuration of bike-sharing stations and their models assessed against static population density, mobile phone data and existing public bicycle system.

Although optimal placement design of the bike-sharing systems is a widely studied topic in the current literature, studies on the capacity of the stations in docked bicycle sharing systems are scarce. The maximal covering location model by Frade et al. (2015) determines the optimal design of the bicycle sharing system including locations and the capacity of the stations. However, this study differs from our framework in several aspects. Their work only gives the number of docks and number of bicycle fleets in 29 zones by taking 2,291 potential trips into account. Therefore, this thesis project differs from the current literature, the dataset used, and the search space sought is vast.

The framework developed in this master thesis was applied in Shanghai, China and four different placement configurations were tested in order to investigate their impact on spatial variation of maximal coverage model. Further analysis was performed to demonstrate the performance of capacity allocation at stations. Findings of the capacity allocation indicate the requirement of enormous areas to build stations since the average size is around 230 lockers desired to cover maximum demand at stations. Due to the fact that the large bicycle flow dynamics and the high bicycle density in the study area led to the requirements of the huge station design. Therefore, the capacity of stations can be limited to maximum lockers though the demand coverage performance will be reduced.

Above all, in the current literature, several studies utilized different aspects such as the type and the amount of dataset, implementations, optimization algorithms and frameworks to solve site allocation problems. This master thesis proposes a particular configuration of optimal bike-sharing system using a specific combination of variables from the previous studies.

5.1 Recommendations for further work

In this study, only a genetic algorithm is used to optimize the placement of bicycle sharing stations, other heuristic optimization techniques can be utilized to attempt producing quick and easy results. Therefore, results can be compared among optimization methods and the one best satisfies the objectives can be implemented further.

The optimization model is created and implemented by certain rules that comply with the inner road of the city of Shanghai. A model developed by considering common aspects of cities/regions can be used as an input for creating a multi-usable script that is applicable for different scenarios. Furthermore, a script created with a lead of this generic model will be easily applicable by planners and engineers, who are seeking an optimization solution in a short time and without an effort so that the requirement of considerable time and energy in the process of the coding by an expert is diverted to another part of the project. Such generic scripts cannot be intuitively understandable, a graphical user interface software can be proposed for improved user experience.

The cost function applied in this study can be further extended to include other cost functions such as the building cost of stations and lockers. Likewise, applying the cost function including the land price of candidate stations can even result in a more realistic and accurate design. Also, some candidate stations derived from the model require considerable space to locate since they were optimized larger than 100 lockers. The investigation regarding the locker size is not conducted in this study.

Anomaly removal on the dataset is performed based on distance, speed and duration of trips. However, spatial anomalies in some areas are not investigated, thus optimized sites might be biased by some infrequent events. As the dataset comprises a two weeks period, some areas, or venues within the inner road of Shanghai are likely to accommodate irregular multi/single day large gatherings like festivals. Such anomalies tend to manipulate existing dynamic rider patterns since a large amount of demand will invade those certain areas within a certain time period. Therefore, the optimization algorithm based on the dataset including uneven activities may yield an imprecise design of placement and capacity.

6 Conclusion

This thesis aims to develop a model to design an optimal bicycle sharing system that considers certain objectives and constraints in order to help decision-makers to establish a cost-effective docked bicycle sharing scheme. Investigation of spatio-temporal rider patterns showed that trips are clustered in the north of Shanghai, and the highest number of demands was observed during the rush hour. Based on the assessment including a comparison of the GPS-based demand coverage rate with population density-based optimization and POI-based optimization, it can be concluded that this model slightly outperforms traditional optimization approaches. As noted earlier regarding genetic parameters, sensitivity analysis demonstrated that population size and mutation rate significantly affect the optimization while the impact of crossover rate is insignificant. It should be also noted that one of the most useful outcomes of this study specifies station capacity optimization, which differs from existing studies. However, considering a limited number of studies that paid attention to station capacity optimization, assessment of the capacity results with the existing literature is not attainable.

References

Advani, M., and Tiwari, G. 2006. Bicycle as a feeder mode for bus service. *In Proceedings of the Velo Mondial conference: third global cycling planning conference* 1-8.

Bachand-Marleau, J., Lee, B. H., and El-Geneidy, A. M. 2012. Better understanding of factors influencing likelihood of using shared bicycle systems and frequency of use. *Transportation Research Record* 2314(1):66-71. DOI: 10.1007/s11116-015-9669

Bouktif, S., Fiaz, A., Ouni, A., and Serhani, M. A. 2018. Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* 11(7):1636. DOI: 10.3390/en11071636

Campbell, A. A. 2016. "Factors influencing the choice of shared bicycles and shared electric bikes in Beijing." *Transportation research part C: emerging technologies* 67:399-414. DOI: 10.1016/j.trc.2016.03.004

Church, R. L., and Davis, R. R. 1992. The fixed charge maximal covering location problem. *Papers in Regional Science* 71(3): 199–215. DOI: 10.1007/BF01434264

Conrow, L., Murray, A. T., and Fischer, H. A. 2018. An optimization approach for equitable bicycle share station siting. *Journal of Transport Geography* 69:163–170. DOI: 10.1016/j.jtrangeo.2018.04.023

Di Gaspero, L., Rendl, A., and Urli, T. 2013. A hybrid ACO+CP for balancing bicycle sharing systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7919:198–212. DOI: 10.1007/978-3-642-38516-2_16

Dong, J., Liu, C., and Lin, Z. 2014. Charging infrastructure planning for promoting battery electric vehicles: An activity-based approach using multiday travel data. *Transportation Research Part C: Emerging Technologies* 38:44–55. DOI: 10.1016/j.trc.2013.11.001

El-Assi, W., Salah Mahmoud, M., and Nurul Habib, K. 2017. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation* 44(3): 589–613. DOI: 10.1007/s11116-015-9669-z

Faghih-Imani, A., Anowar, S., Miller, E. J., and Eluru, N. 2017. Hail a cab or ride a bike? A travel time comparison of taxi and bicycle sharing systems in New York City. *Transportation Research Part A: Policy and Practice* 101:11-21. DOI: 10.1016/j.tra.2017.05.006

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., and Haq, U. 2014. How land-use and urban form impact bicycle flows: Evidence from the bicycle sharing system (BIXI) in Montreal. *Journal of Transport Geography* 41:306–314. DOI: 10.1016/j.jtrangeo.2014.01.013

Fishman, E., Washington, S., Haworth, N., and Watson, A. 2015. Factors influencing bike share membership: An analysis of Melbourne and Brisbane. *Transportation Research Part A: Policy and Practice* 71:17–30. DOI: 10.1016/j.tra.2014.10.021

Ghaheri, A., Shoar, S., Naderan, M., and Hoseini, S. S. 2015. The applications of genetic algorithms in medicine. *Oman Medical Journal* 30(6):406–416. DOI: 10.5001/omj.2015.82

García-Palomares, J. C., Gutiérrez, J., and Latorre, M. 2012. Optimizing the location of stations in bike sharing programs: A GIS approach. *Applied Geography*, *35*(1–2):235–246. DOI: 10.1016/j.apgeog.2012.07.002

Goldberg, D. E. 1989. *Genetic algorithms in search, optimization, and machine learning*. Boston: Addison-Wesley Longman Publishing.

Gu, T., Kim, I., and Currie, G. 2019. Measuring immediate impacts of a new mass transit system on an existing bike-share system in China. *Transportation research part A: policy and practice* 124:20-39. DOI: 10.1016/j.tra.2019.03.003

Hale, T. S., and Moberg, C. R. 2003. Location Science Research: A Review. *Annals of Operations Research* 123(1–4):21–35. DOI: 10.1023/A:1026110926707

Hamilton, T. L., and Wichman, C. J. 2018. Bicycle infrastructure and traffic congestion: Evidence from DC's Capital Bikeshare. *Journal of Environmental Economics and Management* 87:72–93. DOI: 10.1016/j.jeem.2017.03.007

Kennedy, J., and Eberhart, R. 1995, Particle swarm optimization, *Proceedings of ICNN'95 - International Conference on Neural Networks* 4:1942-1948. DOI: 10.1109/ICNN.1995.488968.

Jensen, P., Rouquier, J. B., Ovtracht, N., and Robardet, C. 2010. Characterizing the speed and paths of shared bicycle use in Lyon. *Transportation Research Part D: Transport and Environment* 15(8):522–524. DOI: 10.1016/j.trd.2010.07.002

Jia, Y., Ding, D., Gebel, K., Chen, L., Zhang, S., Ma, Z., and Fu, H. 2019. Effects of new dock-less bicycle sharing programs on cycling: A retrospective study in Shanghai. *BMJ Open* 9(2):1–9. DOI: 10.1136/bmjopen-2018-024280

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220(4598): 671–680. DOI: 10.1126/science.220.4598.671

Kou, Z., Wang, X., Chiu, S. F. (Anthony), and Cai, H. 2020. Quantifying greenhouse gas emissions reduction from bike share systems: a model considering real-world trips and transportation mode choice patterns. *Resources, Conservation and Recycling* 153:104534. DOI: 10.1016/j.resconrec.2019.104534

Kumar, M., Husain, M., Upreti, N., and Gupta, D. 2020. Genetic Algorithm: Review and Application. *SSRN Electronic Journal*, 2(2) 451–454. DOI: 10.2139/ssrn.3529843

Li, X., He, J., and Liu, X. 2009. Intelligent GIS for solving high-dimensional site selection problems using ant colony optimization techniques. *International Journal of Geographical Information Science* 23(4):399–416. DOI: 10.1080/13658810801918491

Li, Y., Soleimani, H., and Zohal, M. 2019. An improved ant colony optimization algorithm for the multi-depot green vehicle routing problem with multiple objectives. *Journal of Cleaner Production* 227:1161–1172. DOI: 10.1016/j.jclepro.2019.03.185

Liu, J., Li, Q., Qu, M., Chen, W., Yang, J., Xiong, H., Zhong, H., and Fu, Y. 2016. Station site optimization in bike sharing systems. *2015 IEEE International Conference on Data Mining* ICDM, 883–888. DOI: 10.1109/ICDM.2015.99

Liu, Q., Liu, J., Le, W., Guo, Z., and He, Z. 2019. Data-driven intelligent location of public charging stations for electric vehicles. *Journal of Cleaner Production* 232:531–541. DOI: 10.1016/j.jclepro.2019.05.388

Mateo-Babiano, I., Bean, R., and Corcoran, J. 2016. How does our natural and built environment affect the use of bicycle sharing? *Transportation Research A Policy Practice* 94:295–307. DOI: 10.1016/j.tra.2016.09.015.

Otero, I., Nieuwenhuijsen, M. J., and Rojas-Rueda, D. 2018. Health impacts of bike sharing systems in Europe. *Environment International* 115:387-394. DOI: 10.1016/j.envint.2018.04.014

Pandey, H. M., Ankit, C., and Deepti, M. 2014. A comparative review of approaches to prevent premature convergence in GA. Applied Soft Computing 24:1047-1077. DOI: 10.1016/j.asoc.2014.08.025

Romero, J. P., Ibeas, A., Moura, J. L., Benavente, J., and Alonso, B. 2012. A Simulation-optimization Approach to Design Efficient Systems of Bike sharing. *Procedia - Social and Behavioral Sciences* 54:646-655. DOI: 10.1016/j.sbspro.2012.09.782

Shaheen, S. A., Guzman, S., and Zhang, H. 2010. Bikesharing in Europe, the Americas, and Asia Past, Present, and Future. *Transportation Research Record* 2143(1):159-167. DOI: 10.3141/2143-20

Meddin, R., DeMaio, P., O'Brien, O., Rabello, R., Yu, C., and Seamon, J. 2021. The Meddin Bike sharing World Map. Retrieved January, 27th 2021. from http://bikesharingworldmap.com.

Tran, T. D., Ovtracht, N., and D'Arcier, B. F. 2015. Modeling bike sharing system using built environment factors. *Procedia CIRP* 30:293–298. DOI: 10.1016/j.procir.2015.02.156

Tu, W., Li, Q., Fang, Z., Shaw, S. lung, Zhou, B., and Chang, X. 2016. Optimizing the locations of electric taxi charging stations: A spatial-temporal demand coverage approach. *Transportation Research Part C: Emerging Technologies 65*(3688):172–189. DOI: 10.1016/j.trc.2015.10.004

Van Wee, B., Rietveld, P., and Meurs, H. 2006. Is average daily travel time expenditure constant? In search of explanations for an increase in average travel time. *Journal of Transport Geography* 14(2):109–122. DOI: 10.1016/j.jtrangeo.2005.06.003

Vasconcelos, J. A., Ramírez, J. A., Takahashi, R. H. C., and Saldanha, R. R. 2001. Improvements in genetic algorithms. *IEEE Transactions on Magnetics* 37(I5):3414–3417. DOI: 10.1109/20.952626

Yang, J., Dong, J., and Hu, L. 2017. A data-driven optimization-based approach for siting and sizing of electric taxi charging stations. *Transportation Research Part C: Emerging Technologies* 77(2):462–477. DOI: 10.1016/j.trc.2017.02.014

Yu, V. F., and Lin, S. Y. 2015. A simulated annealing heuristic for the open location-routing problem. *Computers and Operations Research* 62:184–196. DOI: 10.1016/j.cor.2014.10.009

Yu, V. F., Redi, A. A. N. P., Agustina, Y., and Wibowo, O. J. 2016. A Simulated Annealing Heuristic for the Hybrid Vehicle Routing Problem. *Applied Soft Computing Journal*,53:119-132. DOI: 10.1016/j.asoc.2016.12.027

Zelenkov, Y., Fedorova, E., and Chekrizov, D. 2017. Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Systems with Applications* 88:393–401. DOI: 10.1016/j.eswa.2017.07.025

Zhang, S., Xiang, G., and Huang, Z. 2018. Bike sharing Static Rebalancing by Considering the Collection of Bicycles in Need of Repair. *Journal of Advanced Transportation* 2018:18. DOI: 10.1155/2018/8086378

Zhang, W., Cao, K., Liu, S., and Huang, B. 2016. A multi-objective optimization approach for health-care facility location-allocation problems in highly developed cities such as Hong Kong. *Computers, Environment and Urban Systems* 59:220–230. DOI: 10.1016/j.compenvurbsys.2016.07.001

U.S. Space Force. 2021. GPS Accuracy: How accurate is GPS. Retrieved August, 29th 2021, from https://www.gps.gov/systems/gps/performance/accuracy.

Appendix A – Data cleaning

Python script is hosted on https://github.com/bircl/optibss. Pseudo code of the data pre-processing script:

data_process.py

```
read pickled data frames

for data frames:
        project(trips)
        df[‘distance] = df[‘destination_location’] - df[‘origisn_location’]
        df[‘duration’] = df[‘destination_time’] - df[‘origin_time’]
          df[‘speed’]= df[‘distance’] / df[‘duration’]

          delete if df[‘distance’] >= 200 and df[‘distance’] <= 500
          delete if df[‘duration’] <= 0.5 and df[‘duration’] <= 40
          delete if df[‘speed’] <= 4 and df[‘speed’] <= 25

          clip(df,innerroad)

concat(dataframes).sample(500000)
```

Appendix B – Genetic optimization

Python script is hosted on https://github.com/bircl/optibss. Pseudo code of the genetic optimization script:

genetic_optimization.py

```
search_space(boundary(df))
random_population(search_space)

for g in generation:
        fitness_score(population)
        elit(population)
        crossover(population)

mutate(population)
```

Appendix C – Locker optimization

Python script is hosted on https://github.com/bircl/optibss. Pseudo code of the locker optimization script:

locker_optimization.py

```
For station in stations:

        allocate trips to stations within 500m radius
        group trips into 5minutes intervals

        for group in 5 minutes intervals:
                calculate incoming and outgoing bikes
                abs(incoming-outgoing)


       max(abs(incoming-outgoing))
```