Traffic Analysis, Prediction and Optimization using Machine Learning Algorithms

Pedro Lopes
Instituto Superior Técnico
Guarda, Portugal
pedromtlopes@tecnico.ulisboa.pt

***Abstract*—A balanced public transport system is essential for environmental, social and economic sustainability and is a prerequisite for ensuring the quality of life of a city's inhabitants. In such manner, it is crucial to compose a traffic network capable of match passengers' expectations. However different and conflicting interests have to be addressed. Passengers want to travel as fast as possible, waiting as little as necessary and making as few transfers as possible. On the contrary, operators seek to minimize costs in order to achieve higher profits.**

Determining customised routes and their frequency adjusted to demand can highly affect the costs and efficiency of the urban transportation system. A change in transit frequency can influence not only route capacity but also waiting time. Variation in waiting time will affect the passengers' route choice strategy.

In this thesis, we extend previous work of other authors, that were focused on origin, destination and interchange inference and transit network design. In this context, our work tackles simultaneously the two first stages of the overall operational planning process for public transportation networks, network design and frequency setting.

Moreover, it will be studied different metrics regarding public transportation activity, especially waiting and travel time, transfers number, but also bus empty seats that constitute an unproductive cost to transit companies.

It will be considered transit data from the city of Lisbon to assess the efficacy of the computed methods.

***Index Terms*—Transit Network Design, Frequency Setting, Genetic Algorithm**

I. Introduction

In the 21st century, urban mobility has become one of the key issues to metropolitan areas. The continuous rural exodus and the increased preoccupation about the environmental issues has conducted many different governmental policies, mainly focused on the sustainable growth of urban areas.

It is safe to say that public transportation is considered a crucial backbone of it, due to the fact it constitutes a reliable mean of transportation at an affordable fare in high density population zones, whilst emitting low carbon emissions.

Public transportation operators face a wide range of challenges related to its transit operation, such as unpredictability of events or extreme weather conditions. At the same time, these companies often struggle to strike the right balance between its level of service at a competitive cost and the user's expectations, who greedily demand reduced waiting times at stations, accuracy in bus or train arriving time, but also a modern fleet with sufficient capacity.

Reducing users' costs usually leads to increase in operator costs, and vice versa. However, determining the most efficient operation for both users and operators, for given levels of service and cost requirements, is essential for the success of such systems. [1]

Frequency setting is one of the the main activities in the tactical planning of public transport operations. Allocating frequencies of bus services in a city network is a multi-criteria problem that typically considers the operational costs, the passenger demand coverage and the service reliability. [2]

This thesis will address Transit Network Design (TND) Problem and Frequency Setting (FS) Problem together. It will be applied Automatic Data Collection System (ADCS) data divided by different parts of the day during a week in order to assess the appropriateness of the given route set.

The objective function contains several metrics regarding passenger and operator costs. The latter includes empty seats. As far as passenger costs is concerned, waiting and in-vehicle time, transfers and unsatisfied demand will be measured.

The data sources treated in this thesis, were kindly provided by Carris and Metropolitano de Lisboa, both transit operators running in Lisbon area.

II. Background

A. Transit Network Design

The aim of the transit route network design is to determine a route configuration that achieves a certain desired objective, subject to the constraints previously highlighted. As noted by [3], network design elements are part of the overall operational planning process for public transportation networks. The process includes four steps: design of routes, setting frequencies, developing timetables and scheduling buses and drivers

The first stage is optimization of bus routes based on the demand matrix. The next stage involves proper determination of bus frequencies on each route with respect to the demand matrix. Scheduling optimal fleet to the routes based on the predetermined timetables in second stage, budget limits and location of the depots will be considered in this stage. In the final stage, fleet crew and their roster table will be assigned. [4]

B. Frequency Setting

The Frequency Setting problem determines the number of trips for a given route to provide the best level of service as

possible in a planning period. Including service level, cost and fleet size restrictions. Since frequency changes affect directly route waiting and travel time, so this parameter impacts the passengers' perception of the level of service which may lead to an increment or decrement of the system usage.

Bus transport planners have to consider the bus network demand, fleet capacity and route characteristics in determining bus service frequencies. At the same time, the bus operators have to provide the bus service operation and management based on the cost and benefit analysis of their investments.

The time between two vehicles offering a service, which conveys exactly the same information as frequency, is known as the headway. The service headway is the inverse of the frequency.

$$Headway = \frac{1}{Frequency} \tag{1}$$

The passenger's arrival follows the uniform distribution, so the average waiting time is assumed to be equal to half of the headway of a route.

$$WaitingTime = \frac{Headway}{2} \tag{2}$$

We define a trip as a set of intermediate journeys, depending whether there are interchanges thorough trip progression. In the case of multiple routing options, the waiting time is estimated by the average of the sum of the half of the headways of all possible routes.

$$\text{Journey Waiting Time} = \frac{\frac{\text{Alternative 1 Headway}}{2} + ... + \frac{\text{Alternative n Headway}}{2}}{n} \tag{3}$$

For the trips that include transfers, the waiting time is calculated as the sum of all the intermediate journeys headways that concern the trip.

$$\text{Trip Waiting Time} = \sum \text{Journey Waiting Time} \tag{4}$$

Frequency on each route should lie within a range. For instance, high frequency on a route increases crowding or congestion in the route [5]. On the other side, low frequency leads to rise in waiting time and travel time [6].

C. Transit Route Network Design with Frequency Setting

The Transit Network Design and Frequency Setting Problem (TNDFSP) is a NP-hard, multi-constrained combinatorial optimization problem with a vast search space for which the evaluation of candidate solutions route sets can be both challenging and time consuming, with many potential solutions rejected due to infeasibility. [7]

The TNDFS Problem addresses the conflicting interests of passengers and transit agency. The key objectives of the problem are to minimize the passenger time and the operating cost. For improving the service level, it is essential to reduce the in-vehicle travel time by providing a shortest path between each pair of stops, to reduce the waiting time by increasing the frequency of the vehicles and mitigate the number of transfers by considering more direct trips for passengers. For operators, it is economical to reduce the fleet cost and carbon tax by restraining the fleet size and imposing a penalty for emissions, respectively. There are two objective functions in this problem: the first objective function represents the passengers time, and the second objective function comprises of the operating cost. Passenger time in the first objective varies with respect to multiple parameters of waiting time and transfers [8].

In the first phase, the transit network design (TNDP) part consists of finding a set of transit routes that together forms an efficient transit network, serving users' travel demand represented by an origin–destination matrix with minimum total travel time and transfers, which include penalties for transfers and trips which contain more than 2 transfers (unsatisfied demand). Transit routes take into consideration existing transportation network available and predefined stop points. During the second phase, assigning frequencies (FSP) to transit routes establishes waiting times and total capacity levels for routes, and makes possible the calculation of the total necessary fleet for network operation and also vehicles utilization.

D. Public Transportation Planning

The planning problem of designing public transport networks is highly complex. The main goal of most transit agencies is to offer to the population a service of good quality that allows passengers to travel easily at a reasonable fare. However, they are subject to budgetary and management restrictions (number of buses or drivers, as well as bus capacity) that make difficult meet costumers' expectations.

By saying that, this organisation procedure treats a wide range of settlements, from longstanding ones to real-time choices. There are four distinct stages: strategic, tactical, operational and real-time.

At the **strategic** level, transit planning is concerned not only with the design of transit routes and networks, but also with types of vehicle and stop spacing. This involves long-term decisions focused on designing a network of routes to meet passenger demand.

Operational-level planning aims at constructing vehicle and crew schedules that minimize total costs. This includes to vehicle scheduling, driver rostering, maintenance planning, as well as parking and dispatching. [9]

Tactical planning is typically performed on an annual or biannual basis. The aim is continually to address the tradeoff between service quality and system costs, e.g., by reflecting changes in demand or the available budget. [10]

Even when a solution for the planning process is given, operation of transport systems is affected by uncertainty of travel times. Extreme weather, accidents or passenger demand at bus stops may cause such constraints. To address these situations, **real-time control** strategies are implemented to guarantee an efficient service during the operation of the system. For instance, that may be holding vehicles in a station or even skipping specific ones. [11]

E. Genetic Algorithm

Genetic algorithm (GA) is a search heuristic that mimics the process of natural selection and genetics. It belongs to a group of evolutionary algorithms. This procedure combines a Darwinian survival-of-the-fittest with a randomized, yet structured information exchange among a population of artificial chromosomes. The main attractions of the GA approach involve simplicity of procedures, global perspective and inherent parallelisation. [12]

GA begins with an initial population of strings (chromosomes), created randomly. Each chromosome is formed by a set of variables (genes), encoded in binary values. In fact, each string represents all the problem variables to the solution context according user criteria. The creation of strings in the initial population of GA is as simple as tossing an unbiased coin. The successive coin flips (head=l, tail=0) can be used to decide genes (bits) in a string. [13]

After initial population of possible solutions obtained, it is necessary to determine how fit a gene is. This method, named fitness function, is the most important feature of GA, because it provides comparable scores among chromosomes.

In general, a fitness function F(x) is first derived from the objective function and used in successive genetic operators. For maximization problems, the fitness function can be considered to be the same as the objective function i.e.,

$$F(x) = f(x) \tag{5}$$

However, for minimization problems, the goal is to find a solution having the minimum objective function value. Thus, the fitness can be calculated as the reciprocal of the objective function value so that solutions with smaller objective function value get larger fitness. [12] [13]

$$F(x) = \frac{1}{[1 + f(x)]} \tag{6}$$

As stated above, GA simulates procedures that actually happen in nature. Improvements, from one generation to the next, are achieved using a set of three operators, **reproduction/selection**, **crossover** and **mutation**. Each one plays a crucial role and serves a well-defined objective.

Reproduction/selection operator is the first applied on a population. Basically, chooses the best chromosomes, in other words, the ones with higher score in fitness function, and copies them to the upcoming generation. This is how we guarantee to strings (individuals) with high scores to be selected more often than those with low scores, which may not be selected at all. [14]

The **crossover** operator is the most significant phase in GA, it combines two chromosomes to originate new offspring (children) to the next generation. Crossover is performed by randomly choose a bit which separates head and tail of the string. Eventually, the tail of the two parents are swapped to get new children.

In certain new offspring formed, some of their genes can be subjected to a **mutation** with a low random probability. This implies that some of the bits in the chromosome can be flipped. Mutation operator occurs to maintain diversity and avoid premature convergence.

In the end, after the completion of the previous procedures, there is an evaluation of the new chromosomes. If the termination criteria are not met, the population is again operated by above three operators and evaluated.

F. Data Formats and Tools

This thesis will be developed using a wide range of data structures, such as GTFS (General Transit Feed Specification), Automatic Fare Collection (AFC), Automatic Passenger Count (APC), and Automatic Vehicle Location (AVL).

OpenStreetMap (OSM) will be used to design the optimized network, analyse traffic aspects, compare current and shortest paths between stations on the same route and compute travel durations.

1) GTFS: General Transit Feed Specification (GTFS) defines a common format used for transit operators to publish public transportation schedules and related geographic information. This format was created in 2005 by a Google Engineer, Chris Harrelson, with the aim of incorporating transit data into Google Maps.

GTFS comprises two different components:

- **Static**, that contains schedule, fare and geographic information
- **Real-time**, that presents predictions, service alerts and vehicle positions

A GTFS dataset is described by a collection of at least 6, and up to 13 CSV (comma separated values) files in .txt format. Files agency.txt, calendar.txt, routes.txt, stops.txt, stop_times.txt and trips.txt are considered mandatory to a valid GTFS data feed [15].

The main advantage of using GTFS is to access the detailed schedule (stop_times.txt) of each trip ID. [16]

Today, many operators publish publicly their routes and schedules in GTFS format.

2) Automatic Data Collection: Operators have three sources of automatically collected data obtainable:

The **AFC (Automatic Fare Collection)** system includes several functionalities to monitor and control operations concerning issuing, sale and validation of transport tickets. In Lisbon, AFC system is denominated VIVA and integrates all the transit operators in Lisbon area.

This system records the fare related information whenever a passenger pays for a trip. It can be at a ticket vending machine, in a fare gate or, in case of Carris, in a validator on-vehicle. Its information comprehends card identifier and fare category (standard, student, senior), timestamp and location.

In the matter of AFC systems, they can be classified as open, hybrid or closed. While the Carris bus network is considered an open system, as the passenger only needs to tap in, in the case of Metro network, it is mandatory to tap at the entrance and exit of the stations. In respect of a hybrid example, in Lisbon area also, which has station from both categories,

there is Comboios de Portugal, state-owned company which operates passenger trains.

Typically, open systems have been the main research interest in the development of a traveler's trip chain because the closed system provides both origin and destination (O-D) information of a trip. [16]

The **APC (Automatic Passenger Count)** includes technology devices, such as sensors or CCTV cameras which accurately log boardings and alightings. Each record indicates timestamp, card ID and trip information.

AVL (Automatic Vehicle Location) system periodically indicates real-time vehicle localisation. This is used to inform passengers bus arrival time estimation. AVL data notifies about date, time, speed, trip information and of course, GPS coordintates.

3) OpenStreetMap: OpenStreetMap is an open collaborative project to create a free editable map of the world. Volunteers gather location data using GPS, local knowledge, and other free sources of information and upload it.

G. Data Sources

1) Carris: Carris is the main public transportation operator at the surface directed by Lisbon's City Hall. Its fleet is comprised by 706 buses from different sizes and fuel types, 48 trams, 3 funiculars and even an elevator (Santa Justa lift). As far as human resources are concerned, Carris employs nearly 2450 collaborators.

During the previous year (2019), a total of 139.5 million of passengers travelled with Carris. [17]

2) Metro: Metropolitano de Lisboa is the public transit operator responsible for the metro system of Lisbon. The system includes 56 stations and 4 lines (Green, Yellow, Red and Blue) summing up 44.5km of extension. Regarding its capacity and dimension, Metropolitano de Lisboa assumes 1452 employees and 333 train carriages.

Throughout 2019, 173 million of passengers used Lisbon's underground system. [18]

III. ODX Inference

A. Origin Inference

Origins are inferred by matching fare transactions to locations through transaction time stamps and card-reader identifiers. When a card reader is installed at a fixed location, such as a fare gate in a station, the reader's location is simply assigned to the transaction. When the reader is installed in a vehicle, the transaction time stamp is compared with vehicle location data to determine boarding stops. [19]

B. Destination Inference

Following the assumptions presented on Chapter 2 concerning bus alighting times and locations, spatial information about the passenger's possible alighting location and subsequent origin location is needed. By saying that, the distance between the passenger's next boarding location, referred to as target-location, and each stop served by the current vehicle trip need to be calculated, in order to determine which stop is closest.

In the event of a transaction of a bus boarding, the stop and route field must be valid and cannot be on the last stop of the route . Otherwise, the origin is assumed unknown and the destination cannot be inferred.

If there are no other records in the card's daily history, the closest-stop rule cannot be applied. On the opposite, if the transaction is the last transaction of the day, the daily symmetry rule is applied, and the target location is defined as the first origin of the day. Otherwise, the target location is the next stage's origin.

If the distance between the alighting stop candidate and the next tap location is greater than a pre-established maximum interchange distance, the destination can´t be inferred, because is assumed that the passenger will not walk a significant distance.

C. Bus Alighting Time Estimation

[20] exploits the GTFS file *stop times.txt*, which has, for every trip, the scheduled arrival and departure times for every serviced stop. In the case of the bus GTFS, the departure time and arrival time have the same value in every entry. By calculating the difference between the departure time for two consecutive stops, the result is the time the bus takes between those two stops. So, by adding up each consecutive stops plus a bus stop time per stop, it is possible to deduce route or trip duration.

D. Destination Inference Results

The percentage of destinations deduced is less than the half (47.70%). The reasons rely on a significant percentage of cards (40.28%) that are tapped only one time that day, which does not enable us to consider closest stop rule. Furthermore, a substantial distance between the alighting stop candidate and the next tap location (12.02%) does not allow us to infer destinations. It is assumed that passenger travelled by other means of transport.

As it would be expected, this result is quite similar to [20], where 45.24% of the destinations were inferred.

E. Bus Alighting Stop Distribution

The destination inference process computed by [20] was capable of deducing 47.70% of bus alighting times and locations. This is quite inconclusive, as the majority of the journeys do not have a destination.

In this thesis, one metric of our work is empty seats in a bus during a route, which represents the difference between entrances and alightings in a route stop. In pursuance of the best optimization solution as possible, it is essential to have the idyllic overall estimate of the demand.

By stating that, in order to overcome this lack of destinations inferred, we do propose to assume an uniform distribution to the trips that do not have destinations deduced. This means that the number of passengers that enter in a specific stop, which do not have a destination, are distributed equally to the stops ahead of that route.

IV. Transit Network Design with Frequency Setting

As demand varies throughout the day, operator resources vary as well. In the same manner that during peak hours there is a heavy pressure on transit activity, during the night period this tends to become residual.

In our case, day will be divided in 4 parts, which means 4 OD matrices that acknowledge demand alterations. The time interval studied is from 7 (Monday) to 11 (Friday), October 2021. Each limit value corresponds to transaction timestamp, that is bus entry time:

- 6:30 a.m to 10:59 a.m, which will be named first matrix or morning matrix
- 11 a.m to 3:29 p.m, that will be titled as second matrix or lunch matrix
- 3:30 p.m to 7:59 p.m, denominated as third matrix or afternoon matrix
- 8 p.m to 0:30 a.m, designated as fourth matrix or night matrix

In this chapter, we begin to present some preprocessing steps in the OD matrix as well as problem representation and its difficulties. Then, some procedures about frequency assignment and adjustment. In the end, it will be pointed up details regarding Genetic Algorithm.

A. Inputs

1) Input Matrices: Each matrix divides the day, so it becomes easier to match resources according to demand.

Matrix	Number of Validations	Percentage
6:30 a.m - 10:59 a.m	158064	**32%**
11 a.m - 3.29 p.m	131580	**27%**
3:30 p.m - 7:59 p.m	165439	**33%**
8 p.m - 0:30 a.m	39079	**8%**
Total	**494162**	**100%**

Empirically, it is demonstrated that during peak hours there are way more passengers compared to night hours. This divergence on the values has to be considered regarding frequency assignment.

B. Initial route set

In the first step of the genetic optimization process, it is necessary to produce an initial solution (route set).

Firstly, one single route set of R (input value) routes is obtained. Each of the R routes is a shortest path (based on travel time) between a selected pair of stops. Using a greedy algorithm, it is selected the R pairs (i, j) that have the highest number of passengers that travel along the shortest path between stops i, j. We will address this value as ds_{ij}.

$$ds_{ij} = \sum_{m \in S} \sum_{n \in S} d_{mn} \tag{7}$$

where S is the set of stops that are in the shortest path between i and j. d_{mn} is the number of entries for pair (m, n) in the OD matrix.

The greedy algorithm is summarised as:

1. Input the value of number of routes, R.
2. Initialise matrix DS, where $DS = \{ds_{ij} \mid i, j \epsilon [0, 1, 2, \ldots, |N - 1|]\}$
3. Find the pair (i, j) in DS with the highest ds_{ij} value.
4. i and j become the terminals of a new route.
5. Add every node in the shortest path between i and j to the route.
6. Remove every node pair (m, n) that are satisfied by the newly added route from DS
7. Check if the number of routes reaches N, stop. Otherwise go to Step 3

Secondly, an initial headway is assigned to each route created, equal across matrices. This value indicates the time between consecutive services in a route. In section 5.4, it will be specified this operation.

In the of the described procedure, there is an initial route set and its headways.

C. Genetic Algorithm

In this chapter, we will describe Genetic Algortithm [20] [21], but with some improvements to include frequency setting .

Each gene is a route and its respective headway. Each population is a possible solution to the problem, which is comprised by a route set and its headways that vary throughout the day. In our case, a possible solution includes a route set and four headways, for each route, since we have four OD matrices.

The objective function attempts to minimize costs of both parties which take part on public transportation activity, passengers and transit operators.

1) Objective Function: The objective function defines the aim of the optimization process. In our case, this equation describes the quality of a solution, which intends to minimize costs of all stakeholders associated with public transportation operation. This costs can be given by T:

$$T = passenger\_weight * (ATT + AWT + w * ATR) + operator\_weight * ($$

Where each metric stands for:

- ATT: Average In-Vehicle Time. The average time a passenger spends in-vehicle when travelling
- AWT: Average Waiting Time. The average time a passenger has to wait at route stop
- ATR: Average Number of Transfers. The average number of transfers a passenger has to experience, w is a penalty equal to AWT
- AES: Average Empty Seats. The average number of empty seats across all network

Metrics are measured considering entity weight, which admits divergent interpretation of cost. While from passengers' point of view, cost means time taken travelling or waiting for the bus in the station, but also annoyance or discomfort on transferring between buses and routes. From operators perspective, empty seats are a sign of prejudice, as buses are

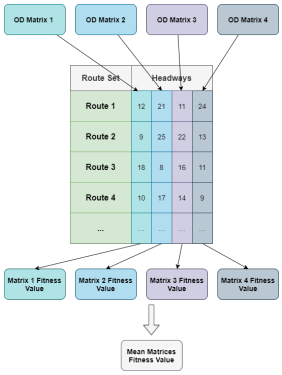

Fig. 1. Solution Topology

not operating at a reasonable occupation, although fuel cost and vehicles depreciation remain stable.

As we are dealing with a route set which contains different headways across matrices, there is a fitness value for each matrix (part of the day), hence we define that the most suitable solution is the route set and its respective headways that have the lowest mean fitness value.

2) Operators: Genetic operators guide the algorithm towards a solution to a given problem. They mimic natural world processes, such as survival of the fittest, reproduction or mutation.

3) Selection: In each generation, individuals are selected from the population to mate and generate offsprings for the next generation. Tournament Selection was used as the selection mechanism. The probability of selecting each individual i is given by

$$p_i = \frac{\Sigma_{j=1}^{N} f_j}{f_i},$$

where N is the population size and f_i is the value of the objective function of individual i.

4) Crossover: The crossover operator swaps routes at each route index, between the two selected parents with probability $p_{swap} = \frac{1}{R}$, where R is the number of routes.

5) Mutation: The mutation operator is applied to the two offsprings that result from the crossover step. The mutation process is designed to select routes that have less demand more often, as the probability of select one route is: [20]

$$p_l = \frac{\frac{1}{ds_{ij}}}{\sum_{q \in L} \frac{1}{ds_{rs}}}$$

where i and j are the terminals of route l, L is the route set, r and s are the terminals of route q, and ds_{ij} is the total number of entries in the OD matrix that are satisfied, without any transfer, by using route l, that connects terminals i and j.

After route selection, it is applied a small modification with a high probability p_{ms} and a big modification with a low probability $1 - p_{ms}$.

- ***Small modification:*** selects one of the route terminals with the same (0.5) probability and applies one of two operators:
  - **deletes** the selected route terminal, with probability p_{delete}.
  - **extends** the selected route, adding a new stop in the selected route terminal, with probability $1 - p_{delete}$ The new stop is chosen at random from among the adjacent stops to the chosen terminal, in the road network graph. If the selected route terminal is the route's first stop, the new stop is prepended to the route. If the selected terminal is the route's final stop, the new stop is appended to the route.
- ***Big modification:*** Selects one of the route terminals (say terminal i) with the same (0.5) probability, and then selects a new terminal k, with the following probability:

$$P_k = \frac{ds_{ik}}{\sum_{r \epsilon N} ds_{ir}}$$

The selected route will be replaced by a new route that is the shortest path between terminal i and terminal k.

6) Elitism: One of the inputs of Genetic Algorithm parameters is $elite\_size$. This value ensures that the most fitting individuals from the current generation are copied to the next one.

D. Frequency Setting

In this section, we will describe frequency setting procedures during Genetic Algorithm progress. At start, headway interval is defined between variables ($min\_headway$) and ($max\_headway$).

After that, two separated frequency setting methodologies can be followed:

- **Random Search**: Randomizes all the headway values associated to the route set between all .
- **Local Search**: progressively vary all the headway values associated to the route set.

1) Initial and Update Frequencies Process: At the procedure start, as regard to random search, initial headways are completely randomized. As Genetic Algorithm evolves, for each new generation, it is maintained random change in headway values. On the opposite, neighbour search starts off all headway with mean value between limit interval values, ($min\_headway$) and ($max\_headway$). After that, each new headway value is subject to a much smaller change operation. Consequently, each number may vary a difference of $[-5, 5]$.

2) Service Level across matrices: Between operator cost values associated with the same route set, our optimization algorithm does not allow disparities above a certain percentage throughout day planning.

In such option, our work ensures that there is a balance between demand and given level of service. Empirically, waiting times, in other words route headways, are way higher at night than during peak hours, as a consequence of decline on demand. Accordingly, keeping the same level of service (route headways) becomes financially unsustainable to transit company.

Peak hours are more profitable to operators than calm seasons.

As a way of simulating resources management, we considered this operation in this thesis.

3) Small Frequency Adjustment: When there is a disparity between two matrices operator costs above a pre-determined value, there is a method that attempts to converge these values.

On the one hand, by subtracting an integer $\in [0, 2]$ to the route headways (increase frequency) associated with the matrix that has higher fitness value and, on the other hand, by summing an integer $\in [0, 2]$ to the route headways (diminish frequency) of the lower fitness value matrix, we intend to normalize both values. The operatation is common to both frequency searches.

This means that through boosting route bus appearances (route frequency), matrix fitness value tends to decline.

E. Routes Overlapping

During genetic algorithm progress, when calculating fitness value of a matrix, there is a decisive step that aims to check whether exists equal segments among route set. By doing this, we intend to check if there are alternatives to an OD trip. It is not considered an option even if it includes that exact OD pair, but when there is the same OD segment. This option is due to the fact that the route generation method already contemplates shortest paths.

The aforementioned constraint is quite suitable as it approximates our work with the reality, particularly when calculating some metrics, such as empty seats and waiting time (method explained on II-B).

As far as empty seats is concerned, routes overlapping determination is absolutely vital, because can give us the number of bus appearances of routes that cover an intended OD segment.

Firstly, empty seats are calculated by iterating the bus capacity (in our case 85 seats, following Carris fleet capacity [1]), with the difference between boardings and alightings in each stop throughout a route progession. For each route, it will be computed the mean empty seats value of consecutive stops journeys.

In the end, the final average empty seats value to all the network corresponds to the mean value between every route empty seats value. Below a small illustration of Average Empty Seats metric calculation.

V. Results

This chapter presents the results of applying the optimization methodology earlier described to the bus network in Lisbon.

A. Evaluation and Metrics

In the first set of experiments, the routes were generated from scratch, using the initialization method described in Section 4. Several tests were conducted for the number of routes $R = 150$ in order to assess headway adjustment method (Local and Random) and metric weights variation (passenger and operator weight).

[1]https://www.carris.pt/descubra/frota

Additionally it is analysed in detail most optimized solution among experiments conducted.

Later, it is conducted another set of tests, where the existing (real) bus network was used as a starting point to the optimization process, serving as the initial solution. The analysis method is equal to the described on the above paragraph.

All tests do accept a maximum difference among matrices operator costs of 20%. We opted for this value because on small experiments, with less routes and iterations, the biggest disparity on matrices empty seats was frequently above this threshold.

Several combinations of values for the following parameters were used in both tests:

- **Operator Weight** ($operator\_weight$)
- **Passenger Weight** ($passenger\_weight$)
- Search Method:
  - **Local Search**
  - **Random Search**

To evaluate the results, the following metrics of objective function (described on Section 4) were considered:

- **Fitness Value**
- **Average In-Vehicle Time** (ATT)
- **Average Waiting Time** (AWT)
- **Average Number of Transfers** (ATR)
- **Average Empty Seats** (AES)
- **Passenger Cost** = ATT + AWT + AWT*ATR
- **Average Headway**

Genetic algorithm parameters were not the aim of this work, so we fixed them across all tests: pop_size=16, $elite\_size$=4, t=4, p_{ms}=0.7, p_{delete}=0.6

Across all tests, each route headway lies within an interval between $min\_headway$=7 and $max\_headway$=25.

B. Optimization from Scratch

In the initial tests with a smaller network of 150 routes, which are generated from scratch, we evaluated search frequency models and metric weights changes. We opted to assign values of 0.4 and 0.1 to operator metric, together with 0.6 and 0.9 to passengers', because we would like to test two situations. Firstly, where metrics have a balanced value (first case, 0.4 and 0.6), and the second, where there is a heavy importance given to passengers (0.1 and 0.9).

It is worth to mention that in our tests, passengers have always an higher importance assigned, since we have larger amount of information available associated with them and consequently more metrics to assess, than operators cost, which is limited regarding matrices values already. Another relevant reason is that the absolute value of empty seats is way greater than passenger costs putted together.

Search	Objective Function	ATT (min)	AWT (min)	Empty Seats
Random	39.83 (−10.62%)	6.68	5.725	77
Local	41.9 (−5.11%)	8.31	6.12	76

TABLE I
FINAL METRIC VALUES - $operator\_weight$=0.4, $passenger\_weight$=0.6

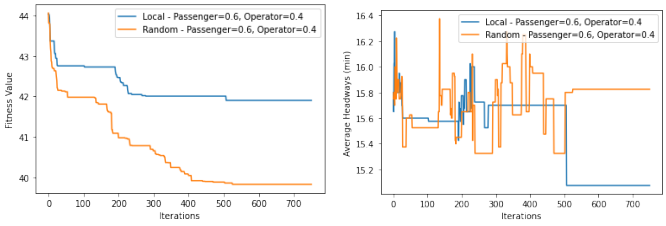

Fig. 2. Fitness Value and Average Headways - $operator\_weight$=0.4, $passenger\_weight$=0.6

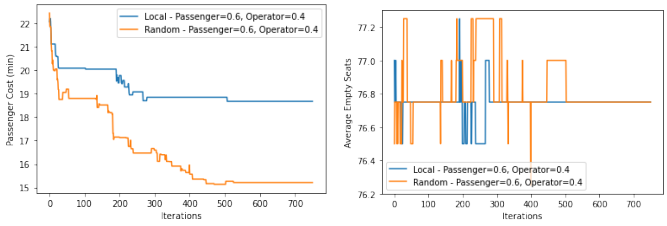

Fig. 3. Passenger Cost and Average Empty Seats- $operator\_weight$=0.4, $passenger\_weight$=0.6

Search	Objective Function	ATT (min)	AWT (min)	Empty Seats
Random	21.79 (−27.35%)	6.92	5.54	78
Local	23.17 (−21.75%)	7.45	6.01	77

TABLE II
FINAL METRIC VALUES - $operator\_weight$=0.1, $passenger\_weight$=0.9

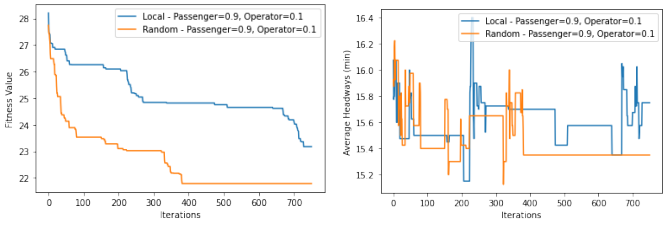

Fig. 4. Fitness Value and Average Headways - $operator\_weight$=0.1, $passenger\_weight$=0.9

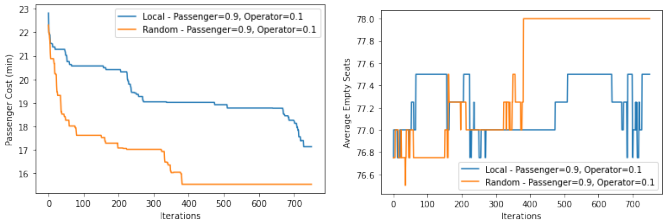

Fig. 5. Passenger Cost and Average Empty Seats - $operator\_weight$=0.1, $passenger\_weight$=0.9

First experiments show that Random Search method has a better performance than Local Search.

Random Search computes optimal solution after around 400 iterations, as well as, it rapidly decreases fitness during the first 200 iterations. While Local Search on 0.4/0.6 test soon reaches it optimal solution and barely changes during the rest of the process. On contrary, during 0.1/0.9 test it gradually adjusts it optimal solution, however nearly in the end it has a quick fitness decrease.

Regarding cost weight changes, on the test where it is given a greater importance to the operator cost (0.4), there is a reduction on average empty seats, which is natural given the bigger preoccupation to the company cost.

1) Optimized solution in detail: Using Random Search and $operator\_weight$=0.1, $passenger\_weight$=0.9, which was the best parameters combination (−27.35% of improvement on fitness value), we present in detail metrics for each matrix.

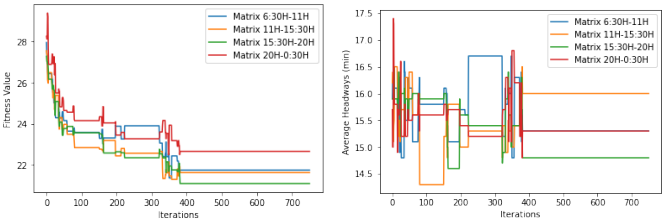

Fig. 6. Fitness Value and Average Headways - $operator\_weight$=0.1, $passenger\_weight$=0.9

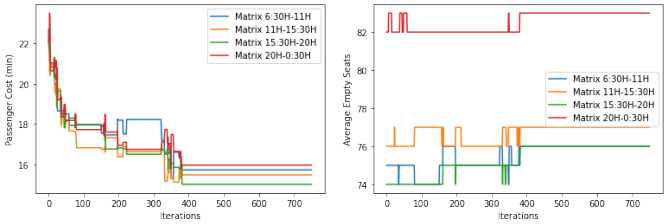

Fig. 7. Passenger Cost and Average Empty Seats - $operator\_weight$=0.1, $passenger\_weight$=0.9

Night period clearly has a greater fitness value, mostly because it has a higher average empty seats value, which is somehow acceptable given demand level. However, passenger costs are rather balanced among all matrices.

The huge average empty seats difference between fourth matrix and the others is not mirrored on the increase of matrix

fitness value, because in this experiment operator weight is reduced.

C. Optimization of the existing network

Similarly, we tested the methodology using the real route set of Carris as a starting point. Carris network comprises R = 308 routes.

The variations on passenger and operator weights were equal to the above experiments performed. Firstly, by assigning 0.6 to passengers' and 0.4 to operators metrics. On the following test, 0.9 to passengers costs and 0.1 to operator metric.

Search	Objective Function	ATT (min)	AWT (min)	Empty Seats
Random	41.49 (−1.45%)	10.01	7.58	72
Local	41.55 (−1.20%)	10.16	7.6	72

TABLE III
FINAL METRIC VALUES - $operator\_weight$=0.4, $passenger\_weight$=0.6

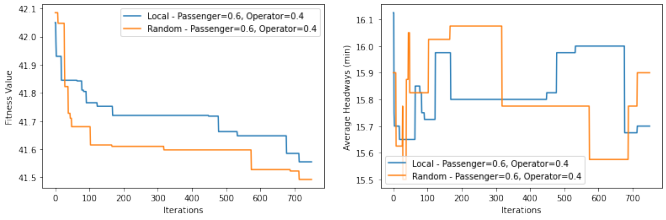

Fig. 8. Fitness Value and Average Headways - $operator\_weight$=0.4, $passenger\_weight$=0.6

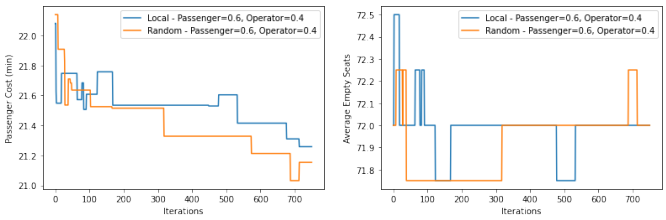

Fig. 9. Passenger Cost and Average Empty Seats - $operator\_weight$=0.4, $passenger\_weight$=0.6

Search	Objective Function	ATT (min)	AWT (min)	Empty Seats
Random	26.6 (−1.17%)	10.36	7.49	72.75
Local	26.64 (−0.53%)	10.4	7.57	72.5

TABLE IV
FINAL METRIC VALUES - $operator\_weight$=0.1, $passenger\_weight$=0.9

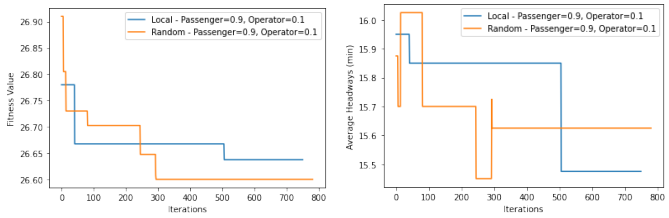

Fig. 10. Fitness Value and Average Headways - $operator\_weight$=0.1, $passenger\_weight$=0.9

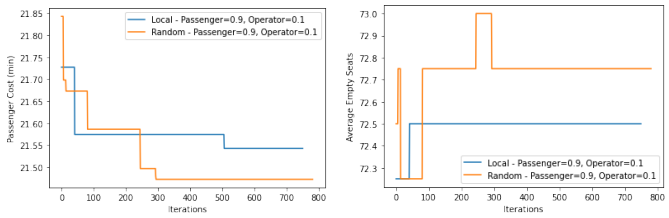

Fig. 11. Passenger Cost and Average Empty Seats - $operator\_weight$=0.1, $passenger\_weight$=0.9

Similarly to optimization from scratch, Random Search had a better performance than Local Search.

Curiously, the fitness improvement, in equal conditions of weights assignment, was pretty low compared to experiments with R = 150 (1.17 <<27.35). The reason may be the wider range of routes and headways number, which can lead to a more difficult optimization procedure.

1) Optimized solution in detail: Likewise, as it made on optimization from scratch analysis, we detail best solution matrices.

Using Random Search and $operator\_weight$=0.4, $passenger\_weight$=0.6, which had the best performance (−1.45%), we present in detail metrics for each matrix.

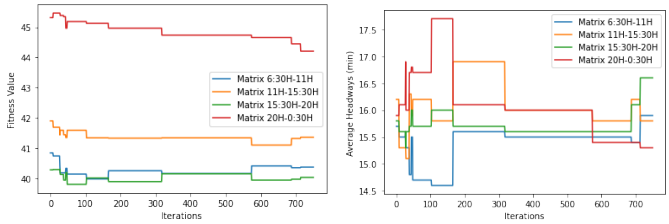

Fig. 12. Fitness Value and Average Headways - $operator\_weight$=0.4, $passenger\_weight$=0.6

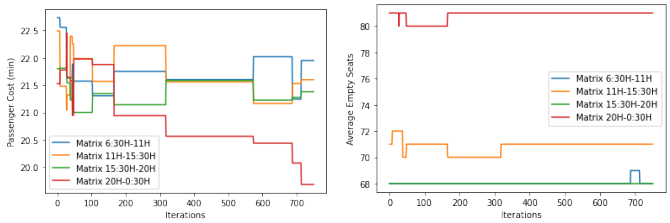

Fig. 13. Passenger Cost and Average Empty Seats - $operator\_weight$=0.4, $passenger\_weight$=0.6

The two matrices that include peak hours (blue and green lines) had practically similiar results. Moreover, the third matrix (orange) produced a slightly higher fitness value, due to higher average empty seats metric.

Fourth matrix (red) performed the highest number of empty seats and the lowest on passenger costs and headways as well, providing that more bus appearances result in emptier buses.

D. Overall Analysis

We may conclude that Random Search is the best performing method as it minimizes best the ojective function across all tests performed. The possible reason is that by randomizing values, we can rapidly test a large scope of possible solutions, with diverse headway values. Contrarily, Local Search slightly varies headway values through new generations. As a result of having a large set of routes and its associated headways, Local Search method leads to an homogeneity of headway values tested. Correspondingly, graphs plotted demonstrate the effective reduction on fitness values on Random Search experiments.

As far as metric weights is concerned, the heavier operator weight is, the higher fitness value is, which is somehow obvious, given absolute metric value differences. However, taking a closer look at the graphs, it is demonstrated on all experiments that when there is a certain balance between passenger costs (AWT + ATT + AWT*ATR) and average empty seats (AES). While there is a gradual reduction on passenger costs we can check also a compensation by increasing empty seats.

In perspective, it is consistent. A reduction on waiting time, caused by a decrease on headway (increase on bus appearances) and consequently leads to more empty seats. Plots of fixed max/min headway confirm that. Moreover, it is clearly observed that on matrices that include peak hours (first and third matrices, morning and afternoon, respectively), there are less average empty seats, as there is way more demand. On experiments, which was given a greater importance to operator, resulted in a reduction on empty seats.

Empirically, on our perspective as passengers, every public transportation operator adjusts service level, given demand. On the one hand, during peak periods, there are low route headways (high frequency) and reduced waiting time. On the other hand, during night hours there is an adjustment of the service level, mainly for management reasons. It is not financially feasible to keep high frequencies all day, as demand hugely vary as well.

By stating that, on both solution matrices plots, the previously mentioned does not occur. While there is a reduction on demand, mirrored by average empty seats metric matrices differences, the same adjustment on headways did not take place. As can be noted on plots, there is few difference among matrices average headways. Curiously, on peak times matrices, the headway is increased to compensate a drop on average empty seats. On the night matrix, the inverse happens. It is reduced average headways and consequently passenger costs, in order to balance the high number of empty seats.

The reason to all this is that our methodology computes the most optimised solution as the most balanced among matrices fitness values (Figure 1), hence the difference in headways (also waiting times) is low, which is further balanced by the operator's cost control during the day. So the set of headways that harmonize passenger and operator costs are considered the most appropriate to the optimization process.

E. Comparison to existing network

As far as real network comparison is concerned, *frequencies.txt* file present on GTFS data, describes trips headways of CARRIS real routes. Thus, we analyzed that file in order to assess our results, using same intervals.

Matrices	Optimized Average Headway (min)	Real Average Headway (min)	Difference (%)
6:30H - 11H	15.9	16.2	-1.8
11H - 15:30H	15.8	17.2	-8.8
15:30H - 20H	16.6	15.5	$+7$
20H - 0:30H	15.3	26	-70

TABLE V
COMPARISON BETWEEN REAL AND COMPUTED HEADWAYS ON RANDOM SEARCH - $operator\_weight$=0.4, $passenger\_weight$=0.6

On the first three periods, the computed headway values are somewhat similar to real ones. In the first, practically the same result, and on the second and third, a relevant decrease and increase, respectively.

Notwithstanding, on fourth (night) matrix there is an enormous reduction of 70% on average headway values.

VI. CONCLUSION

This thesis developed a methodology to tackle Transit Network Design and Frequency Setting Problem simultaneously. The computed method was applied to real transit data, provided by CARRIS, the bus operator of the city of Lisbon. The data sources utilized were GTFS and AFC data.

Firstly, it was inferred trip destinations, given that passengers only validate fare ticket at boarding on Carris Network. 47.7% of destinations were deduced. After that, a method to uniformly allocate uninferred destination trips throughtout routes was employed, with the intent of handle all available data to the forward steps. In the end of this process, we divided the demand in four different matrices that represent periods of a week day: morning, lunch, afternoon and night. These computed matrices were used as input data to approach the problem of the work.

With the resulting matrices, it was described network and road representations, as well as the Genetic Algorithm aspects, namely its operators and objective function to the optimization problem. Furthermore, it was explained a couple of actions about frequency adjustment, such as two different approaches regarding frequency optimization searching of solutions computed by GA: Local Search and Random Search. The results of the conducted experiments proved that Random Search has a better performance than Local method. One of the reasons is that small variations on headways, lead to an homogeneity on values and a small scope of probable values. On contrary,

Random Search rapidly tests a wide range of values and finds optimal values.

The methodology managed to find the optimized solution with lowest average matrices values, which resulted on matrices metric values quite similar to each other. Besides that, it has not been expressed demand fluctuation on metric values. For instance, waiting time and headways are not increased on night matrix, to compensate demand decrease.

Under these circumstances, we do believe that computed methodology represents a great step on real public transportation planning solution. Not only by how metrics are calculated, but also by how methodology was developed. Providing that, there is more data, namely from transit companies activity, this methodology can preview a wider range of situations and give a closer picture of reality. Perhaps by assigning day periods different weights to optimization process, this work can be a useful tool when tackling Transit Network Desing and Frequency Setting problems.

A. Limitations

The main limitation identified was the way operator costs were computed. Empty seats alone are not a good parameter to represent operator costs by itself, it is a rather simple model. Without real information, for instance about fuel costs or fleet maintenance, it became quite difficult to develop a more robust representation. Given available data, we fancy a way to estimate empty seats to symbolize it. Even though, having only access to AFC data, sometimes imprecise and inconclusive, it could not depict a real picture of boardings and alightings still. A more feasible way of calculating it would be desirable, namely taking advantage of Automated Passenger Counters (APC) system, if available.

Another relevant absence of data observed was AVL. From the Transit Network design, it would represent higher destinations and interchanges inferred rates. Moreover, as far as frequency optimization is concerned, AVL data available would mean a more comprehensive notion regarding bus appearances and a more meticulous analysis on waiting time, due to the fact that provides real traffic impact on bus activity.

Ultimately, the lack of matching between route stops on opposite directions forms an additional barrier. Since all of opposite stops are on the other side of the street, providing that there are many one-way streets. Without further information, it becomes quite difficult to give a closer picture of a real network which is composed by symmetric routes.

B. Future work

The natural path to this work could be to continue optimization process to the next steps of the public transportation planning, particularly timetabling development and resources allocation, included on tactical and operational choices, respectively.

As mentioned in the subsection of Limitations, if we had at disposal a matching between opposite route stops, a procedure that generates symmetric routes would promote a real bus network representation. In this manner, it would be quite natural to add another metric, fleet allocation to network. By calculating round trips duration this could result on number of buses required to each route.

In addition, other relevant aspect to consider in the future could be assign economic costs to the objective function. Besides fuel costs that are directly impacted by network length, it would favorable to add some metrics, such as salaries or fleet maintenance costs. On revenues, ticket fares revenues would be applied. These data could be kindly provided by the operator to further researches.

REFERENCES

[1] R. O. Arbex and C. B. da Cunha, "Efficient transit network design and frequencies setting multi-objective optimization by alternating objective genetic algorithm," *Transportation Research Part B: Methodological*, vol. 81, pp. 355–376, 2015.

[2] K. Gkiotsalitis, "Exact optimization of Bus Frequency Settings considering Demand and Trip time variations," *Presented at 96th Annual Meeting of the Transportation Research Board*, no. January, 2017.

[3] A. Ceder and N. Wilson, "Bus Network Design," *Bus Network Design*, vol. 1, no. 4, pp. 331–344, 1986.

[4] Sh. Afandizadeh, H. Khaksar, and N. Kalantari, "Bus fleet optimization using genetic algorithm a case study," 2011.

[5] G. Fattouche, "How to improve high-frequency bus service reliability through scheduling," *Proceedings of the ITRN*, no. 833, pp. 45–47, 2011.

[6] J. Walker, *Human transit: How clearer thinking about public transit can enrich our communities and our lives*. 01 2012.

[7] L. Fan, C. L. Mumford, and D. Evans, "A simple multi-objective optimization algorithm for the Urban transit routing problem," *2009 IEEE Congress on Evolutionary Computation, CEC 2009*, pp. 1–7, 2009.

[8] M. Wardman, "A review of British evidence on time and service quality valuations," *Transportation Research Part E: Logistics and Transportation Review*, vol. 37, no. 2-3, pp. 107–128, 2001.

[9] W. Wang, J. P. Attanucci, and N. H. Wilson, "Bus passenger origin-destination estimation and related analyses using Automated Data Collection Systems," *Journal of Public Transportation*, vol. 14, no. 4, pp. 131–150, 2011.

[10] J. J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda, "Origin and Destination Estimation in New York City with Automated Fare System Data," no. 02, pp. 183–187, 2002.

[11] J. B. Gordon, H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci, "Automated Inference of Linked Transit Journeys in London using Fare-Transaction and Vehicle Location Data," *Transportation Research Record*, vol. 2343, pp. 17–24, 2013.

[12] K. Deb, "Introduction to Genetic Algorithms for Engineering Optimization," vol. 1975, pp. 13–51, 2004.

[13] F. A. KIDWAI, B. R. MARWAH, K. DEB, and M. R. KARIM, "A Genetic Algorithm Based Bus Scheduling Model for Transit Network," tech. rep., 2005.

[14] A. Tošić and A. Horvat, "Optimization of traffic networks by using genetic algorithms," Tech. Rep. 4, 2012.

[15] "Gtfs google." https://developers.google.com/transit/gtfs/reference. Accessed: 2020-12-05.

[16] N. Nassir, A. Khani, S. G. Lee, H. Noh, and M. Hickman, "Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system," *Transportation Research Record*, no. 2263, pp. 140–150, 2011.

[17] "Carris." https://www.carris.pt/en/carris/company/our-numbers/. Accessed: 2020-12-05.

[18] "Metropolitano de lisboa." https://www.metrolisboa.pt/institucional/conhecer/metro-em-numeros/. Accessed: 2020-12-05.

[19] G. E. Sánchez-Martínez, "Inference of public transportation trip destinations by using fare transaction and vehicle location data," *Transportation Research Record*, vol. 2652, pp. 1–7, 2017.

[20] R. Loureiro, "Machine Learning Methods for the Optimisation of Urban Mobility," Master's thesis, Instituto Superior Técnico, 2020.

[21] M. A. Nayeem, M. K. Rahman, and M. S. Rahman, "Transit network design by genetic algorithm with elitism," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 30–45, 2014.