Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

PROJECT

A WASTE COLLECTION CASE STUDY

GONÇALO TORRES DUARTE

MARCH - 2022

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
PROJECT

A WASTE COLLECTION CASE STUDY

GONÇALO TORRES DUARTE

SUPERVISION:
LEONOR ALMEIDA LEITE SANTIAGO PINTO

MARCH - 2022

To my grandparents,

whom I lost during this

master's program.

Acknowledgements

I would like to express my deepest appreciation to Valorsul S.A. for providing me the data needed for this research and for having been helpful and caring through the project.

I am also grateful to my professor, Leonor, for the unwavering support, knowledge and extreme patience shown from the start of the project. I will forever cherish the effort I received from you and the way you always made me feel assisted. My gratitude goes beyond words.

Lastly, to my parents and my brother, whom I could not have undertaken this journey without.

Abstract

Innovative waste collection strategies have been established to replace conventional systems with dynamic systems that respond to the actual fill-level of waste containers. This work proposes a method for route design involved in the waste collection, focusing on minimizing the total distance travelled, for a case study of the Portuguese municipality of Caldas da Rainha. The process uses historical data of the filling percentage of each container to determine the number of weekly collections required per container to determine the number of weekly collections required per container and therefore create collection routes. More conventional approaches are based on a single average fill-up rate common to all containers, which may not represent the reality and therefore lead to inefficiencies. By analysing the available data, is concluded that all containers only require one weekly collection. A data clustering algorithm is then applied to aggregate container in groups corresponding to the same route, based on proximity. Then, routes are designed within each cluster. Finally, the selection of pairs of routes for the same day is done through a matching problem. Data limitations lie on the manual introduction of containers' fill level and on most historical data being from the pandemic.

The established methodology is applied to the glass waste collection and transportation system of Valorsul S.A. The present project aims at proposing a method for waste collection planning and explores critical considerations and future improvements based on the difficulties faced.

Table of Contents

Table Index

1. Introduction

Nowadays, the theme of waste management is crucial to society's well-being due to the implications it takes on public health and overall quality of life. Having a proper waste management is key to ensure functional quality cities as waste collection is directly related to air quality, street maintenance and traffic intensity.

Surely urbanization and changing waste generation levels make it vital, and also a big challenge, for responsible entities to innovate and develop waste management plans that respond to the current needs. The process of waste collection is expensive due to the various costs associated with both collection vehicles (from maintenance to fuel) and employees, and so trying to minimize inefficiency on the process as much as possible, must be a priority. Conventional routes, i.e., pre-defined circuits not taking into consideration neither containers' waste level nor any other factor that might influence waste generation, lead to containers that consistently are not at a collection required waste level when vehicles stop by. To overcome this problem, it is important to have both a data-driven approach towards route planning, including having carefully registered waste level data, and an understanding of the corresponding area at hand and its characteristics. Previous work (Bijsterbosch, J., & Volgenant, A., 2010) has shown that waste management is a complicated matter to generalize – i.e., it is a mistake to build a plan that is efficient to a certain city and simply apply it to cities with different characteristics, without the proper adjustments. Relatively simple differences may have a significant impact on how to design waste collection routes.

In fact, waste collection routing optimization brings positive impacts not just on a business perspective, but also on environmental matters, which, by living in a society where climatic changes are dangerously speeding up, should not be neglected. By reducing total duration of a set of routes, companies are also reducing pollutant gas emissions, which will always be one positive outcome of a structured waste management.

This project aims at contributing to the study on route optimization for waste collection by clustering containers, for a given area, and designing routes trying to minimize distance travelled by collection vehicles.

The report is organized as follows. In section 2, the motivation and problem definition are presented. Section 3 contains a literature review. Next, in section 4, data of a case study is described, followed by its analysis, in section 5, to determine the number of weekly collections each container requires. In section 6, the methodology used is detailed. Section 7 includes the application of the methodology to the case study that is discussed in section 8. Finally, the last section summarizes the main conclusions.

2. Motivation and Problem Definition

Managing waste collection is an issue that demands a structured and thoughtful plan from the responsible entities. Making sure the balance between implementing routes that are effective regarding costs to the company, and not compromising the quality of the service (no containers getting over filled in between collections) is key for the organization.

The current work was motivated by waste collection routing problem on the Portuguese municipality of *Caldas da Rainha.* To develop the desired analysis, real data was provided by the Portuguese company, Valorsul, S.A., which acts in several business units, such as energetic and organic valorisation, recyclable waste collection and sorting, amongst others. Valorsul, S.A. is the organization in charge of urban waste treatment in nineteen municipalities in Lisbon and West Portugal. The company is also responsible for the waste collection on fourteen of them.

The problem to solve through this project is on both understanding the number of required collections for each container on a weekly basis, given a certain geographic region, and designing the corresponding routes in order to minimize the total distance traveled by collection vehicles.

It is important to take into consideration the business needs and limitations when structuring waste collection. For instance, employees' daily routines stability is fundamental for the results to be attainable. It is relevant to distinguish tactical vs. strategic collection route planning. While a strategy consists of the plan that will lead the company to where it is desired, tactics are the numerous actions and phases required to achieve those goals – more than just a theoretic improvement from current plan, the tactical aspects seek to ensure practical implementation. (Ghiani, Laganà, Manni, Musmanno, & Vigo, 2014) developed a study on the most impactful work on Solid Waste Management (SWM), focusing on tactical and strategic issues. The authors conclude that, since SWM involves several different factors (as social, financial, economic, environmental and institutional), no model could capture all different relevant features, while not becoming extremely hard and time consuming. For example, consider a scenario where a container would have different weekly number of collections needed on three consecutive months. This may lead to different weekly collection routes for the three consecutive months, which brings difficulties on time and human resources management. Taking this into consideration, the granularity of seasonality to be studied on this project is a six-month period, leading to the possibility of having two different weekly plans through a year – arising the question: is the season (Summer and Winter), a significant variable when determining the number of weekly collections required per container?

Valorsul owns a total of around 5,000 containers, considering glass, paper and plastic waste – distributed amongst all geographic regions covered by the company – and 18 collection vehicles. From the fourteen municipalities that represent the company's portfolio, the one motivating this research – *Caldas da Rainha* – is the second biggest in terms of residents, representing 13% of the total population served, in 2019, accordingly to the company data. *Caldas da Rainha* was suggested by the Valorsul's team to be the focus of this project, since it is a diverse municipality, having a coastal zone, urban and non-urban areas.

We decided to develop the study only on the collection of glass waste containers, these consisting of 577 units.

After understanding the needs of weekly collections per container, the problem is to cluster containers based on proximity – each cluster then demanding for one, or more routes. Then, clusters are paired, leading to routes to be scheduled for the same day. Lastly, the challenge is to find a route (or a set of routes) that minimize total distance travelled by the vehicles, starting and ending at the waste sorting centre, while collecting every container. Note that, when designing waste collection routes, containers to collect can be represented either by edges or by nodes of the graph. On the one hand, regions, usually high population density cities, where each building has its own container that must be collected, containers are not considered individually, as they are spread along streets. In that case either all the containers of a street are collected or none and so arc routing problems are adequate. On the other hand, regions where there are common containers installed in specific points apart from each other, containers are represented by the nodes on the graph and modelled by a node routing problem. Regarding the sample described above, in this project the last case should be used.

3. Literature Review

Waste collection management is a challenge that requires careful preparation to be as smooth and efficient as possible. From building forecasting models to predicting the containers' waste level and a city waste generation, to designing vehicle collection routes, several research have been conducted.

Regarding forecasting models, it is crucial to clearly define the dependent variable as well as the predictors, existing different ways to approach this challenge. (Azadi & Karimi-Jashni, 2016), developed both Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) models to predict the Seasonal Municipal Solid Waste Generation (SMSWG) in 20 urban regions in Iran. In this paper, the goal was to analyse the average waste generation taking into consideration seasonality. An ANN model, with the explanatory variables population, solid waste collection frequency, maximum seasonal temperature and altitude, provide accurate forecasts for SMSWG. Regarding the MRL model, results were not as satisfactory, having the difference between predicted and observed values been considered statistically significant.

Note that there is a wide variety of variables to be possibly considered when training and testing a model to predict waste generation. (Kontokosta, Hong, Johnson, & Starobin, 2018), built a model to predict waste generation considering more than 750,000 New York City residential buildings. On this research, three categories of predictors were considered: urban, demographic/socioeconomic and seasonal. For urban variables, the percentage of city-owned buildings was taken into consideration as well as the percentage of commercial, retail, and residential areas and households per building. Regarding demographic and socioeconomic predictors, the employment rate, percentage of population with at least a bachelor's degree, median gross rent, household income and percentage of households living alone were some of the explanatory variables. Lastly, to measure the impact of seasonality, some of the variables were: the average temperature, total precipitation and snowfall amount, average wind speed and number of severe weather events. Studying the building-level waste generation with a considerable level of detail regarding predictors, allowed the model to obtain 93.9% accuracy, concluding that achieving a competent route planning depends on a proper estimation of the amount of waste generated. Another positive implication is the possibility to understand which areas are more and less prone to recycling, opening doors to a possible plan to educate society on its importance.

As mentioned on the two papers described above, seasonality is a factor that can affect waste generation and therefore collection routes planning. (Ferreira, Figueiredo, & Oliveira, 2017) developed models trying to improve forecasts regarding frequency of collection needed on a monthly basis, using seasonal factors for each month, from 2010 to 2014 in six municipalities in Northern Portugal. The goal was to predict the number of collections per year, per container (using ANN and MLR) and then attribute each month a value based on seasonality. Taking seasonality into consideration led to an estimated 10% decrease on stops at empty containers. Reducing the number of visits to empty, or almost empty containers, is an important factor to the company since it represents the reduction of unnecessary costs.

In fact, not only understanding the waste generation is important, but also designing efficient collection routes. Waste collecting vehicles represent an expense for the companies and therefore it is needed to monitor routes' costs. (Abdallah, Adghim, Maraqa, & Aldahab, 2019), aimed at developing a model to monitor these costs, by considering fuel and labour expenses. The following metrics were considered:

$$C = C_t * D + N_w * C_w * T_c$$

In which C represents total operation cost, in $\mathrm{US\$}$, C_t is the transportation cost per kilometre, in $\mathrm{US\$\ km^{-1}}$, D stands for route driving distance, in km, N_w is the number of employees per route, C_w is the average labour expense per working worker hour, in $\mathrm{US\$\,worker^{-1}hour^{-1}}$ and T_c represents total duration of the collection route, in hours. To determine the total duration of the collection, the following formula was considered:

$$T_c = (N_b * T_b) + T_r$$

being N_b the number of containers to be collected, T_b the time required to unload each container, in hours, and T_r the actual driving time of the trip, also measured in hours. One challenge regarding monitoring collection routes' costs lies on the difficulty to determine, *a priori,* the real distance travelled by vehicles. When planning routes, a simple method to calculate distances is the Euclidian distance – note that, nowadays, there are packages available to determine real travel distances and many municipalities already use them. Also, note that the formulas presented above do not take into consideration all costs, as e.g. the natural vehicle depreciation, which also brings expenses to the company. Nonetheless, (Abdallah, Adghim, Maraqa, & Aldahab, 2019) presents an interesting method to control and keep track of collection costs.

Nowadays, having a social and environmental consciousness is important, not just for each one of us as members of a society, but also for the companies themselves, which must follow restricting pollutant emission laws, depending on the country and industry. (De Marco, Mangano, & Zenezini, 2018) developed a study with the goal of identifying legal measures with regard to urban logistics. They concluded that authorities were intervening mainly through laws regarding urban consolidation centres and low-emission vehicles and zones.

Considering environmental measures when solving a Vehicle Routing Problem (VRP) – Sustainable Vehicle Routing Problem – is a topic that has gained attention over the last years, despite not being discussed in the past. (Dündar, Ömürgönülşen, & Soysal, 2021), performed a research on 148 papers to understand considerations taken when planning vehicle routing on economic, environmental, and social dimensions. Through the paper, it was possible to conclude that sustainability related matters were the least considered and economic matters the most highlighted. Combining social and environmental topics, (Hailin, Tao, & Yang, 2020) developed a study about optimization of vehicle routing for waste collection – to solve the problem of designing optimized circuits, minimizing total distance, gas emissions and costs (both vehicle related and gas emission related costs). Routes designed were sensitive to environmental matters (by controlling pollutant emissions and its costs) as well as social matters (by attributing each container with a priority level, considering proximity to places such as hospitals, schools and gas station, will lead to high priority containers).

(Zdena, Semiao, & Beijoco, 2013), developed a project to measure the impact of dynamic routes on fuel consumption and pollutant emissions – more precisely CO, CO2 and NOx – by comparing results with similar routes consisting of constant and dynamic loads in the municipality of Barreiro, Portugal. The study concluded that considering distinct fulfilment rates for each container, depending on past data, instead of an average for all containers, and using a VRP solving software, would lead to less collection time and less pollutant emissions. More recently, (Simoni, Bujanovic, Boyles, & Kutanoglu, 2018) concluded that opening consolidation centres and choosing electric collection vehicles instead of conventional vehicles might lead to a reduction of total costs by around 20% and CO2 emission cost by around 40%.

Through this literature review, it is important to understand the scope of research on collection routes, some being more waste generation prediction oriented and some more focused on route planning. Note that it is difficult to generalize the explanatory variables to be considered on a waste generation prediction model, since it is important to understand the characteristics of the region in study – some factors might be relevant for a geography and not for other. Lastly, the papers mentioned above also alert to the importance of an effective route planning and the effects on pollutant emissions. Living in a world where climatic changes are becoming more and more intense, should alarm us to be as conscious as possible regarding environmental and social related issues.

4. Data Sources and Description

As mentioned on section 2 of the project, the data used concerns glass waste containers from Portuguese municipality of *Caldas da Rainha*. The current section aims at presenting data provided by Valorsul, including column descriptions and examples of each of the two excel files made available by the company. The first file, *Listagem_Contentores_Caldas_Rainha.xlsx,* stores 577 records, describing each container, having the following columns:

TABLE I

DATA SOURCE I – COLUMNS DESCRIPTION

Column	Description
Codigo_Interno_do_Local	Location internal code.
Tipologia_do_Local	Whether it is an ecopoint or an island.
Descricao	Container id.
Tipo_de_Residuos	Waste category – glass (*Embalagens de Vidro*).
Capacidade	Container maximum waste capacity, in litres.
Tipo_de_Instalacao	Whether the container is above or under ground.
Latitude	Container's location latitude.
Longitude	Container's location longitude.

An example of two rows of the data stored in the document is given in Table II.

TABLE II

DATA SOURCE I – EXAMPLE

Codigo_Interno_do_Local	*Tipologia_do_Local*	*Descricao*	*Tipo_de_Residuos*	*Capacidade*	*Tipo_de_Instalacao*	*Latitude*	*Longitude*
954	Ecoponto	V1051	Embalagens de Vidro	2500	Superfície	39,44301218	-9,069529415
2704	Ecoponto	T_V0271	Embalagens de Vidro	3000	Enterrado	39,40402222	-9,133680556

The second file, Caldas_da_Rainha_Enchimentos.xlsx, stores 107,467 records, of 577 containers, from November 1st 2019 to march 1st 2021. Each row corresponds to a new record of a container's waste level, registered on a given day/time. Each time a container is collected the worker, before emptying it, records its filling level as one of the following five possible values: 0%, 25%, 50%, 75% or 100%. The document has five columns, as displayed in Table IV. The description of the columns is in Table III.

TABLE III

DATA SOURCE II – COLUMNS DESCRIPTION

Column	Description
Contentor	Container id.
Tipo_de_Residuos	Waste category – glass (*Embalagens de Vidro*).
Capacidade	Container maximum waste capacity, in litres.
Nivel_de_Enchimento	Waste level recorded (0/25/50/75/100 % - 0% corresponding to an empty container and 100% to a completely full one).
Data_da_Leitura	Timestamp of the register.

Similarly to the first data source, an example of two records is presented in Table IV.

TABLE IV

DATA SOURCE II – EXAMPLE

Contentor	*Tipo_de_Residuos*	*Capacidade*	*Nivel_de_Enchimento*	*Data_da_Leitura*
V0288	Embalagens de Vidro	2500	25	2019-12-12 19:36:49
V0345	Embalagens de Vidro	2500	75	2021-02-08 19:13:18

Note that the employees have five options when introducing the level of a container, these being 0% if empty, 25%, 50%, 75% and 100% if full. Data sources described in this section will be the source for sections 5, Data Analysis, and 7, Case Study.

5. Data Analysis

The goal of this section is, initially, for each container and for both summer and winter seasons (defined ahead), to get the number of weekly collections required, for past data. Then, based on past data, forecasts are developed for the future, if possible.

Three different indicators will be presented on the end of this section, per season, per container:

- *Increasing rate 25%* (ir_{25}) – days required, on average, for a container to increase 25% of waste level.
- *Increasing rate 75%* (ir_{75}) - days required, on average, for a container to increase 75% of waste level.
- *Weekly collections* – based on the previous indicators, the number of weekly collections needed.

To develop the previous metrics, the following transformations were performed:

i) Split data in two seasons (boolean column "Season": from April until September, summer – "S" and from October until March, winter – "W"). Splitting data in two seasons will allow the analysis to possibly come up with different conclusions for both – especially since the sample in study corresponds to a geography with beach, which might result in an increase in waste generation on summer season. Note that there was no further research developed on which months would be the most appropriate to split seasons on, to get a significant variable regarding season. The split used was advised by the company based on knowledge and experience on the municipality in study.

ii) Remove two types of records: records with waste level difference from the previous equal to zero (meaning the waste level registered was the same comparing to the previous record) and waste level difference negative (records corresponding to collections). Note that collections records are removed due to the fact that right before collecting, the employee registers the current level (which is not removed) and then once again after collecting (which is removed). Then, two auxiliar metrics are considered on the data-frame, these being:

- $Diferenca\_Tempo_j$ – time between the two (consecutive) observations $j+1$ and j, in seconds.
- $Ponderacao_j$ – weight to be given to each record considering the waste level difference between the consecutive observations $j+1$ and j, following the rule:

TABLE V

RULE TO GENERATE COLUMN *PONDERACAO*

Waste level difference	*Weight*
25%	1
50%	2
75%	3
100%	4

Note that a waste level difference of 100% corresponds to a container going from empty to full on consecutive observations.

The first indicator, *Increasing rate 25%*, is then determined by the formula:

$$ir_{25} = \frac{\sum_{j=1}^{n-1} Diferenca\_Tempo_j}{86400 * \sum_{j=1}^{n-1} Ponderacao_j},$$

$$n = number\ of\ observations\ per\ container, per\ period$$

With 86,400 corresponding to the conversion from seconds (unit on the column *Diferenca_Tempo*) to days (unit on ir_{25}) – i.e., the number of seconds in a day. Note that $86400 = 60 * 60 * 24$, one day having 24 hours, each hour 60 minutes and each minute 60 seconds. An example can be seen on Table VI.

TABLE VI

INCREASING RATE 25% - EXAMPLE

Nivel_de_Enchimento	*Data_da_Leitura*	*Diferenca_Enchimento*	*Ponderacao*	*Diferenca_Tempo*
25	2022-01-01 12:15:00	-	-	-
50	2022-01-04 13:30:00	50	2	263 700
75	2022-01-06 19:45:00	25	1	195 300

$$ir_{25} = \frac{263700 + 195300}{86400 * (2 + 1)} \cong 1.77$$

Then, after calculating ir_{25}, ir_{75} comes as follows:

$$ir_{75} = 3 * ir_{25}$$

Notice that, for summer data, there is only information regarding one period – summer of 2020. Regarding winter season, there are two periods to be considered – October 2019 / March 2020 (winter 2019) and October 2020 / March 2021 (winter 2020). There are seven containers with ir_{25} only for winter 2019 period (no value for summer 2020 or winter 2020 means the containers were most likely removed after winter 2019). Corresponding records are then removed.

iii) Lastly, the third indicator – number of weekly collections required for each container per period – is determined following the rule:

TABLE VI

RULE TO DETERMINE NUMBER OF WEEKLY COLLECTIONS

Increasing rate 75%	*Weekly collections*
Mais de 7 dias	1
Entre 3 e 7 dias, inclusive	2
Menos de 3 dias	3

Seasons with no increasing rates for a particular container will return 0 weekly collections needed for the corresponding container – indicating that the container may be removed or relocated.

Below are two summary tables describing both the increasing rates and the number of weekly collections determined by the process described above. Table VII stores information about indicators $ir25$ and $ir75$ (average number of days required for a container to increase 25% or 75% of waste level). The first column contains the statistic each row refers to. Columns two to four contain $ir25$, column two summer 2020 period, column tree Winter 2019 and column four Winter 2020. The last three columns have the information about $ir75$ for the tree periods considered. The first row identifies the number of containers considered, then each row displays a statistic for the two indicators in the different periods: the second row refers to the mean, the third the standard deviation, the fourth the minimum, then first, second and third quartiles, and, the last row the maximum. Table VIII displays the number of containers requiring each different possible number of weekly collections, per period.

TABLE VII

STATISTICS PER PERIOD – *INCREASING RATE*

	Increasing rate 25% Summer 2020	*Increasing rate 25% Winter 2019*	*Increasing rate 25% Winter 2020*	*Increasing rate 75% Summer 2020*	*Increasing rate 75% Winter 2019*	*Increasing rate 75% Winter 2020*
Count	551.00000	508.00000	513.00000	551.00000	508.00000	513.00000
Mean	12.552940	11.178425	11.256589	37.658820	33.535276	33.769766
Std	14.710878	6.426222	7.256014	44.132634	19.278665	21.768042
Min	1.870000	1.940000	0.870000	5.610000	5.820000	2.610000
25%	6.845000	6.767500	6.320000	20.535000	20.302500	18.960000
50%	9.530000	9.775000	9.580000	28.590000	29.325000	28.740000
75%	13.970000	13.962500	13.700000	41.910000	41.887500	41.100000
Max	180.42000	45.510000	68.190000	541.26000	136.53000	204.57000

TABLE VIII

NUMBER OF CONTAINERS REQUIRING EACH DIFFERENT POSSIBLE NUMBER OF WEEKLY COLLECTIONS, PER PERIOD

	Weekly collections Summer 2020	*Weekly collections Winter 2019*	*Weekly collections Winter 2020*
No information	1	44	39
1 weekly collection	550	506	510
2 weekly collections	1	2	2
3 weekly collections	0	0	1
Total number of containers with information	551	508	513
Total number of containers	552	552	552

The last goal of this section was to train and test different forecasting models to predict the number of weekly collections required for each container, per season.

There are two problems when trying to develop this analysis: on the one hand, there are few historical data to be considered (note that, for summer season, there is only one period of past data) which makes it difficult to develop a fairly accurate forecast; on the other hand, when considering the number of weekly collections per container for the three periods available, 99.62% of the observations indicate one weekly collection is enough.

Taking both points above into account, all containers will be collected once a week through the development of this analysis.

6. Routes Design

After concluding that collections must be done once a week for all containers, this section focuses on designing a set of routes considering the two vehicles available. To do so, containers are aggregated into clusters, then the clusters are matched to find the routes that should be scheduled for the same day. Lastly, within each cluster, a route is designed.

Collections are performed six days per week – from Monday until Saturday – with a total capacity of $150m3$ of raw waste per vehicle. Note that raw waste is the waste on its form as it is on the container. As it is collected, vehicles have a proper mechanism to compact the waste reducing its volume. The maximum duration of a collection route is eight hours, with a break of forty-five-minute rest for collaborators. Regarding total distance, there are no limitations, as long as duration and capacity restrictions are satisfied. Note that distance travelled is related to both duration and number of containers being collected, which by having restrictions on these last two, will consequently be limited.

6.1. Clustering

As mentioned above, the first stage of this section is to aggregate containers into clusters, to then design routes for each cluster. To cluster containers, K-Means constrained clustering algorithm is used.

Cluster analysis aims at splitting data into groups so that, in the same group, objects are related and also different from objects in the other groups.

The determination of k clusters can be formulated as a nonlinear programing problem. Assume that a dataset $D = \{y^i, i = 1, \dots, m\}$ of m points in R^n should be clustered in k groups. For each group h it should be found a cluster centroid C^h so a set of $C = \{C^1, C^2, \dots, C^k\}$ points in R^n must be defined. Another set of variables $X = \{x_{i,h}, i = 1, .., m,\ h = 1, \dots, k\}$ is used to assign each point y^i to a cluster,

$$x_{i,h} = \begin{cases} 1 & if\ y^i\ is\ assigned\ to\ a\ cluster\ h \\ 0 & otherwise \end{cases} \quad , i = 1, \dots, m, \quad h = 1, \dots k$$

The nonlinear programming problem can be stated as follows:

$$\min \sum_{i=1}^{m} \sum_{h=1}^{k} x_{i,h} \cdot \left(\frac{1}{2} \left\| y^i - C^h \right\|^2 \right)$$

$$\text{subject to } \sum_{h=1}^{k} x_{i,h} = 1, \qquad i = 1, \dots, m$$

$$x_{i,h} \geq 0, \qquad i = 1, \dots, m,\ \ h = 1, \dots, k$$

where $\|a - b\|$ stands for the Euclidean distance between points a and b.

(Bradley, Bennett & Demiriz, 2000) proved that variables $X = \{x_{i,h}, i = 1, .., m,$ $h = 1, \dots, k\}$ can be continuous nonnegative instead of binary.

(Bradley, Bennett, & Demiriz, 2000) proposed an algorithm that starts with an initial set of centroids and each iteration contains two main steps. The assignment step, where every y^i point is assigned to the cluster of the closest centroid, and a step that updates centroids based on the assignment. The algorithm end when the update step results in the centroids of the previous iteration.

The solution found this way may include some clusters with a very large number of points and other clusters very small. Since clusters will be the source for routes, it is crucial that a balance between cluster sizes is maintained. This task was accomplished by k-Means constrained in which a minimum and maximum size for each cluster is specified – the Python Package Index (PyPI). The package was developed based on (Bradley, Bennett, & Demiriz, 2000) and includes methods to add restrictions to the already existing K-Means clustering algorithm with the objective of avoiding solutions with considerable discrepancies on cluster sizes. The choice of a K-Means instead of other algorithms was solely based on the existence of an up-to-date package that allows cluster sizes to be constrained a priori.

The input is a set $D = \left\{y^i, i = 1, \dots, m\right\}$ of m points in R^n, k the number of clusters and $\tau_h \geq 0,\ h = 1, \dots, k$, the minimum number of points on each cluster. The centroid of cluster h at iteration t is denoted by $C^{h,t}$.

The constrained K-Means clustering algorithm

0. **Initialization.** Set initial centroids $C^0 = \{C^{1,0}, C^{2,0}, \dots, C^{k,0}\},\ t = 0$.
1. **Cluster Assignment.** With cluster centroids $C^t = \{C^{1,t}, C^{2,t}, \dots, C^{k,t}\}$ solve the linear problem:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{m}\sum_{h=1}^{k} x_{i,h}\cdot\left(\frac{1}{2}\left\|y^i - C^{h,0}\right\|^2\right) \\
\text{subject to} \quad & \sum_{i=1}^{m} x_{i,h} \geq \tau_h, \qquad h = 1,\dots,k \\
& \sum_{h=1}^{k} x_{i,h} = 1, \qquad i = 1,\dots,m \\
& x_{i,h} \geq 0, \qquad i = 1,\dots,m, \;\; h = 1,\dots,k
\end{aligned}$$

2. Cluster Update. Update $C^{h,t}$ as follows:

$$C^{h,t+1} = \begin{cases} \frac{\sum_{i=1}^{m} x_{i,h}^{t} y^i}{\sum_{i=1}^{m} x_{i,h}^{t}} & if \;\; \sum_{i=1}^{m} x_{i,h}^{t} > 0 \\ C^{h,t} & otherwise. \end{cases}$$

If $C^{h,t+1} = C^{h,t}$ for every $h = 1,\dots,k$, STOP

$t = t + 1.$

To finish, the constrained K-Means algorithm takes a finite number of iterations, leading to a solution that is locally optimal, proved in (Bradley, Bennett, & Demiriz, 2000).

6.2. Matching

After aggregating containers into clusters, the next phase is to pair clusters for the same day routes. If previously the goal was to minimize distance between points and their corresponding centroid, on this section the objective is to maximize distances between centroids, so that the two vehicles are as far from each other as possible, on each day.

Indeed, the problem faced is defined on a nonbipartite matching. Considering a complete graph, $G = (V, E)$, in which $V = \{1, .., k\}$ represents the set of centroids (vertices) and $E = \{[i,j]: i, j \in V\}$ the set of edges, a matching M, of the corresponding graph, is a set of edges with no common vertices. The problem of finding a maximum distance (weight) matching may be formulated as an integer programming problem. Consider the binary variables:

$$v_{ij} = \begin{cases} 1, & if\ [i,j] \in M \\ 0, & if\ [i,j] \notin M \end{cases} \qquad , \forall [i,j] \in E$$

The problem can be stated as follows:

$$\max \sum_{[i,j] \in E} w_{ij} v_{ij}$$

$$\text{subject to} \sum_{j:[i,j] \in E} v_{ij} = 1, \qquad for\ i \in V$$

$$v_{ij} \in \{0,1\}, \qquad for\ [i,j] \in E$$

where w_{ij} stands for the Euclidean distance between C^i and C^j, the centroids of clusters i and j.

Note that, in this project the number of clusters is always even. Whenever the number of vertices of the graph is even the inequalities of the formulation may be replaced by equalities.

The optimal solution for the problem may be achieved in polynomial time. The software presented by (Lu, Greevy, Xu, & Beck, 2011) was used and, regarding the distances, it is assumed $w_{ii} = -\infty, i = 1, \dots, k$.

6.3. Vehicle Routing Problem

Having containers aggregated in clusters, this section aims at defining the routes on each cluster, starting and ending at Valorsul waste sorting center – referred as the depot – which can be represented by a VRP. "The VRP is one of the most frequently encountered optimization problems in logistics, which aims to minimize the cost of transportation operations by a fleet of vehicles operating out of a base" (Erdoğan, 2017).

In fact, the problem can be identified as a Capacitated Vehicle Routing Problem (CVRP), which "is the most studied version of the VRP". (Irnich, Toth, & Vigo, 2014) described the CVRP as a problem in which goods are distributed from a single depot – represented by point 0 – to a defined number of n points, denoted as customers $N = \{1,2, \dots n\}$. Customers' demand can be defined as the amount to deliver to each customer, $q_i \geq 0,\ i \in N$. Regarding routing, p vehicles are available at the depot and all are considered to have the same capacity $Q > 0$ and operating at similar costs. A vehicle that delivers customer goods to a subset

$S \subseteq N$ starts and ends its trajectory on the depot, passing once and only once by each customer in S. For the use case in study on this project, customers are represented by containers and customers demand can mirror waste to be collected on each container ($q_i \geq 0$), measured in m^3.

Assuming that p vehicles are available and consequently p routes must be found, the variable $y_{ijk},\ i,j \in N \cup \{0\},\ k \in \{1,2,\dots,p\}$ will have value 1 if the vehicle k goes from container i to j in its route, and zero otherwise.

Considering the subset $S = \{i_1, \dots, i_s\} \subseteq N$ of customers to be visited in a cluster, a route can be defined as a sequence $r = (i_0, i_1, i_2, \dots, i_s, i_{s+1})$ – in which $i_0 = i_{s+1} = 0$ represent the start and end at the depot. For the problem of this project, feasibility requirements for routes lie on respecting the capacity restriction – $\sum_{i \in S} q_i \leq Q$. Additionally to the capacity requirement there is also a time limit L that must be observed. Assuming $t_{ij} > 0$ is the travel time between vertex i and j including time required to collect container j, this constraint is given by: $\sum_{i,j \in N \cup \{0\}} t_{ij} y_{ijk} \leq L,\ \forall k \in \{1,2,\dots,p\}$. If clusters $S_1, S_2, \dots, S_p$ form a partition on N and the corresponding routes $r_1, r_2, \dots, r_p$ are feasible, these set of routes is a feasible solution to the CVRP.

Considering that d_{ij} is the distance between containers i and j, and u_{ik} is the accumulated waste collected by vehicle k when arriving at container i, the problem of minimizing total distance travelled can be stated as:

$$\min \sum_{k=1}^{p}\sum_{i=0}^{n}\sum_{j=0}^{n} d_{ij}y_{ijk}$$

$$\text{subject to} \quad \sum_{j=1}^{n} y_{0jk} = 1 \qquad k = 1, \dots, p \qquad (1)$$

$$\sum_{\substack{j=0,\\ j\neq i}}^{n} y_{ijk} = \sum_{\substack{j=0\\ j\neq i}}^{n} y_{jik} \qquad i = 0, \dots, n, \quad k = 1, \dots, p \qquad (2)$$

$$\sum_{k=1}^{p}\sum_{\substack{i=0\\ j\neq i}}^{n} y_{ijk} = 1 \qquad j = 1, \dots, n \qquad (3)$$

$$\sum_{i=0}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} q_j y_{ijk} \leq Q \qquad k = 1, \dots, p \qquad (4)$$

$$\sum_{\substack{i=0\\ j\neq i}}^{n}\sum_{j=1}^{n} t_{ij} y_{ijk} \leq L \qquad k = 1, \dots, p \qquad (5)$$

$$u_{ik} - u_{jk} + Qx_{ijk} \leq Q - q_j \qquad i, j = 1, \dots, n, \quad k = 1, \dots, p \qquad (6)$$

$$q_i \leq u_{ik} \leq Q \qquad i = 0, \dots, n, \quad k = 1, \dots, p \qquad (7)$$

$$x_{ijk} \in \{0,1\} \qquad k = 1, \dots, p, \quad i, j = 0, \dots, n \qquad (8)$$

Constraints (1) ensure that every vehicle leaves the sorting centre. By the second set of constraints, it is guaranteed that a vehicle arrives and leaves a vertex the same number of times. Restrictions (3) force every node to be visited once and only once. The fourth and fifth groups forbid that the capacity and the time limit is exceeded by a vehicle. Lastly, the sixth set of constraints eliminate possible subtours and is adapted by the MTZ-formulation, proposed by (Miller, Tucker, & Zemlin, 1960). If $y_{ijk} = 1$ then $u_{jk} \geq u_{ik} + q_j > u_{ik}$ which ensures that the route for vehicle k does not include subtours by defining an order on the vertices with variables u_{ik}. Constraints (7) fix the bounds on variables u_{ik}.

Note that clustering containers is the first task proposed to solve the CVRP – which is developed on section 6.1 – allowing to reduce problem's dimension and also being needed due to limitations on the number of nodes to be included on a single run by the software used to solve the CVRP in the case study. If each cluster created on section 6.1 aggregates the containers for only one route, then for each cluster, the problem to solve is the Travelling Salesman Problem (TSP). In that case, the constraints regarding the capacity of the vehicles and the time limit were checked a posteriori. However, while clustering the containers in the first phase an important information provided by the company was used. It is known that the average number of containers required to fulfill a vehicle is 60, and that the duration of routes that collect 60 containers are within the time limit. So in the present case study the maximum number of containers per cluster (if a cluster is to correspond to one route) will be set to 50 (around 83% of its total capacity on average). This approach successfully ensured that the methodology used conducted to routes satisfying the requirements.

7. Case Study

The goal of this section is to apply all methodology described in section 6 to the case study of glass waste of *Caldas da Rainha* – clustering, matching and VRP solver application.

Valorsul covers fourteen municipalities regarding waste collection, *Caldas da Rainha* corresponds to 13% of total population, as may be seen in Table IX.

TABLE IX

DISTRIBUTION OF RESIDENT POPULATION AMONGST MUNICIPALITIES COVERED BY VALORSUL

Municipality	*Resident Population 2019*	*%*
Alcobaça	53 555	13%
Alenquer	44 052	11%
Arruda dos Vinhos	15 412	4%
Azambuja	22 716	6%
Bombarral	12 558	3%
Cadaval	13 673	3%
Caldas da Rainha	***51 912***	***13%***
Lourinhã	25 855	6%
Nazaré	14 165	4%
Óbidos	11 850	3%
Peniche	26 501	7%
Rio Maior	20 379	5%
Sobral de Monte Agraço	10 651	3%
Torres Vedras	78 530	20%
	401 809	**1**

Applying this ratio to the total number of collecting vehicles, 18, a total of 2 vehicles will be considered available for the municipality in study.

As described previously on the report, clustering containers is the first step to then design routes. For comparison purposes on total distance traveled, three different scenarios will be considered, regarding three different possible approaches on cluster sizes and on the number of routes to be designed per cluster. These scenarios will be the base of twelve routes, are defined as follows:

Scenario 1) twelve clusters are created, each corresponding to one route; Scenario 2) four clusters are created, each demanding for three routes;

Scenario 3) three additional clusters are created for each cluster in the scenario 2, generating then a route per cluster.

Common to the three scenarios, routes will then be paired based on maximum cluster centroid distance, to be collected on the same day.

To sustain the current section, the focus will be on scenario 1. As there are 523 containers to distribute across twelve clusters – of the 552 containers to be included on routes, 29 did not have geographic information on data source I, therefore were removed. Through the project, the balance between the number of containers per cluster was achieved due to the use of K-Means constrained algorithm. PyPI's library k-Means constrained receives three parameters: n_clusters, size_min and size_max, representing the number of clusters, and inferior and superior limits for cluster size, having these been fixed at 12, 40 and 50, respectively for scenario 1.

After aggregating containers into twelve clusters – from number zero to eleven – the number of containers and the total waste to collect per cluster, are presented in table X.

TABLE X

TOTAL WASTE TO COLLECT PER GENERATED CLUSTER ON SCENARIO 1

Cluster label	*Number of containers*	*Total waste to collect*	*<Q=150000*
0	50	125000	True
1	40	100000	True
2	40	100000	True
3	41	103000	True
4	40	100000	True
5	42	107500	True
6	40	100000	True
7	50	129500	True
8	40	100000	True
9	40	100000	True
10	50	125000	True
11	50	122740	True

The fourth column in table X stores boolean variables indicating if the total waste to collect on the corresponding cluster does not exceed the capacity of the vehicles. As denoted, every cluster respects the limit.

The next step is to pair clusters, matching them, maximizing the sum of the cluster centroid distances.

In this section, the programming language used was R (not Python as for the rest of the project), due to the existence of a package, nbpMatching, that provides functions to get optimal matchings on non-bipartite graphs. NbpMatching function nonbimatch() receives three arguments: a distance matrix, storing distances between cluster centroids on the case study, a threshold storing the maximum distance required to define matches and a precision value, indicating an upper bound for the digits on the largest value on the matrix. The NbpMatching function will then create a set of pairs of centroids minimizing the sum of distances between them.

As the purpose of this case study is to maximize centroid distances, centroid distances were multiplied by -1 to solve the problem with the minimizing algorithm. Table XI stores the output of the algorithm (i.e., routes to be collected on the same day) for scenario 1.

TABLE XI

ROUTE MATCHING PER DAY ON SCENARIO 1

Day	*Same day routes*
Monday	(0,3)
Tuesday	(1,5)
Wednesday	(2,11)
Thursday	(4,7)
Friday	(6,9)
Saturday	(8,10)

Lastly, the goal is to design collection routes – CVRP problem defined on section 6.3 with $p = 1$ vehicles per cluster – i.e., a TSP.

To solve the problem, the software used is the VRP Spreadsheet Solver (Erdoğan, 2017). VRP Spreadsheet Solver stores data in separate worksheets adopting an incremental flow of information. On a first step, the workbook only stores the worksheet named VRP Solver Console. Then, the remaining five worksheets, 1. Locations, 2. Distances, 3. Vehicles, 4. Solution, and 5. Visualization, are consecutively generated with a limit of 200 points. After updating the Locations sheet with the latitude and longitude of the containers and of the deposit, at sheet number 2 will be determined Euclidean distances (the software cannot determine real distances between two locations for Portuguese geography) between each two points previously defined. On worksheet number 3, information about vehicles is stored, such as the type and number of vehicles to consider on that solution, capacity and driving time limits, cost per unit distance and per trip, as well as distance limit per vehicle. Lastly, Solution sheet presents the output of the software, which is the route, with starting and ending point in the deposit, visiting each container once. All Visualization is developed through software QGIS 3.22.3. An example of a route developed for the case study is presented in Appendix 1, where 40 containers are collected. On the visualization, points 0 and 41 (both representing the depot – start and end of the trip) are not represented, in order to zoom into the containers.

8. Discussion

In this section, results from the case study presented above are discussed and comparisons between the three scenarios considered are made to evaluate the performance of having, or not, pre-determined clusters on the final total distances.

The use of different scenarios regarding number of clusters and routes per cluster was to possibly measure the performance of the methodology. One limitation of the VRP Spreadsheet Solver lies on the maximum number of points – 200 – to be included, which does not allow to create a scenario in which no cluster was initially determined, and all 523 containers would be added in the solver with 12 vehicles, creating the desired 12 routes. Thus the three scenarios were considered. In scenario 1, containers are grouped into 12 clusters, through KMeans constrained algorithm and having one route per cluster. Scenarios 2 and 3 were also considered, in which containers are split into 4 and 4x3 clusters, respectively. In scenario 2, each cluster would lead to three routes at once using VRP Solver with 3 vehicles. To guarantee no cluster would be over 200 containers and there would be a balance in between clusters, the minimum and maximum number of bins imposed through K-Means constrained were 170 and 180 respectively. For comparison purposes, in scenario 3, before assigning routes, three clusters were created within each of the four already existent clusters with one vehicle per run on the VRP Solver. This way, each cluster would lead to one single route, as before.

The total route distances, in kilometers, obtained for each scenario are displayed in table XII.

TABLE XII

TOTAL AND MODIFIED TOTAL DISTANCE PER SCENARIO

Scenario	*Scenario description*	*Total distance*	*Modified total distance*
1	12 clusters	802.417039	562.417039
2	4 clusters (3 routes per cluster)	764.138462	524.138462
3	4 clusters (within each cluster, 3 clusters)	796.068593	556.068593

Note that the waste sorting center is, on average, at around 10km of each cluster centroid. So, for percentual comparisons, it is removed 20km per route on each scenario (240km), returning modified total distances, as displayed above.

Scenario 2 displays the lowest total modified distance, with distances in scenarios 1 and 3 having increases of around 7.3% and 6.1%, respectively. The results translate that predetermining clusters using the K-Means constrained algorithm will lead to a higher total distance travelled when comparing to giving VRP Solver the freedom to determine both clusters and routes.

Overall, some limitations were faced during the project regarding the case study. First, the majority of the data is from the time of the pandemic, which could have affected comparisons in between seasons. These have been unprecedented years with impact on the regular behavior of many different aspects of society, including waste generation. A second limitation lies on the fact that employees observe and record the level of waste in containers, using no other technology, which surely influences data quality. Lastly, as mentioned during the report, distances in between containers are calculated using Euclidean distances. As it is known, these can vary largely from real travel distances.

In fact, to overcome data quality issues, Valorsul already has containers with integrated sensors to measure waste level (these being effective specially for glass since the waste growth in a container is linear). However, as the number of containers with sensors is rather small it was impossible to benefit from this technology in this project.

On a note for future developments, it is recommendable to Valorsul that studies are conducted to improve efficiency on both understating the frequency of collection required per container, and route planning to minimize travelling distances, avoiding collections on empty – or almost empty – containers frequently. It is important to note that, as described above, there are limitations to this project and that, for the purpose that it had, some assumptions that were made may need to be redefined on a more in-depth research. Nonetheless, it is, hopefully, that this work could inspire future research that will benefit company's performance.

9. Conclusion

The present research studied the problem of waste collection management, regarding both understanding the frequency of collection required per week, for a set of containers of given area, and designing routes to minimize total distance traveled by collection vehicles.

Actually, when focusing on route planning, three steps were taken. First, the use of a constrained K-Means algorithm allowed the containers to be aggregated into clusters. Then, a matching problem was solved to pair clusters to ensure that routes performed in the same day, by the two vehicles available, are sufficiently apart. Finally, a routing software was used in each cluster – if each cluster would correspond to a single route, the problem to solve on this last stage was a TSP, otherwise it is a CVRP.

The motivation for the project, was a case study with data from the municipality of Caldas da Rainha, for glass waste. On a data analysis base, all containers were considered to only require one collection per week. Note that these results are directly related to data quality which could have be compromised by two factors: all records are manually, by employees, introduced on the system and the majority of data available is from the pandemic.

Also, throughout the case study, it was possible to conclude that clustering containers a priori, and then solving a TSP per cluster, leads to worse results – i.e., higher value of total distance traveled – than not predetermining all clusters and letting the CVRP heuristic be partially responsible for both tasks on larger clusters.

10. References

[1] Abdallah, M., Adghim, M., Maraqa, M., & Aldahab, E. (2019). Simulation and optimization of dynamic waste collection routes. *Waste Management & Research, 37(8),* 793-802.

[2] Azadi, S., & Karimi-Jashni, A. (2016). Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: A case study of Fars province, Iran. *Waste Management, 48*, 14-23.

[3] Bijsterbosch, J., & Volgenant, A. (2010). Solving the Rectangular assignment problem and applications. *Annals of Operations Research, 181(1)*, 443-462.

[4] Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained k-Means clustering. *Microsoft Research, Redmond, 20(0)*, 0.

[5] De Marco, A., Mangano, G., & Zenezini, G. (2018). Classification and benchmark of City Logistics measures: An empirical analysis. *International Journal of Logistics Research and Applications, 21(1),* 1-19.

[6] Dündar, H., Ömürgönülşen, M., & Soysal, M. (2021). A review on sustainable urban vehicle routing. *Journal of Cleaner Production, 285*, 125444.

[7] Erdoğan, G. (2017). An open source Spreadsheet Solver for Vehicle Routing Problems. *Computers & operations research, 84*, 62-72.

[8] Ferreira, J.A., Figueiredo, M.C., Oliveira, J.A. (2017). Household Packaging Waste Management. In: , et al. Computational Science and Its 42 Applications – ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science(), vol 10405. Springer, Cham.

[9] Ghiani, G., Laganà, D., Manni, E., Musmanno, R., & Vigo, D. (2014). Operations research in solid waste management: A survey of strategic and tactical issues. *Computers & Operations Research, 44*, 22-32.

[10] Hailin, W., Tao, F., & Yang, B. (2020). Optimization of vehicle routing for waste collection and transportation*. International Journal of Environmental Research and Public Health, 17(14)*, 4963.

[11] Irnich, S., Toth, P., & Vigo, D. (2014). Chapter 1: The Family of Vehicle Routing Problems. In P. Toth, & D. Vigo, *Vehicle Routing Problems, Methods and Applications*, 1-23. Society for Industrial and Applied Mathematics.

[12] Kontokosta, C. E., Hong, B., Johnson, N. E., & Starobin, D. (2018). Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. *Computers, Environment and Urban Systems,* 70, 151-162.

[13] Lu, B., Greevy, R., Xu, X., & Beck, C. (2011). Optimal Nonbipartite Matching and its Statistical Applications*. The American Statistician, 65(1),* 21-30.

[14] Miller, C. E., Tucker, A. W., & Zemlin, R. A. (1960). Integer programming formulations of traveling salesman problems*. Journal of the ACM, 7,* 326-329. 43

[15] Simoni, M. D., Bujanovic, P., Boyles, S. D., & Kutanoglu, E. (2018). Urban consolidation solutions for parcel delivery considering location, fleet and route choice. *Case Studies on Transport Policy, 6(1)*, 112-124.

[16] Zdena, Z., Semiao, V., & Beijoco, F. (2013). Operation costs and pollutant emissions reduction by definition of new collection scheduling and optimization of MSW collection routes using GIS. The case study of Barreiro, Portugal*. Waste Management, 33(4)*, 793-806.

11. Appendix. Route Illustration

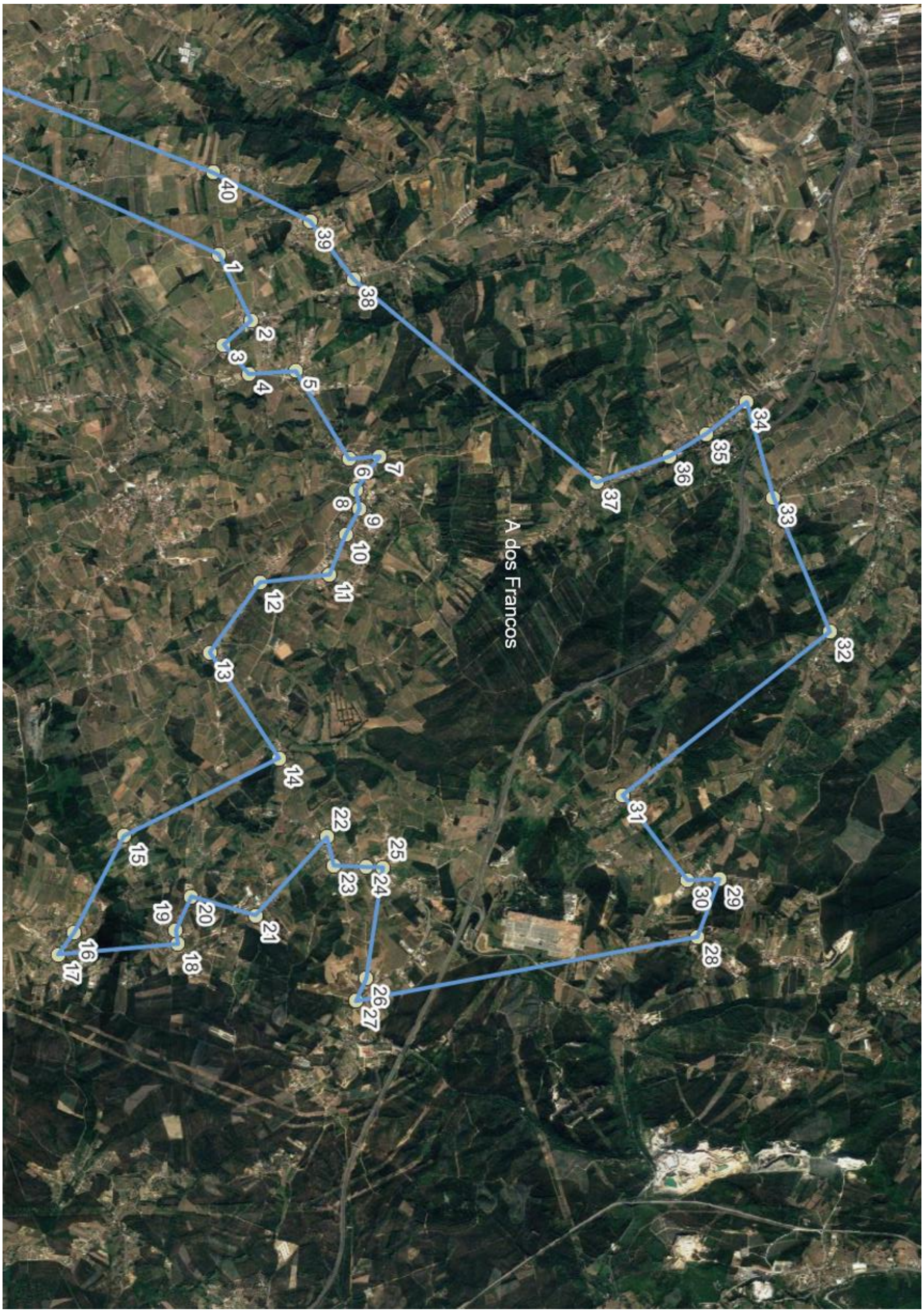

Figure 1 – Route illustration