Master Operations Research, Combinatorics and Optimization
Master Informatique / Master Mathématiques & Applications

Machine Learning in Waste Demand Forecast for Collection Optimization

Antonio López Ropero

August 25, 2023

Research project performed at laboratory G-SCOP

supervised by
Professor Van-Dat Cung

Defended before a jury composed of:
Mrs Nadia Brauner
Mr Pierre Lemaire

August 2023

Abstract

This work addresses the optimization of the waste collection process through the prediction of the filling of bins. The objective is to reduce the distances traveled by collection trucks, which will have a positive environmental and economic impact. To achieve this, Machine Learning techniques and collection optimization strategies were used to predict the filling rates of bins and allow more efficient truck routing in the collection planning.

As part of the methodology, a combination of techniques was implemented, including clustering, reverse inventory management and routing heuristics. The results obtained were assessed in comparison with data from previous years. This has revealed substantial improvements in terms of performance. The application of heuristics in routing was found to contribute significantly to the reduction of distances even not applying the forecasting methods, and the accuracy of the predictions was validated for most of the points of waste.

This study has shown that the integration of prediction and routing techniques is essential to have an more efficient waste collection process. Furthermore, it has been identified that it would be interesting to further investigate the waste collection dates as perspective among others. In summary, this work has provided a complete approach integrating waste prediction and collection planning to assess and improve the waste collection process.

Résumé

Ce travail aborde l'optimisation du processus de collecte des déchets à travers la prédiction du remplissage des poubelles. L'objectif est de réduire les distances parcourues par les camions de collecte, ce qui aura un impact environnemental et économique positif. Pour y parvenir, des techniques de Machine Learning et des stratégies d'optimisation de la collecte ont été utilisées pour prédire les taux de remplissage des bacs et permettre un routage plus efficace des camions dans la planification de la collecte.

Dans le cadre de la méthodologie, une combinaison de techniques a été mise en œuvre, notamment le clustering, la gestion inversée des stocks et les heuristiques de routage. Les résultats obtenus ont été évalués par rapport aux données des années précédentes. Cela a révélé des améliorations substantielles en termes de performances. Il a été constaté que l'application d'heuristiques dans le routage contribue de manière significative à la réduction des distances parcourues, même sans application des méthodes de prévision, et l'exactitude des prédictions a été validée pour la plupart des points de déchets.

Cette étude a montré que l'intégration des techniques de prédiction et de routage est essentielle pour disposer d'un processus de collecte des déchets plus efficace. En outre, il a été identifié qu'il serait intéressant d'étudier plus en profondeur les dates de collecte des déchets, entre autres, comme perspective. En résumé, ce travail a fourni une approche complète intégrant la prévision des déchets et la planification de la collecte pour évaluer et améliorer le processus de collecte des déchets.

Contents

— 1 —
Introduction to the Context of the Study

Since the mid-18th century, waste management has received increasing attention due to the rise of industrialisation and urban population growth. This is especially noticeable in urban areas where sanitation problems have aggravated, leading to major public health problems [6]. In the context of the European Union, different waste management strategies to minimise dioxin emissions have been carried out since they are highly harmful to public health. Today, with increasing pollution and scarcity of raw materials, waste management is more important than ever for what we hope will be a more circular and sustainable economy. Indeed, waste collection is the first step of a Circular Supply Chain [12].

This internship report aims to analyze the application of machine learning techniques in the context of waste demand forecasting to improve waste collection optimization process. The basic framework of this project consists of a set of points of waste locations in the Commaunté de Communes de Serre-Ponçon (CCSP) around the lake Serre-Ponçon where waste must be predicted and collected.

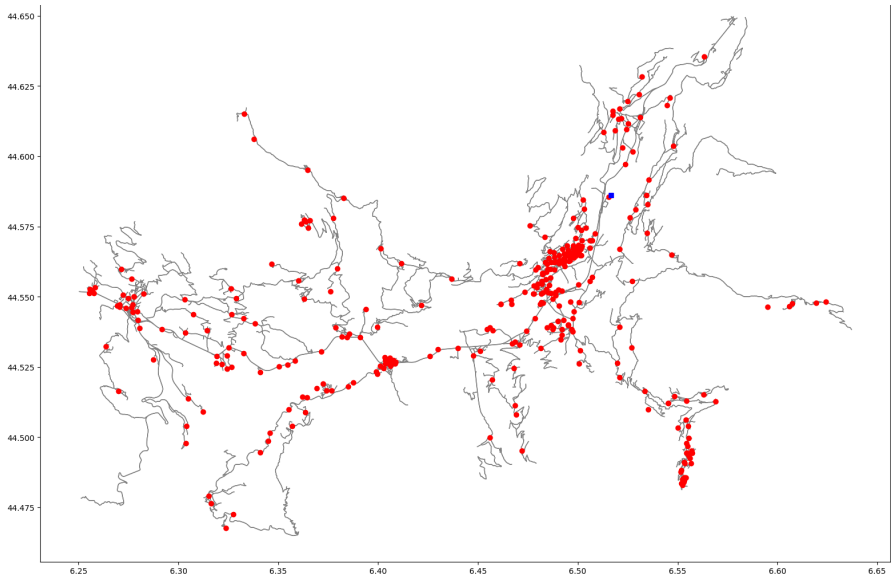


Figure 1.1: Points of waste locations in the Commaunté de Communes de Serre-Ponçon (CCSP)

We provide in section 1.1 an overview on the solid waste management process and its research importance. In section 1.2 we give the problem description, the physical system and

draw the limits of this project. Section 1.3 presents the current situation and solutions applied. Therefore, we will specify the research question and directions. Finally, section 1.4 sets up the objectives of this work.

1.1 The Solid Waste Management Process

To better understand the development of our work, we give an overview on the general solid waste management process.

Apparently the solid waste management process is an intuitive process, but in order to contrast information about it, three articles have been read to get an overview and to describe the process. The [7] is an article from 'WasteTech Engineering' which is an Australian company that designs solutions for sorting different materials using more effective and efficient techniques in order to increase recycling, reuse and reduce waste sent to landfill. The article [22] can be found on the website of the company 'Zaquin' headquartered in Malaysia, they have a fleet of 30 trucks and vehicles for waste collection, i.e. they focus on waste management. Finally, the article [10] is an article found in the Ministry of Environment and Sustainable Development of the Government of Argentina. All three articles focus on defining and describing the phases of integrated municipal solid waste management. After a review we proceed to define them into 5 phases (see Figure 1.2):

1. Waste generation: waste generation is due to residential, commercial or industrial practices. Depending on the origin, more of a certain type of waste will be produced, e.g. households will produce a high quantity of organic waste.

2. On-Site waste management: in this second phase, waste is stored in certain bins. In the past it used to be common to have one bin per block of flats or per house, but nowadays larger bins are often placed in specific locations. These bins store different types of waste (glass, organic waste, cardboard, etc.) and a group of houses or neighbourhoods deposit their waste there. The reality is that this type of waste storage facilitates the collection process since different wastes are stored in different bins, then there are less bins to manipulate and less points of waste collect.

3. Collection and Transport: this phase consists in collecting the waste in the containers or bins where it was stored in the previous phase. Collection can be of two types, general if the different types of waste are not discriminated or differentiated if the different types of waste are discriminated. In our project, we are in the case of differentiated collection.

4. Treatment: in this phase the collected waste is treated, whether general collection or differentiated collection was used. It consists of a second separation of the collected waste depending on the type of waste (organic waste, glass, cardboard, etc.) for recycling and subsequent use in new products.

5. Disposal: this is the final stage of the process, where waste that cannot be recycled for reuse will be stored in landfills or incinerated. Sanitary engineering methods are used at this stage to minimize public health risks, as there are liquid and gaseous emissions, among others, that can be harmful to health.

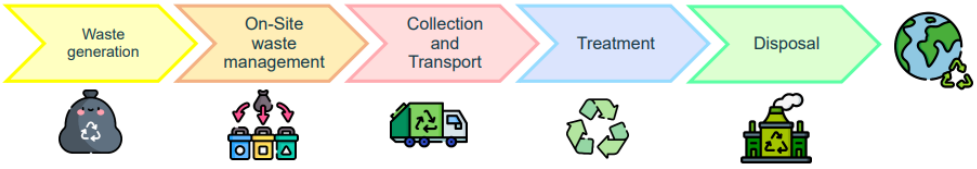


Figure 1.2: General solid waste management process

In the remaining of the report, we will mainly focus on phases 2 and 3. Regarding phase 2 we have to make predictions of the waste that will accumulate in the points of waste while, in phase 3, we have to develop some strategies for collecting and transporting the waste.

An optimized strategy in waste collection could lead to significant cost and resource savings: use of less fuel and reduction of CO2 or time spent by workers. But finding an optimized waste collection strategy is scientifically challenging since many parameters and constraints have to be taken into account (good demand forecast, routing optimization, traffic congestion, drivers schedule balancing, etc.)

1.2 The Physical System of CCSP

This section presents the existing physical system in the current project. This project examines the problem of waste collection, mainly domestic, in the *Commaunté de Communes de Serre-Ponçon (CCSP)* (see Figure 1.1), in the departments of Hautes-Alpes and Alpes-de-Haute-Provence, in the Provence-Alpes-Côte d'Azur region. The population of the commune is around 17000 inhabitants, but can triple during the high season, in summer around the lake Serre-Ponçon, and in winter at ski resorts such as Les Orres. In this geographical area of CCSP there is a set of 22 towns, in each of these there is a set of points of waste and for each of them there is a set of bins. Regarding waste accumulation behaviour, it may vary among the different points of waste as some are located in urban areas where more or less constant amounts of organic waste are produced, while other points located in ski resorts or campsites will depend on seasonality and tourism.

Each bin has the function to collect one specific type of different wastes, either organic, biflux (mix waste), cardboard or glass. From now on we will focus on organic waste since it is the one with the highest quantity of data available and it is of interest for the CCSP collection waste planners. Not necessarily in each point of waste there will be bins for all types of waste, and in fact, it is important to know that in some of these points we can have more than one bin for a specific type of waste, for example, some points have 2, 3 or even 6 bins to collect organic waste.

Regarding the capacity of the bins, we have to highlight there are different sizes: 1000, 3000, 4000 and 5000 litres. The last ones are the most prevalent, that is, most of the bins have this last capacity.

Therefore, we will talk about the existing resources available to collect waste. Since June 2022 we have a set of Excel files where we can find the routes carried out each day, the kilometres travelled and, specifically, in each of the routes, the points visited, the filling rates and the volume collected in each bin. According to the data provided by the company UNICO France who handles the waste management system of CCSP, they have 6 trucks for the collection of waste, 4 of which have a capacity of 14000 litres for organic waste and the other 2 have a ca-

pacity of 18000 litres for organic waste.

In addition to the waste collection points there is a special point called **depot**, from this point all the trucks depart and arrive, that is, the routes begin and end. Also, every time the truck reaches its capacity as it collects the waste, it goes to this point to empty its load and continue with the collection work.

1.3 Current Situation of the Problem and Research Question

In this section it will be described (1) the current solving and decision levels handled by the CCSP planners and (2) the research question raised at the tactical decision level.

Firstly, it will be explained how they are collecting each point of waste nowadays and, later, how the routing is being performed. The different periods considered will also be explained.

In order to decide which points of waste will be collected each day, the CCSP planners base their decision on their experience and their own historical data, thanks to which they have created a variable called *week frequency* that determines when to visit each point of waste. For example, if a point has frequency equal to 1, this means that it will be visited once per week and equal to 0.5 once every two weeks and so on for all points of waste.

The second decision is how to collect the points of waste, i.e. which routing strategy is used? For this, an operational level strategy is used, which means that each day the collection routes will be chosen according to the specific needs and constraints of that day. An example for a better understanding would be to imagine that if a specific waste collection point has not been collected today because a road is announced to be closed, then, it will be collected later on another day varying the initial route that had been set. According to the Excel file with distances and volumes collected, they usually collect a set of points per route, which will vary according to the needs that arise.

The general periods of time planned by CCSP are:

1. Summer period: 16th June - 31st September.
2. Autumn period: 1st September - 15th December.
3. Winter period: 16th December - 14th April.
4. Spring period: 15th April - 15th June.

Autumn and Spring periods are also called *off-season* periods. These periods are set according to different behaviours of the accumulated waste. The CCSP planners think that there are more gain opportunities with a better forecast of the filling rates during the off-season periods. This leads us to focus mainly our work on some off-season periods.

Consequently, the research question is: **Could we reduce the travelled distance by designing a waste collection strategy at the tactical level compared to their current solution in an off-season period?** In other words, we could also ask: **Could we increase the ratio of waste filling rate collected per kilometer run by the trucks?** This means whether or not we could collect more waste over the same distance.

Now, in an attempt to answer the research question, the key will be to break the problem down into two parts directly related to phases 2 and 3 of the section 1.1:

Waste demand forecast Prediction of the daily filling rates for each point of waste (prediction of the quantity of organic waste generated in each bin), for this part machine learning techniques such as time series and polynomial regression will be used.

Collection optimization Once the daily filling rates for all the points of waste are known we should optimize the collection and transportation. For this part, the techniques of clustering and vehicle routing will be used to decide which points should be gathered and create routes minimizing the overall distance traveled by trucks. Heuristics will be applied to find good solutions.

The work that we will develop is at the tactical level, that is, we focus on planning the collection of waste in a medium time frame for phases 2 and 3. Indeed, we make demand predictions and optimize routes for the next 2-3 months, or a specific seasonal period. Nevertheless, once tactical planning was created, some things could still be improved on a day-to-day operational level, as CCSP planners do today.

1.4 Main Objective and Methodology

The main objective of this research project is to analyze the application of Machine Learning techniques in waste demand forecasting and, subsequently, their impact on the waste collection optimization process. To achieve this general objective, we have adopted the methodology step-by-step as follows:

1. Reviewing the current State-of-the-Art Machine Learning techniques for demand prediction, and some problems beneath the waste collection optimization such as "reverse" inventory management, clustering and vehicle routing.

2. Developing some appropriated predictive models using Machine Learning algorithms to forecast waste demand.

3. Assessing the performance and accuracy of the predictive models developed and applied to the historical data.

4. Development of some planning strategies to determine the garbage collection dates at each point. These strategies are related to clustering and the "reverse" inventory problem and might depend on the filling rates pattern for each point.

5. Evaluation of current planning solution of CCSP and comparison with results got after the strategies developed to collect organic waste.

6. Identifying potential limitations and future research directions for waste collection optimization and Machine Learning in this field.

— 2 —

State of the Art in Waste Demand Forecast and Collection Optimization

As explained in the previous section, we focus on phases 2 and 3 of the general waste collection planning process in Figure 1.2 at tactical level, i.e. we will look for a strategy in a medium time frame for one of the off-season periods explained in section 1.3. Firstly, we need to predict filling rates in the different bins for each point (as mentioned above we focus on organic waste) and secondly, we have to design some waste collection planning strategy by making use of two problems: clustering and reverse inventory management. Therefore, relevant articles will be reviewed concerning the problems of these two phases in order to understand the existing information and studies up to the present day. But prior to the article review, a brief overview of studies on waste generation will be provided to give a broader understanding of the factors influencing its generation.

2.1 Influencing Factors and Seasonality in Organic Waste Generation

As indicated above, this section will provide a summary of articles to understand the main factors influencing the generation of waste, in particular organic waste, which is the main focus of this project. Organic waste includes food waste, gardening waste and other biodegradable materials.

Intuitively, there are several factors that can influence organic waste generation, such as population size and density. A larger population results in higher organic waste generation. Economic factors, lifestyle and activities in an area could also influence organic waste generation. For example, an area with a high concentration of restaurants and food establishments would theoretically produce a higher amount of waste.

As mentioned in the introduction, the Serre-Ponçon lake area has locations that are frequented or populated only during certain seasons or months of the year because of tourism. Therefore, some articles will also be looked for to explore the influence of seasonality on the generation of organic waste.

The article [3] presents the findings of a study conducted in a suburban area in Sri Lanka to understand solid waste generation and its composition. It was found that the amount of waste generated is related to the population and the standard of living of the people. Additionally, other factors such as climate, lifestyle habits, educational level, and cultural beliefs were iden-

tified to influence the quantity and composition of the waste, with population being the most significant factor. The study revealed that waste generation also varies based on these socioeconomic factors. In addition, the amount of waste was analyzed and found to differ according to seasonality and socio-economic patterns. These results demonstrate the importance of taking these factors into account.

Regarding the most produced waste, in the municipality of Moratuwa, the average amounts of waste generated per capita per day were: 374g of organic waste, 18.9g of paper, 14.1g of plastic, 66.9g of glass, and 29.7g of metal. Biodegradable organic components are the largest category (approximately 90%) in the household waste generation. It is therefore confirmed that organic waste is the most significant type of waste produced, and our project will focus on data related to this type of waste.

Still, it was observed a seasonality pattern in terms of amount. This could provide a clue to understand that different amounts of organic waste are generated depending on the season of the year. Furthermore, it was observed as well that the most important factor in organic waste generation is the population and, in this project, there are bins where the population varies depending on the season of the year due to the impact of tourism, such as campsites or ski resorts.

According to [5] tourism is often seen as a promising sector for growth and development, but it also brings with it some negative consequences. This paper aims to examine the environmental costs of tourism in terms of solid waste generation efficiency in Tuscan municipalities, with a focus on the seasonality of tourism as presented in our project. The study fills a gap in the literature by exploring how seasonality due to tourist presence affects waste generation and, furthermore, the results confirm the significant impact of tourism on waste generation and collection. The paper also reveals that the seasonality of tourism greatly amplifies the impact of tourism on waste management and hinders solid waste management at an optimal level. The study provides evidence on the mechanism relating the presence of tourists to the efficiency of solid waste management, identifying a strong negative effect of seasonality of tourism and short stays. The implications of these results are clear: more effort is needed to flexibly manage solid waste collection, as predictions of waste generation are more difficult.

In conclusion, the study reveals a significant impact of tourism on waste generation and, consequently, on waste management. Therefore, this article contributes to our understanding of how seasonality will affect certain points of waste in our project, presenting challenges especially in the prediction of waste generation and, subsequently, in its collection.

In summary, the article [3] has made the following key contributions: firstly, organic waste is the most commonly generated waste and will therefore be the subject of the experiments in this project as more historical records are available for analysis; secondly, population size is identified as the most influential factor and finally it is introduced that seasonality influences the generation of organic waste. The article [5] has provided a comprehensive insight into how tourism and seasonality affect the generation of waste and how it complicates predictions and collection.

This fulfils the first objective of the state of the art which consists in knowing in general terms the main factor of waste generation, the population, deducing that a greater amount of waste will be generated in urban areas than in rural areas. Likewise, for seasonal areas like ski resorts or campsites, the impact of seasonality and tourism was explained, which is necessary for the analysis of the available historical data and the behaviour of the different points of waste.

2.2 Prediction of Waste Filling Rates

One of the main challenging objectives in organic waste generation is to accurately predict the filling rate at which organic waste is generated and the filling rates that will be generated in the future. As discussed in the previous state of the art, waste generation is a complex process influenced by various factors such as population, seasonality and others. Accurate prediction of waste generation is essential to ensure efficient waste collection practices.

In these last years, the development and application of advanced data analysis techniques, such as Machine Learning (linear regression, time series, neural networks, etc.), have considerably improved the accuracy of waste generation predictions. By analyzing historical data on waste generation, these techniques can provide valuable predictions for the future. Furthermore, depending on the available databases, Machine Learning models can be improved to further improve predictions with other variables, such as economics or demographics. Accurate prediction of waste generation is highly beneficial, as it subsequently will allow waste collection to be optimized, avoiding excessive bin emptying frequency and preventing bin overflows at the same time.

The following is a review of a set of articles in which predictions of the waste generated are performed.

In the article [1], a literature review of 88 articles is carried out. These articles deal with the challenge of solid waste management and its predictions. The aim of the papers is to predict solid waste generation and its influencing factors using different mathematical models and techniques. According to the authors, waste collection presents a significant challenge, and the lack of information and historical data on waste generation complicates the task, as discussed previously. The approach adopted is to estimate waste generation through socio-economic factors by reviewing existing literature. Most studies use these variables and analyze effects through linear regression, with recurring factors such as GDP, population, income, household size, energy consumption and water consumption. Identifying and understanding these factors can improve the prediction of waste generation and facilitate more efficient waste collection although our aim is to make predictions with historical records and not with other factors as it is done in these articles.

Here, several modeling techniques are used in solid waste generation studies. Researchers often use different methods, which can be classified into several classes, such as regression models (used in 37 published studies), Machine Learning (e.g. neural networks are the most common, but require a larger amount of data for model training), time series analysis and others, such as geographically weighted regression, which was used in 7 country-level studies.

In summary, the choice of modeling techniques depends on the available data and the goals of the study. Researchers typically apply different techniques to complete and compare results.

This literature review provides a comprehensive overview of techniques used in this type of problems and the variables that have the most effect on waste production (population, GDP, etc.). Since in this project we are interested to exploit the data from the historical waste collections, a review of articles will be carried out on the waste production forecast with historical data.

In article [11] a short-term prediction model for waste generation in New York City is presented. By integrating historical waste collection data with several variables, a Gradient Boosting Regression Tree model was developed, which achieved an average accuracy of 88% in predicting the weekly generation of three types of waste: solid waste, paper recycling and

waste petroleum gas management. The study concluded that the regularity of weekly waste generation predictions allowed a good prediction accuracy in the model. Furthermore, the model proved robust when incorporating additional external features, in particular weather conditions. The collaboration with the New York City Department of Sanitation was aimed at improving operational efficiency and long-term planning. The forecasting capability provided by the model allows for optimisation of waste collection, vehicle allocation and the possible development of specific waste reduction strategies and recycling strategies. This research could be used by other cities to optimise their waste collection in the future. The work in [11] provides a first contribution predicting the waste generation using historical data instead of socio-economic factors like the literature review carried out in [1]. This work also provides us a set of techniques used in some other studies such as correlation analysis, multiple regression analysis, time-series analysis, etc.

At this point, it is confirmed that there are hundreds of articles related to waste generation using socio-economic variables, such as GDP, and even literature reviews like [1]. However, to the extent of our knowledge, the work presented in [11] is the only article found regarding the prediction of waste generation using historical data after an extensive literature review.

Let's explore another article on linear regression, this time focusing more on understanding how to measure the accuracy of the trained model, which will be useful for our project. The article [9] deals with the need to develop a predictive model to estimate the amount of waste generated in kg (dependent variable) from organics, plastics and other components of the waste. The results present a comparative analysis of two statistical techniques: a statistical regression technique and an averaging technique, for modeling organic waste generation. The accuracy of both techniques is evaluated using R^2, $RMSE$, MSE, $MAPE$ and the t-test (these indicators are measured on both training and test sets). The results indicate that the statistical regression technique shows slightly better accuracy than the averaging technique in terms of R^2 and $RMSE$. Furthermore, the t-test reveals a significant difference between the two techniques, with the statistical regression technique demonstrating superior accuracy.

In summary, this study contributes to the expanding body of literature in waste generation by comparing two statistical techniques and providing insights into the measures employed for evaluating model accuracy.

Continuing on techniques for predicting waste generation, the article [18] highlights the importance of accurately predicting solid waste generation to facilitate efficient waste collection. It points out that most waste prediction models are based on demographic and socio-economic factors, as exemplified in the aforementioned article [9]. Therefore, this paper proposes a prediction technique based on non-linear dynamics, comparing its performance with a seasonal autoregressive moving average methodology (SARIMA) for short- and medium-term forecasting. The models are applied to five real data sets, the first of which comprises residue data from Valencia (Spain) from October 1982 to December 1999, with 207 observations. Both SARIMA and non-linear prediction techniques show promising results in terms of prediction accuracy of solid waste generation. These techniques allow predictions to be made with a low relative error, which facilitates the optimisation of resources and the medium-term planning of solid waste collection.

Let's go deeper into the investigation of articles that employ time series analysis to forecast waste generation. In the article [17], time series analysis and forecasting techniques such

as ARMA and ARIMA models were used to examine solid waste generation in seven states in Malaysia. Quarterly data were used to predict future waste quantities and evaluate model performance, as the country faces a waste storage problem.

The ARMA and ARIMA models showed to be effective in forecasting solid waste generation in the seven analyzed states. Model evaluation was conducted using metrics such as Mean Squared Error (*MSE*), Akaike Information Criterion (*AIC*), and Bayesian Information Criterion (*BIC*), which are also used in linear regression analysis.

We can conclude that this study highlights the key role of ARMA and ARIMA models in the analysis of solid waste generation in Malaysia. The results obtained from these models and evaluation using statistical measures provided valuable information for more effective waste management planning and prediction.

Another paper on the application of time series to waste generation is [16]. It presents a detailed time series analysis conducted to predict solid waste generation in the city of Arusha, Tanzania. A total of 66 monthly data points, obtained from records of the amount of solid waste collected by the municipal authorities, covering a five-year period from July 2008 to December 2013, were used. Of these data points, 60 were used for model training, while the remaining 6 were reserved for model testing.

After applying several models, it was determined that the ARIMA(1, 1, 1) model produced the best results in terms of accuracy and precision in predicting the amount of solid waste generated. These results are of great relevance for effective solid waste planning and management in the city of Arusha, as they provide useful information on trends and patterns of waste generation, allowing municipal authorities to make informed decisions and design effective strategies for a proper waste collection. The ARIMA(1, 1, 1) model demonstrated good performance on several accuracy evaluation metrics, such as Mean Absolute Error (*MAE*) or Root Mean Square Error (*RMSE*). These evaluation metrics indicated the accuracy and precision of the model in predicting the amount of solid waste generated, which supports its capability as a forecasting tool for waste generation in the city of Arusha.

In the annexes, we can find table 5.1 where it is shown a summary of these contributions, models applied, accuracy of the models, etc. Looking at the table we can see that in all the articles found in the literature, socio-economic variables are used except for the paper [11]. In this later one, in addition to other factors (economic activity, demographics, land use or climate) the historical data were used to train a model which is based on decision trees to create a more accurate model.

Now, regarding to the articles [18], [17] and [16], the application of ARIMA and SARIMA models have shown promising results in predicting solid waste generation. However, it is important to notice that these studies differ from our approach as they focus on estimating the amount of waste generated in a city (in kilograms or tones) at a strategic level, while our goal consists in determining the filling rate (%) of specific waste collection points in the Serre-Ponçon Lake area at a tactical level. Additionally, in the mentioned articles, the data are spaced at a regular time step, unlike our current situation where our data are paced by the collection dates. Therefore, it is crucial to explore and gather information on proper data preparation techniques to apply these models effectively in our specific context.

Since the data from CCSP in our project are not paced regularly in time, we looked for papers to have a deeper understanding of time series data, on what they should look like in

terms of characteristics and model features, and what are the possible solutions when the data do not satisfy some conditions.

The article [8] aims to go deeper into the field of time series data mining, covering aspects such as definition, main tasks, representation techniques, distance measures, etc. The first thing you read is the definition of a time series, which is defined as an ordered sequence of n variables. They are usually the result of observing a process in which data values are collected at evenly spaced time intervals. This implies that the time between data points should always be the same. However, in our data set this condition is not satisfied. The article also discusses data representation (how to decompose a time series into other components), similarity measurement (how to check whether two time series behave in a similar pattern) and the indexing method (techniques used to organise time series for more complex data sets) as the main aspects that characterise time series. This article contributes to the idea that our data need to be spaced at the same time interval, and so we need to explore as well some interpolation techniques.

The article [13] provides a detailed review of interpolation methods for filling the missing gaps in time series data, evaluating efficiency criteria and quantifying uncertainties. It highlights the need to take prediction uncertainties into account and discusses the lack of their estimation in interpolated/extrapolated data. The article suggests new lines of research and introduces a new interpolation method. Let's examine some of the interpolation methods described to deal with this issue in our data. Deterministic interpolation methods perform a basic estimation between the last known point before the lag and the first known point after the lag. They are based on a deterministic relationship and fill in missing values using linear or nonlinear interpolation schemes (without taking into account probabilistic or stochastic factors). Among these methods, the nearest neighbour interpolation method assigns the value of the nearest known neighbour to the missing value. Polynomial interpolation, on the other hand, looks for a straight line between two known points, while polynomial interpolation methods use polynomial functions to fit the data more accurately. Spline methods, such as B-splines, are common in interpolation. Other methods include distance weighting or Fourier-based methods.

Stochastic methods take probability into account and include regression methods such as linear regression (which can be used to estimate unknown values) and polynomial regression (this technique should be used carefully since we can lead into overfitted models). Autoregressive methods use previous values in the data series to predict future values, assuming that a missing value can be calculated only from previous values. For example, an AR(2) model uses a linear combination of the two previous values and an error term to predict the next value. Advanced methods related to Machine Learning, such as Artificial Neural Networks, can also be found.

The article concludes that no exhaustive comparative studies have been published so far, which can be problematic for researchers, engineers and professionals who need to choose interpolation methods for their purposes and data. Each interpolation method has its limitations and assumptions, so it is important to carefully consider the context and characteristics of the data before selecting a specific method. An incorrect choice of method or failure to take uncertainties into account may lead to inaccurate or unreliable results in data analysis and future predictions.

This article will be essential for the time series in this project, as the data are not always equally spaced in time interval. Applying time series models to datasets with irregularly spaced data can present additional challenges and considerations. The absence of constant time interval data can introduce irregularities in the data and make temporal analysis and modeling difficult.

While predictions can still be made, the exact timing may be uncertain, as the assumption of constant time interval is not satisfied.

In our dataset, for each point where we want to predict waste generation, we have a set of observations with a time interval that is not always regular. These records contain a variable indicating the percentage of waste collected at a particular point and date. In addition, there is a variable that accumulates this amount assuming that there is no waste collection, so it is an increasing function. Therefore, the objective variable of the prediction is the filling rate, without taking into account the accumulation. However, if the models demonstrate superior performance in terms of MSE, the accumulated filling rate can also be used.

In summary, thanks to this article, a specific interpolation method for time series will be developed in section 3.3, and the polynomial regression technique will also be applied in section 3.4. We also remark that we will not apply neural networks, as they require a larger amount of data than we currently have available at most points of waste.

2.3 Waste Collection: Clustering, Reverse Inventory and Routing Problems

In this section a literature review on problems related to the waste collection problem in the second part of this project will be carried out. It will be subdivided into different sub-problems: clustering points of waste, managing "reverse" inventory to decide when to collect the points of waste and optimizing the routing of the trucks.

2.3.1 Clustering Points of Waste

Once the predictions at each of the point of waste are obtained, we will have to start with the collection optimization and, for this, the first part consists in performing a clustering, so we will carry out a research on this method for our planning strategy at the tactical level.

In the article [14] clustering is defined as an unsupervised learning technique used to find structures in unlabelled datasets. The aim is to group similar objects into clusters based on certain characteristics. Most important among different types of clustering are:

- Hierarchical: it can be agglomerative (bottom-up) or divisive (top-down) and uses distance measures to merge or split clusters based on their similarity.
- Density-based: there algorithms use the notion of density to cluster nearby data points that are surrounded by areas of low density or noise. Some examples are DBSCAN or SSN.
- Partition-based: they consist in fixing the number of clusters k desired and iteratively reallocate the elements until convergence is reached. Some examples are k-means or k-medoids.

Determining the optimal number of clusters is a challenge in clustering. Common techniques include setting a distance threshold, analyzing the dendrogram structure in hierarchical

clustering, and using information criteria or external validation methods. There is no universally accepted formula for determining the number of clusters, and the choice depends on the context and objectives of the analysis.

On the other hand, some methods require as input the number of clusters k and this is a challenge. For this there are some techniques such as distance threshold, information criteria or external validation methods but the most correct strategy is to review the context and understand well the objectives of the problem before choosing any integer k.

It should be noted that obviously clustering does not have an unique solution, it depends on the method chosen and the decisions made by each person in each possible problem. Regarding the analysis of the results, care must be taken to understand the advantages and limitations of the chosen method.

This article provides some basic notions into the types of clusters and the problem of determining the number of clusters k. Then, a literature review more specifically on clustering in waste collection and transport process will be carried out.

For this, the article [21] on municipal waste collection and transport is presented. It explains that currently the total area of their project is divided into geographical regions marked by districts to facilitate organisation and operation, each district has a treatment plant and, as a result of capacity restrictions and inflexibility with vehicles, the current model is not very efficient.

To improve this efficiency it is proposed to do away with the current splitting (removing the district constraints) and to cluster the bins using the improved hierarchical agglomerative clustering algorithm (IHAC) which takes into account the distances between bins and the operational costs of cluster merging. This improved algorithm and collection strategy showed better results than the original hierarchical agglomerative clustering algorithm (HAC).

This article provides an initial idea about a possible clustering that could be used in this project.

Now, a research paper [2] presents a bibliographical review on Cluster Analysis in waste management. It helps to identify the most commonly used techniques available so far. Sixty-one articles were analyzed and classified into nine categories, although the relevant category for us is "waste collection process" which focuses on waste collection routes.
Reviewing this category we can see that there are several techniques applied: 27 heuristics and metaheuristics, 4 machine learning based and 3 linear programming based. The predominance of heuristics is expected due to the complexity of the problem and the high computational cost induced. It is also explained that the clustering that occurs most frequently is k-means, although hierarchical clustering also appears.

Finally, a suggestion is made in the article: explore techniques in reverse logistics applications. The reverse logistics application is defined as the flow of products from the point of consumption to the point of collection or recycling (the opposite direction of traditional logistics).

Therefore, this article has provided an overview of the studies conducted so far on waste management, and especially on waste collection. Furthermore, it proposes further research on reverse logistics for better waste management and also it is remarked that a lot of studies are theoretical and most practical and real studies should be carried out as it is our case. That is why we propose section 2.3.2, but first let's do a review about how to balance and cluster a set of points, i.e. how to take into account the distances and, the weight in each point of waste (quantity of waste generated).

Before we could obtain k geographical clusters but they will not be balanced in terms of amount of waste generated in each one of them. In other words, an attempt will be made to improve the geographic clustering of the 2.3.1 section by balancing waste collection generation with the goal of creating future collection routes that are similar in terms of distances and/or time. This is related to the Bin Packing Problem (BPP) which according to the article [19] is a combinatorial optimization problem in which n objects of different sizes $(w_1, w_2, ..., w_n)$ that have to be packed into bins of capacity C. The objective function will try to minimize the total number of bins used by meeting the following constraints: all objects will be packed in some bin and the capacity C of the bins will not be exceeded. The BPP has applications in several areas such as: logistics, transportation and storage system design. The formulation of the BPP can be found in the paper although its main objective is to create a flow arc-based model with a heuristic algorithm for solving it. This problem is NP-hard, which implies that finding an optimal solution for large instances is computationally expensive and, therefore, heuristics that provide near-optimal solutions in reasonable times will be used. The relationship of the BPP to our problem is:

BPP	**Our problem**
The objective is minimizing number of bins used	For us, the number of used bins (in this project bins are clusters) is fixed. The objective is minimizing the difference between the bin with the maximum weight and the minimum weight
Pack all the items with weight w_i	Pack all the items with weight w_i
Capacity C of the bins	There is no capacity for the bins

Table 2.1: Relation between BPP and our problem

These BPP model similitude will help to adapt a model according to our necessities in section 4.

2.3.2 Reverse Inventory Management for Collections

In this section we will review reverse logistics, this refers to the flow of products from the points of waste in our project to the depot. But inside the reverse logistics process, we need to manage carefully the inventory level of the waste bins. The objective is to create a planning strategy that allows us to know on which dates we should collect each bin taking into account that:

- Not every day a route will be carried out for each of the clusters, i.e. each cluster will have some specific dates of collection.
- Not all the points of waste of the same cluster will be collected, some conditions will have to be satisfied to collect each point of waste.

The article [20] deals with reverse logistics which is defined as the process of collecting and managing waste once it has reached the end of its useful life. The study proposes to approach solid waste collection and logistics by defining it as an inventory control problem with stochastic demand, the demand in our problem would be equivalent to the organic waste generated,

and it is stochastic because the demand generated can vary unpredictably, this demand is the first part of the project, the predictive models of organic waste generation.

In this article we gain understanding of reverse logistics and also in heuristic which defines some strategies to minimise waste collection distances. Now, let's have a look on the *reverse* inventory problem to define later our strategy for the collection.

In the book [15] we find the definition of reorder point. It is a notion of inventory management in which, going below a certain inventory level (defined in units, percentage, etc.) the stock will be replenished, this level is fixed. Referring to our problem it could be a good idea to use it in a *reverse* way, i.e. before reaching a level (filling rate) in each point of waste, the collection of this waste should be done to avoid overflowing of the bins at that point. In addition, this level could vary depending on the point to be collected and depending on how good KPI's (R^2, MSE, etc.) were obtained in the model prediction. In [15] the reorder interval is also defined as the time that passes between two stock replenishments, in our case it is the time that passes between two collections. It should be noted that since the demand varies and is not constant every day this interval will vary, but it will not be a problem in this part since the daily filling predictions are generated for each of the points.

In our case, once the dates of possible collection (taken as assumption) have been fixed, it will be decided when to visit each point of waste.
Once decided the dates on which each point of waste will be visited and the demand generated in terms of organic waste are known, the routing problem of the section 2.3.3 will be carried out.

2.3.3 Capacitated Vehicle Routing Problem and Heuristics

According to the article [4] the Capacitated Vehicle Routing Problem ($CVRP$) is one of the fundamental problems with a set of applications in transports or logistics. Like our case, we need to collect the waste from a set of points of waste satisfying: each route begins and ends at the depot, each point of waste is visited exactly once and, the waste generated per point of each route does not exceed the capacity of the trucks. Thus, the CVRP is suitable to model our collection routing problem.

The first formulation was proposed by Dantzig and Ramserc in 1959 and, five years later Clarke and Wright proposed the first heuristic to solve this problem since it is NP-hard and it still has no algorithm found so far to solve it optimally with big instances.

We can find both the original formulation of the $CVRP$ and the formulation based on the idea of combining individual routes and optimising them to form feasible and near-optimal solutions. Initially, each customer is considered as a separate route, and then the algorithm combines these routes based on vehicle capacity and distance between customers. Route combination is performed using a savings approach, which identifies pairs of customers that, when combined, generate a significant reduction in the total distance travelled. In this way, the algorithm constructs more efficient route to achieve solutions closer to the global optimum. In conclusion this Clarke and Wright heuristic will be applied to solve our problem.

— 3 —
Waste Demand Forecast

3.1 Data Processing: Consolidation and Analysis

In this document, we present a detailed description of the procedures used to consolidate the data for experiments with mathematical models. Data consolidation is a crucial stage in scientific research, as incorrect or incomplete data can significantly impact the accuracy and validity of the obtained results, especially when using mathematical models for predictions.

3.1.1 Introduction to the Provided Data

Regarding the data provided by CCSP for predicting waste generation at each point, there are primarily three Excel documents available.

- The first Excel document contains the filling rates at different points from January 2020 to March 2023. The columns included in this Excel are as follows:
    1. Date: records the date of organic waste collection.
    2. Commune: municipality or town where the waste point is located.
    3. Points OM: identifies the organic waste points.
    4. Taux: percentage of collected waste.
    5. Débordement: indicates the percentage of overflow at the waste point, if any.
- The second Excel document contains the following important fields:
    1. Identifiant: unique identifier for each waste collection point.
    2. Adresse: exact address of the waste point.
    3. Commune: municipality where the waste point is located.
    4. Code_postal: postal code corresponding to the waste point.
    5. Latitude: indicates the latitude coordinate of the waste point.
    6. Longitude: indicates the longitude coordinate of the waste point.
- The third Excel document contains the following columns:

1. Nom: format XXX-YY, where XXX corresponds to the unique identifier of the waste point and YY represents the number of bins since there can be more than one bin at a single waste point.
2. Adresse: exact address of the point of waste.
3. Commune: municipality where the waste point is located.
4. Type: type of container, with two different types.
5. Flux: type of waste. In our results, we will focus on the 'OM' waste since it is the most abundant, as reviewed in the State-of-the-Art, and therefore, we have more data available.
6. Volume: volume in liters of the points of waste capacity.

3.1.2 Data Consolidation Process

The objective of this process is to create a consolidated Excel document that contains all the historical records from the first Excel file. In addition, the following steps will be taken for data integration:

- The variables "Adresse," "Code_postal," "Latitude," and "Longitude" from the second Excel document will be added. To merge the data, the "Points OM" column from the first Excel document must match the "Identifiant" column in the second Excel document.
- The variables "Type," "Flux," and "Volume" from the third Excel document will also be included. To merge the data, the "Points OM" column from the first Excel document must match the "Nom" column (part XXX) in the third Excel document.

By performing these integration steps, a single Excel document will be created, encompassing all the available information. The integrated dataset will provide a comprehensive view of waste generation, including the historical records as well as additional relevant variables such as address, postal code, geographical coordinates, container type, waste type, and volume. This consolidated dataset will serve as a valuable resource for subsequent data analysis and modeling, enabling a more comprehensive understanding of waste generation patterns and facilitating the development of accurate predictive models.

3.1.3 Exploratory Analysis of the Consolidated Data

Exploratory analysis of the consolidated data is crucial to gain an initial understanding of the data and identify any patterns or trends. It can also help to detect potential errors or inconsistencies in the data which can then be addressed in the data cleaning process.

A summary of the exploratory analysis will be explained below.

1. Variable Identification: Each column in the dataset was examined to identify the variables present, including their data types (numeric, categorical, etc.).
    a) Datatime64[ns]: Date.
    b) Object: Commune, Points OM, Adresse, Type.
    c) Float64: Taux, Débordement, Commentaires

d) Int64: Volume (L), code_postal, latitude, longitude

2. Count of registers and columns: Each column in the dataset was examined to identify the variables present, including their data types (numeric, categorical, etc.).

3. Count of registers and columns: there are 12 variables and 29508 registers.

4. Unique values per variable: "Date" has a range of 1185 days (more than 3 years of historical data), 22 "Commune" and 364 different "Points OM".

5. Relation between "Taux" and "Débordement": during the analysis, it was observed that whenever a non-null value exists in the "Débordement" column, the corresponding "Taux" value is 1. Therefore, to accurately represent the actual collection rate in such cases, the "Taux" value is updated by adding the overflow rate. For instance, if "Débordement" is 0.2, then "Taux" is adjusted to 1.2, reflecting the true collection rate in that particular record. "Taux">1.4 are outliers values and they are removed (only 25 registers).

3.1.4 Adding New Columns to Historical Data

These addings will facilitate the data handling in the waste demand prediction and collection programs.

1. Nom1: This column is used to predict the waste generation at each point, regardless of whether it has one or more bins. Without this column, separate predictions would be made for each combination of point and bin (e.g., Point_OM = 1-1, Point_OM = 1-2). By introducing the Nom1 column, a single prediction is made for each unique point, taking into account the number of bins.
   Also, the column 'Volume' indicating the capacity in litres of each bin will be merged if there are more than 1 bin (e.g., Point_OM = 1-1 had a capacity of 5000 and Point_OM = 1-2 too, then Nom1 = 1 has a capacity of 10000 litres.
   Now, there are 274 different points of waste (some of them with more than 1 bin for organic waste) to predict and collect. "Nom1" with less than 10 historical registers are removed since it is not possible training any model.

2. Taux_AC (% accumulated): This column accumulates the waste generated for each value of Nom1, assuming that no collection takes place and waste simply accumulates over time. The Taux_AC value for each record is obtained by summing the Taux values from previous records with the same Nom1 value. This allows for the consideration of waste accumulation in the prediction models.

Taux	Taux_AC	Nom1
0.4	0.4	224
0.2	0.6	224
0.3	0.9	224

Table 3.1: Construction of Taux and Taux_AC columns

3. Number of bins: there may be more than one bins at certain collection points. To account for this, a column indicating the number of bins at each waste collection point is included.

4. Date_numerical: To convert the date variable into a numerical format, a reference date of December 31, 2019, is chosen. For example, if the date is January 5, 2020, the corresponding value for Date_numerical would be 5. This transformation allows the use of the date variable in numerical calculations.

3.1.5 Some Remarks on the Consolidated Data

There are 274 different waste points for which the filling rates need to be predicted. Each of these points has a varying number of historical records, as observed during the exploratory data analysis (ranging from 10 to 888 records). These consolidated data are now used in three different prediction techniques: linear regression, time series and polynomial regression.

3.2 Linear Regression Prediction

Algorithm 1 Pseudocode Linear Regression (LR)

Input: Database for all points of waste 'Nom1' with 'Date_numerical' as dependent variable X and 'Taux_AC' as independent variable y.
Output: model of linear regression $y = ax + b$ where a and b are stimated and we are able to make predictions $\hat{y}_i$ for any day.

for point = 1,...,$|Nom1|$ **do**
  Filter database by $|Nom1|$. // Select the registers for each 'Nom1'
  Split_list = [70, 90, 1] // From 70% to 90% from 1% to 1% (possible divisions of train and test sets)
  Best_MSE $\leftarrow +\infty$; Best_split $\leftarrow$ 70%
  for split = 1,..., $|Split\_list|$ **do**
    X_train $\leftarrow$ [Date_numerical] [:split] // First split % to train
    y_train $\leftarrow$ [Taux_AC] [:split] // First split% to train
    X_test $\leftarrow$ [Date_numerical] [split:] // Last split % to test the model
    y_test $\leftarrow$ [Taux_AC] [split:] // Last split % to test the model
    lr = LinearRegression.fit() // Train the model and get predictions
    $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ // Compute MSE
    if $MSE < Best\_MSE$ **then**
      $Best\_MSE \leftarrow MSE$; $Best\_split \leftarrow split$
    end if
  end for
  X_train $\leftarrow$ [Date_numerical] [:Best_split] // Choose best split %
  y_train $\leftarrow$ [Taux_AC] [:Best_split]
  X_test $\leftarrow$ [Date_numerical] [Best_split:]
  y_test $\leftarrow$ [Taux_AC] [Best_split:]
  lr = LinearRegression.fit() // Train model with best_split and obtain predictions $\hat{y}_i$
  $MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
  $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ // Measured in test (validation) test
end for

A Python program was developed (based on Algorithm 1) to perform Linear Regression (LR) technique using historical data for each point of waste ('Nom1'). The program uses the **sklearn library** and incorporates various functions from it. For instance, r2_score, mean_squared_error and mean_absolute_percentage_error functions are used to compute R^2, *MAPE* and *MSE* respectively. During this section it will be shown the pseudocode for linear regression models, show some results and explain the different behaviours of the points.

Program Results

After executing the Python code and once we have got the R^2, *MSE*, and *MAPE* for each point of waste 'Nom1', a set of eight points are selected to be analyzed, considering their main characteristics in the historical data, the model fit and the accuracy measure (see Table 3.2). The selection of these points is based on the obtained results, taking good and poor outcomes, points with more limited historical data, and those with extensive historical data as can be observed in Table 3.3.

Point	#Registers	#Bins	%Train	Daily % filling	R^2 test	*MSE*	*MAPE*(%)
1	152	2	89	11.26	0.98	0.40	0.43
8	38	1	90	1.40	0.00	0.75	4.96
9	204	3	72	25.00	0.90	46.71	2.26
51	34	1	70	1.97	0.30	3.67	9.46
129	74	1	72	3.95	0.99	0.10	0.67
179	59	1	85	3.46	0.96	0.10	0.70
224	748	1	90	29.50	0.00	424.31	6.31
267	235	4	80	24.19	0.00	1774.58	10.99

Table 3.2: LR results in the chosen 8 points of waste

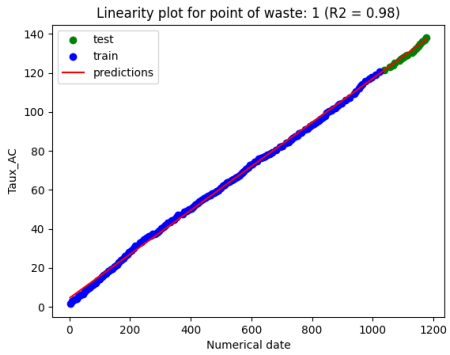
	Good accuracy	Bad accuracy
More than 150 observations	1 and 9	224 and 267
Less than 150 observations	129 and 179	8 and 51

Table 3.3: Chosen points of waste 'Nom1'

Observation: R^2 could result negative when measured in the validation set (like in the previous table), in this case it is truncated to 0 in the results.
According with R^2 points 1, 9, 129 and 179 have a good result since its value is close to 1.
According with R^2 points 8, 51, 224 and 267 have a bad result since its value is close to 0.

Plot Results

In the following figures we can see in blue the points of the train set, in green those of the test set and in red the fitted line for each linear regression model.



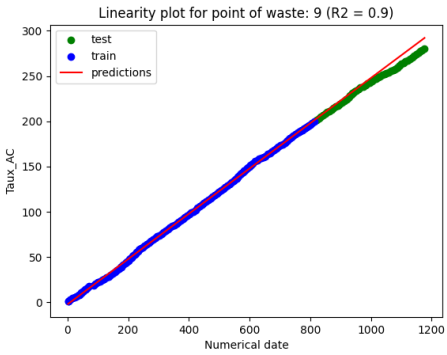
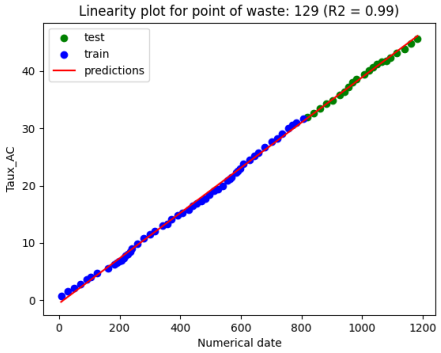


Figure 3.1: Chosen points with good results (Part 1)



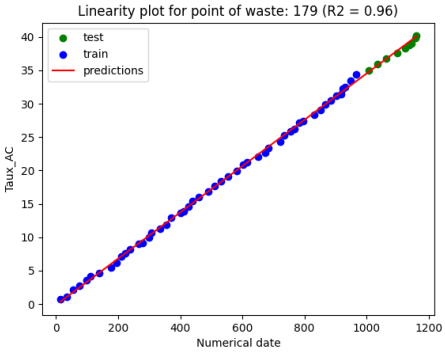
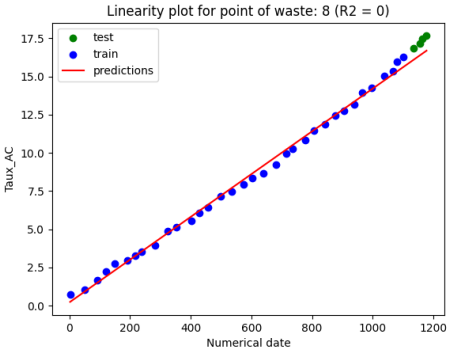


Figure 3.2: Chosen points with good results (Part 2)



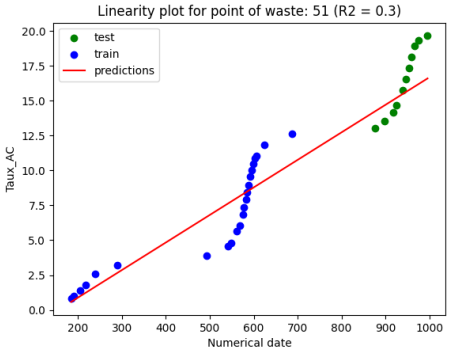
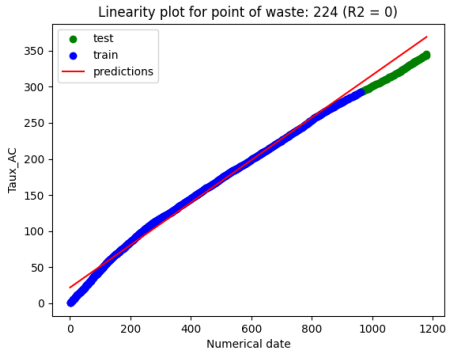


Figure 3.3: Chosen points with bad results (Part 1)



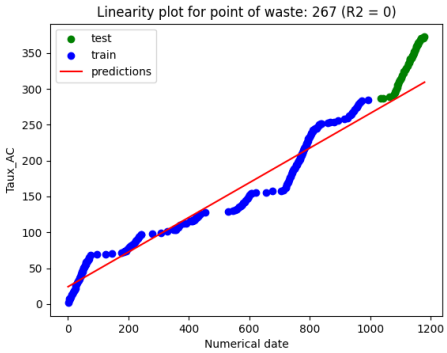


Figure 3.4: Chosen points with bad results (Part 2)

Behaviour Across Different Points

In general, it can be observed that the points with good results are characterized by being easily represented by a line, that is, the accumulation of waste is relatively constant. On the other hand, when the linear model does not fit properly, the behavior of the points becomes more complex. For instance, in Figure 3.3, point 51 can be observed, where the model performs poorly due to highly seasonal behavior. Specifically, the waste generation in this point is highly dependent on the time of year, with a significant increase during the tourist season, as it represents the waste from a camping site. Point 224 represents a hospital, and apparently, the quantity of waste is decreasing as can be seen in test set (green points). Therefore, when the model is trained using the data from the training set, it overestimates the waste generation. This could be attributed to a decreasing trend in waste accumulation.
Point 267 also exhibits seasonal behavior and is located in the municipality of Les Orres, a mountainous area. Again, this can be attributed to seasonal fluctuations due to tourism, specially in winter.
Point 8, on the other hand, seems to underestimate the predictions. Upon examining the plot of historical records, a milder seasonality can be observed. There may also be an increasing trend in waste accumulation, leading to a poor model fit.

Types of behaviours
- **Points with no seasonality:** in general they do not fit good to LR models.
- **Points with seasonality:** in general they fit good to LR models.

We should notice that there may be waste points such as point 224, the hospital, where there is a decreasing or increasing trend for some reason. At these points seasonality is not considered to exist, but neither linear regression models fit good.

General Results and Bloxplot

Let's analyze some statistics regarding the results obtained in the set of the 274 points.

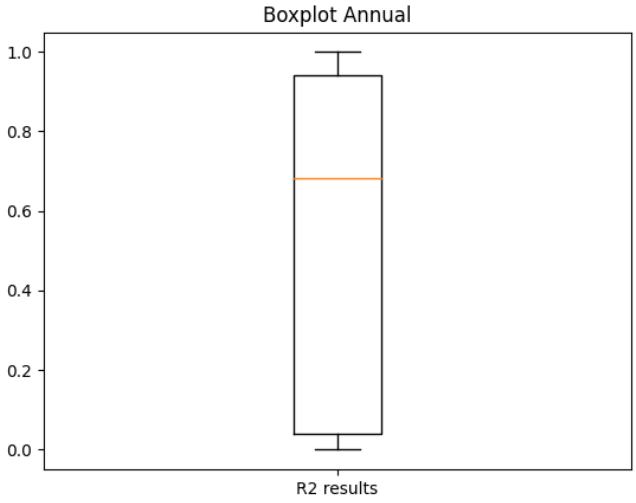


Figure 3.5: Boxplot R^2 results (minimizing MSE of the models)

In the boxplot with R^2 results measure in the test set (Figure 3.5) we can see that the results present a high variability, which means that we have some points with no seasonality and another ones with seasonality. In addition, the median of the results is around 0.7, meaning that there are more or less the same number of points with $R^2 \geq 0.7$ and $R^2 \leq 0.7$ (Table 3.4).

# Points s.t. $R^2 \geq 0.7$	136 points $\longrightarrow$ 49.6%
# Points s.t. $R^2 < 0.7$	138 points $\longrightarrow$ 50.3%

Table 3.4: General R^2 results **minimizing** *MSE*

General results (minimizing *MSE*) $\begin{cases} \textbf{Average } R^2 \textbf{ for all the points: } 0.57 \\ \textbf{Average } R^2 \textbf{ for points s.t. } R^2 \geq 0.7\textbf{: } 0.91 \end{cases}$

Observation: in addition of the results minimizing the *MSE* we also got the results maximizing the R^2 (Table 3.5).

# Points s.t. $R^2 \geq 0.7$	187 points $\longrightarrow$ 68.2%
# Points s.t. $R^2 < 0.7$	87 points $\longrightarrow$ 31.8%

Table 3.5: General R^2 results **maximizing** R^2

General results (maximizing R^2) $\begin{cases} \textbf{Average } R^2 \textbf{ for all the points: } 0.75 \\ \textbf{Average } R^2 \textbf{ for points s.t. } R^2 \geq 0.7\textbf{: } 0.93 \end{cases}$

Map with Bad and Good Models in Linear Regression

The following map show the good and bad models in the whole area from the lake Serre-Ponçon when minimizing *MSE*.

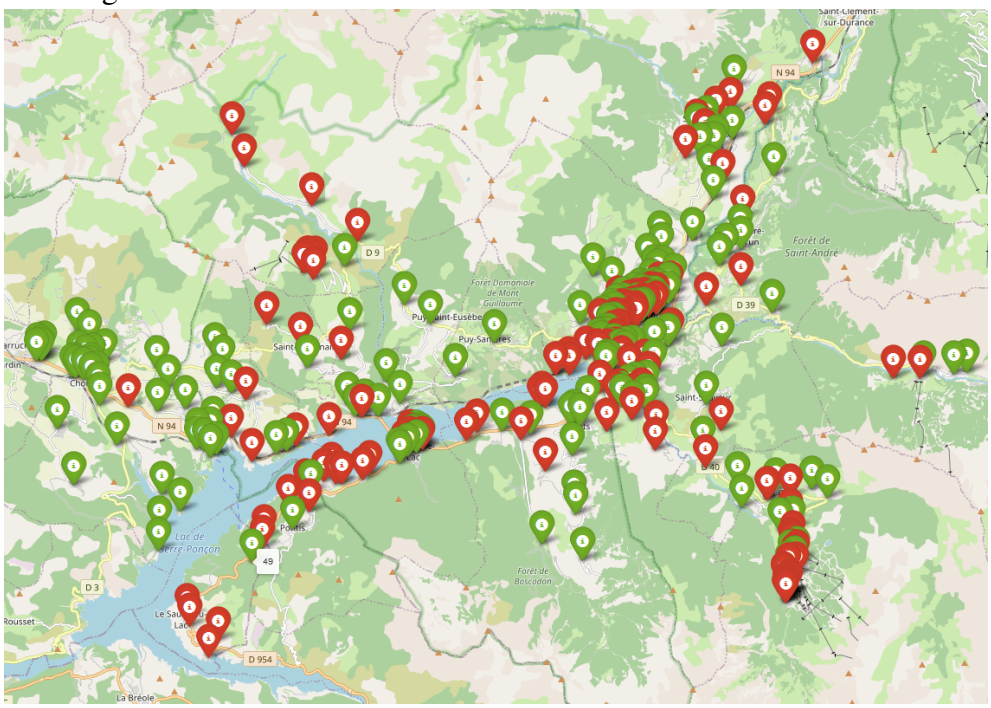
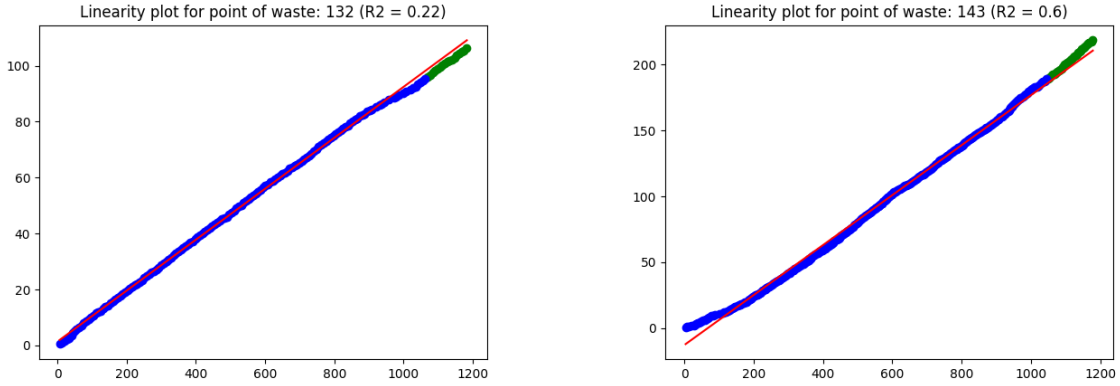


Figure 3.6: Map of Points with $R^2 \geq 0.7$ (green points) and $R^2 < 0.7$ (red points)

In the previous map, it can be observed that the red points (poor models using linear regression) are concentrated in certain areas. In fact, the generated map was inspected to identify and understand which points yield worse predictions, and as expected, these points are found in more seasonal areas (ski resorts in the southeast of the map, campgrounds near the lake area, etc.). Furthermore, some points in the densely populated area (Embrun city) also exhibit poor performance in the linear regression model. Therefore, it was decided to examine points 132 and 143 within the city to observe their trend (Figure 3.7).



(a) Point 132

(b) Point 143

Figure 3.7: Poor models example at Embrun city (not touristic place)

It seems that the models are providing a low R^2 due to a decreasing trend in organic waste at point 132 and an increasing trend in organic waste at point 143.

Predictions and Daily Filling Rates

This section is a consequence of working with the variable 'Taux_AC', remember that this variable accumulates the waste, taking the assumption that it is never collected as seen in the example of the Table 3.3. The linear regression models have the formula $y = ax + b$ where the coefficients a and b are estimated and, therefore, for any day x ('Numerical_date') we can obtain the accumulated waste y ('Taux_AC') and, hence, by calculating the difference between the accumulated waste of two days in a row, we can compute the daily filling rate as we can see in Table 3.6.

Numerical_date	Taux_AC	Taux	Nom1
1888	60.51	-	2
1889	61.23	$61.23 - 60.51 = 0.72$	2
1890	62.01	$62.01 - 61.23 = 0.78$	2

Table 3.6: How to compute daily filling rates

We can conclude that with these models we have the daily filling rates for all the points of waste in order to continue with the second part of this project on the waste collection.
Observation: since we have the daily filling rate $Taux(d)$ and the capacity for each point of waste denoted as $V(p)$, then, the filled volume $FV(d,p)$ for any day d in litres is also known and computed like follows:

$$FV(d,p) = Taux(d) \cdot V(p) \tag{3.1}$$

3.3 Time Series Prediction

3.3.1 Data Issue: Interpolation Method

As we discussed in the section 2.2, it is crucial to have data records with the same time interval. This is because Time Series (TS) models rely on the assumption of a constant and regular time interval between observations. Having a consistent frequency ensures that patterns, trends, and seasonality can be accurately captured and analyzed. Therefore, maintaining a uniform frequency in time series data is essential or otherwise we would have to find an alternative to use data not satisfying this main assumption in time series.

This is what happens with our dataset, there exist some points that, in general, are collected every week, but sometimes it is done every 5 or 10 days and therefore the solution to this problem is the **interpolation**. In Tables 3.7 and 3.8 on data point 224, we can understand a specific interpolation method which was developed for this specific problem and context.

Date	Taux	Taux_AC
1/4/2023	0.3	0.3
2/4/2023	0.4	0.7
3/4/2023	0.3	1
4/4/2023	0.6	1.6
5/4/2023	-	-
6/4/2023	0.6	2.2
7/4/2023	0.7	2.9

Table 3.7: Point 224 (no interpolation)

Date	Taux	Taux_AC
1/4/2023	0.3	0.3
2/4/2023	0.4	0.7
3/4/2023	0.3	1
4/4/2023	0.6	1.6
5/4/2023	0.3	1.9
6/4/2023	0.3	2.2
7/4/2023	0.7	2.9

Table 3.8: Point 224 (with interpolation)

It is a basic interpolation method where, as can be easily seen in the example, the available data point is divided equally among the missing days and the preceding days that had no record. It was expected to be applied on the input set for all the points of waste, but it was found to be computationally very expensive due to the high number of new records to be introduced. Therefore, the alternative was to train the models with the data for 2021 and 2022 and make predictions for the full year 2023. Therefore, the following assumption is made: *The predictions obtained in 2023 are on the same days on which there was collection in 2022 (historical register).*

3.3.2 Prediction by Time Series

We have developed a Python program that applies the *SARIMA* time series model to each point of waste 'Nom1', again using the **sklearn library** and its functions. During this section it will be shown the pseudocode for time series models and results will be shown and analyzed.

Algorithm 2 Pseudocode Time Series (TS)

Input: Database (years 2021, 2022) for all points of waste 'Nom1' with 'Date_numerical' as dependent variable X and 'Taux_AC' as independent variable y.
Output: *SARIMA* model of time series with predictions for 2023.

for point = 1,...,$|Nom1|$ **do**
  Filter database by $|Nom1|$. // Select the registers for each 'Nom1'
  Split_list = [70, 90, 1] // From 70% to 90% from 1% to 1% (possible divisions of train and test sets)
  Best_MSE $\leftarrow +\infty$; Best_split $\leftarrow$ 70%
  for split = 1,..., $|Split\_list|$ **do**
    X_train $\leftarrow$ [Date_numerical] [:split] // First split % to train
    y_train $\leftarrow$ [Taux_AC] [:split] // First split% to train
    X_test $\leftarrow$ [Date_numerical] [split:] // Last split % to test the model
    y_test $\leftarrow$ [Taux_AC] [split:] // Last split % to test the model
    ts = SARIMA.fit() // Train the model and get predictions for 2023
    $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ // Compute MSE, measured in the first three months of 2023
    if $MSE < Best\_MSE$ **then**
      $Best\_MSE \leftarrow MSE$; $Best\_split \leftarrow split$
    end if
  end for
  X_train $\leftarrow$ [Date_numerical] [:Best_split] // Choose best split %
  y_train $\leftarrow$ [Taux_AC] [:Best_split]
  X_test $\leftarrow$ [Date_numerical] [Best_split:]
  y_test $\leftarrow$ [Taux_AC] [Best_split:]
  ts = SARIMA.fit() // Train model with best_split and obtain predictions $\hat{y}_i$
  $MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
end for

Program Results

The program has been executed, *MSE* and *MAPE* values were computed for each point of waste. Let's analyze and compare the results from the same 8 points from LR in Table 3.2:

Point	#Registers	%train	R^2 test	*MSE*	*MAPE*(%)	%train	*MSE*	*MAPE*(%)
1	152	89	0.98	0.40	4.22	85	0.49	0.40
8	38	90	0.00	0.75	5.25	85	128.17	15.26
9	204	72	0.90	46.71	11.32	90	1.38	0.36
51	34	70	0.30	3.67	14.23	89	5.36	16.07
129	74	72	0.99	0.10	12.21	71	0.08	0.56
179	59	85	0.96	0.10	5.29	89	0.09	1.60
224	748	90	0.00	424.31	7.50	88	0.49	0.19
267	235	80	0.00	1774.58	13.56	79	10.28	3.25

Table 3.9: Results for 8 points (Linear regression: blue, Time series: orange)

In Table 3.9, it can be observed that, in general, by looking at the *MSE*, the results for

points of waste 9, 129, 179, 224, and 267 have improved but not in points 1, 8 and 51. Point 1 gives a similar *MSE* results, but points 8 and 51 have a bigger difference. Here we noticed that the reason could be the low number of historical data (38 and 34 respectively), later a deeper analysis on the number of data needed will be provided, first we will have a look to the plots obtained with these models.

Plot Results

In Figures 3.8 to 3.11 we can see in blue the training data, and, in orange, the predictions for 2023.

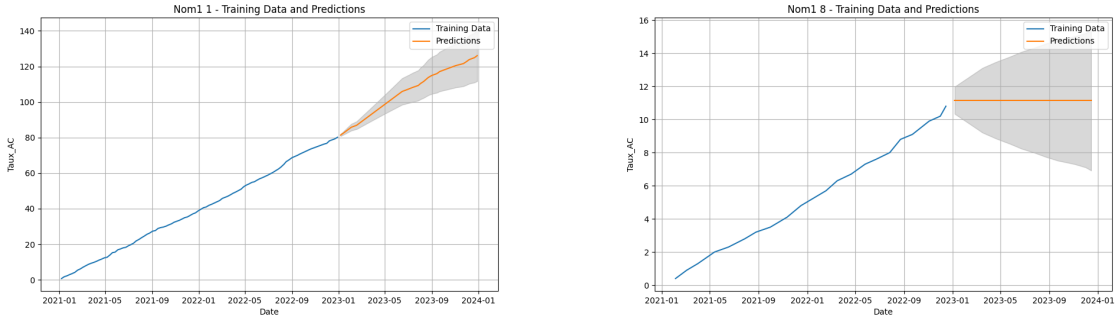


Figure 3.8: TS models (part 1)

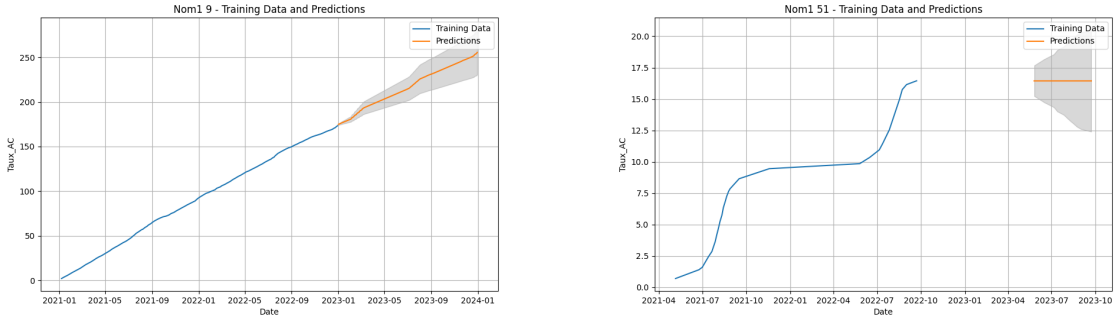


Figure 3.9: TS models (part 2)

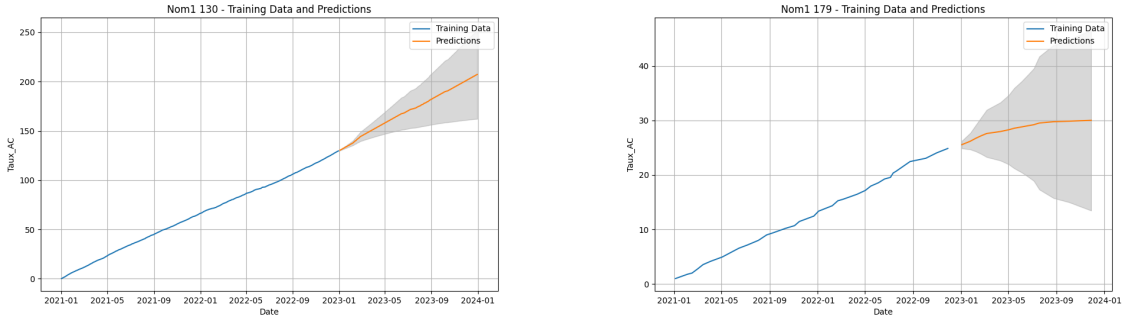


Figure 3.10: TS models (part 3)

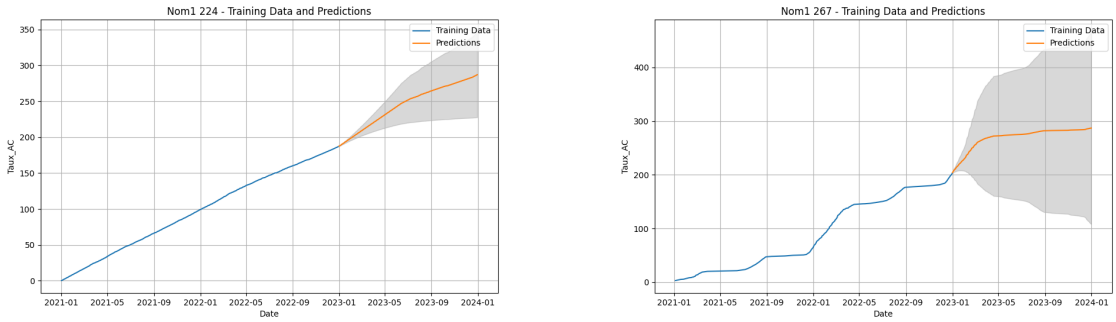


Figure 3.11: TS models (part 4)

It is important to notice that the predictions of models 8 and 51 are constant, and that is the reason why they have a significantly high MSE, so an analysis will be done to see how many points of waste provide such models and try to find a relationship with the number of historical data.

Points of Waste with Bad Predictions

As we saw in Figures 3.8, 3.9, 3.10 and 3.11 points 8 and 51 exhibit poor predictions, and the cause seems to be that they have the least amount of data among the 8 selected points: 36 and 34, respectively. Using interpolation would involve filling a significant number of missing values, which could lead to inaccurate predictions.

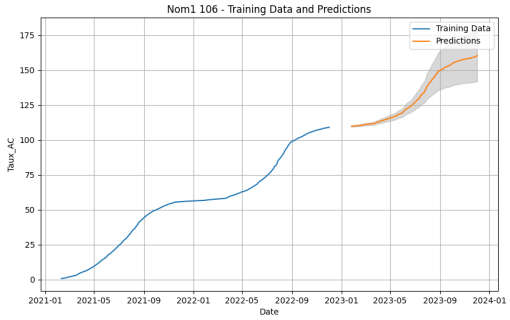
A thorough review of the prediction plots reveals that approximately 18% of the time series models do not yield satisfactory results, while the remaining 82% do. In the points where satisfactory results are not achieved, a common characteristic is observed, they all have fewer than 200 observations.

< 50 registers $\longrightarrow$ 21 points (55% bad models and predictions)
$50 - 100$ registers $\longrightarrow$ 23 points (45% bad models and predictions)
$100 - 150$ registers $\longrightarrow$ 4 points (7.4% bad models and predictions)
> 150 registers $\longrightarrow$ 2 points (1.8% bad models and predictions)

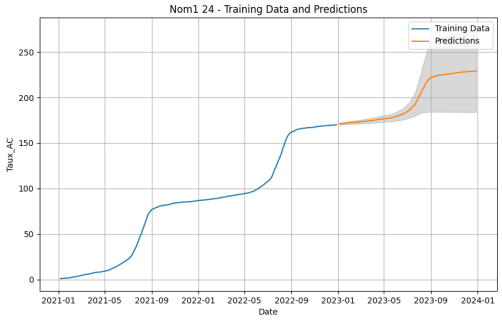
Table 3.10: Time series bad models (number of registers)

As we can see in Table 3.10, the more historical records one has to train a model, the less likely one is to get poorly trained models with constant predictions. It is not possible to generalize a minimum number of observations to train a time series model, as it will depend on how complex the model is.

The predictions obtained with the time series models will be shown for points 24 and 106 in Figure 3.12, where the number of points has been sufficient to train the model. These points of waste have been chosen because they show a clear seasonal trend.



(a) Point 106, 258 observations



(b) Point 24, 213 observations

Figure 3.12: Time series for seasonal points of waste

General Results and Comparison with Linear Regression

In general, the results obtained from time series analysis are better than those obtained from linear regressions (see Table 3.11), especially when the data points exhibit high seasonality. Time series models are capable of fitting the data more effectively than linear regressions. However, there are also cases where linear regressions produce better models than time series analysis when comparing the Mean Squared Error (MSE). In some situations, time series can give worse results than linear regressions due to the presence of complex and non-linear patterns in the data. Time series often involve seasonal components, trends and autoregressive behaviour, which can make them difficult to model and predict. In addition, linear regressions may work well when the relationship between variables is approximately linear, while time series may exhibit non-linear changes over time. The choice of the appropriate model depends on the nature of the data and the ability of the model to capture the specific characteristics of the time series.

LR better than TS	110 points $\longrightarrow$ 40,1%
TS better than LR	164 points $\longrightarrow$ 59.9%

Table 3.11: Comparison of LR and TS using MSE

Daily Filling Rates and Strategy to Collect the Waste

Again, the models have been trained with the variable "Taux_AC" because they showed a better accuracy than using "Taux". Regarding the interpretation of the results, we recall the assumption made with the time series models "the predictions obtained in 2023 are on the same days on which there was collection in 2022 (historical registers). Once we have obtained our predictions for 2023, we will apply the interpolation method of the table, and then we will obtain a daily prediction for each day and point of waste of 2023 as needed for the second problem of the project: the waste collection optimization process.

Observation: the filled volume $FV(d, p)$ for any day and point of waste is known and measured in litres.

3.4 Polynomial Regression Prediction

As we saw in the State-of-the-Art, interpolation could have negative effects when there are too many missing values, and the possibility of using Polynomial Regression (PR) was suggested. Thus it was decided to apply this technique for polynomials of degree 1, 2 and 3, the polynomials of degree 1 being linear regressions with the same solutions as in the section 3.2. Again, a Python program is developed using sklearn library.

Algorithm 3 Pseudocode Polynomial Regression (PR)

Input: Database for all points of waste 'Nom1' with 'Date_numerical' as dependent variable X and 'Taux_AC' as independent variable y.

Output: model of polynomial regression of degree 1 ($y = ax + b$), 2 ($y = ax^2 + bx + c$) or 3 ($y = ax^3 + bx^2 + cx + d$) where a, b, c and d are estimated to get predictions $\hat{y}_i$ for any day.

for point = 1,...,$|Nom1|$ **do**
  Filter database by $|Nom1|$. // Select the registers for each 'Nom1'
  Split_list = [70, 90, 1] // From 70% to 90% from 1% to 1% (possible divisions of train and test sets)
  Best_MSE $\leftarrow +\infty$; Best_degree $\leftarrow 1$; Best_split $\leftarrow 70\%$
  for split = 1,..., $|Split\_list|$ **do**
    Degrees = [1, 2, 3] // Possible degrees
    X_train $\leftarrow$ [Date_numerical] [:split] // First split % to train
    y_train $\leftarrow$ [Taux_AC] [:split] // First split% to train
    X_test $\leftarrow$ [Date_numerical] [split:] // Last split % to test the model
    y_test $\leftarrow$ [Taux_AC] [split:] // Last split % to test the model
    for degree = 1,..., $|Degrees|$ **do**
      pr = PolynomialRegression.fit(degree) // Train the model and get predictions
      $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ // Compute MSE
      if $MSE < Best\_MSE$ **then**
        $Best\_MSE \leftarrow MSE$; $Best\_degree \leftarrow degree$; $Best\_split \leftarrow split$
      end if
    end for
    X_train $\leftarrow$ [Date_numerical] [:Best_split] // Choose best split %
    y_train $\leftarrow$ [Taux_AC] [:Best_split]
    X_test $\leftarrow$ [Date_numerical] [Best_split:]
    y_test $\leftarrow$ [Taux_AC] [Best_split:]
    pr = PolynomialRegression.fit(Best_degree) // Train model with the best degree and obtain predictions $\hat{y}_i$
    $MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
    $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
  end for
end for

Program Results

The program is executed and R^2, *MSE* and *MAPE* values are obtained for each point of waste. Let's see in Table 3.12 the results of polynomial regression on the points selected earlier.

Point	#Registers	#Bins	%Train	Best polynomial degree	R^2 test	*MSE*	*MAPE*(%)
1	152	2	89	1	0.98	0.40	0.43
8	38	1	90	3	0.96	0.00	0.25
9	204	3	87	3	0.99	1	0.32
51	34	1	71	2	0.53	2.11	7.88
129	74	1	90	3	0.98	0.04	0.41
179	59	1	85	1	0.96	0.10	0.69
224	748	1	90	2	0.58	96.70	2.88
267	235	4	84	3	0.78	91.43	2.44

Table 3.12: Polynomial Regression results in 8 points

According with R^2 points 1, 8, 9, 129, 179 and 267 have a good result since its value is close to 1. Point 224 (hospital) and 51 provide a better model with a polynomial regression of degree 2 and 3 respectively, but still we get a significant amount of unexplained variability.

Plot Results

In Figures 3.13 to 3.16 we can see in blue the points of the train set, in green those of the test set and in red the fitted polynomial regression for each point:

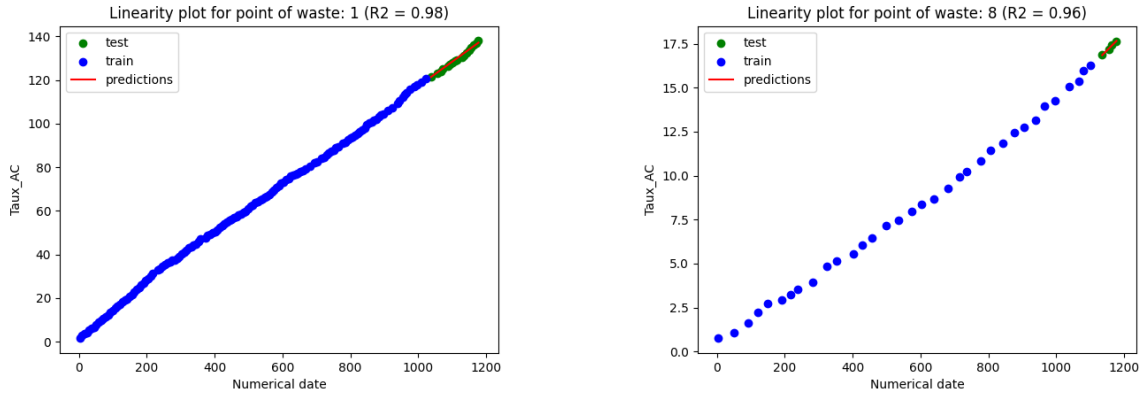


Figure 3.13: PR models (part 1)

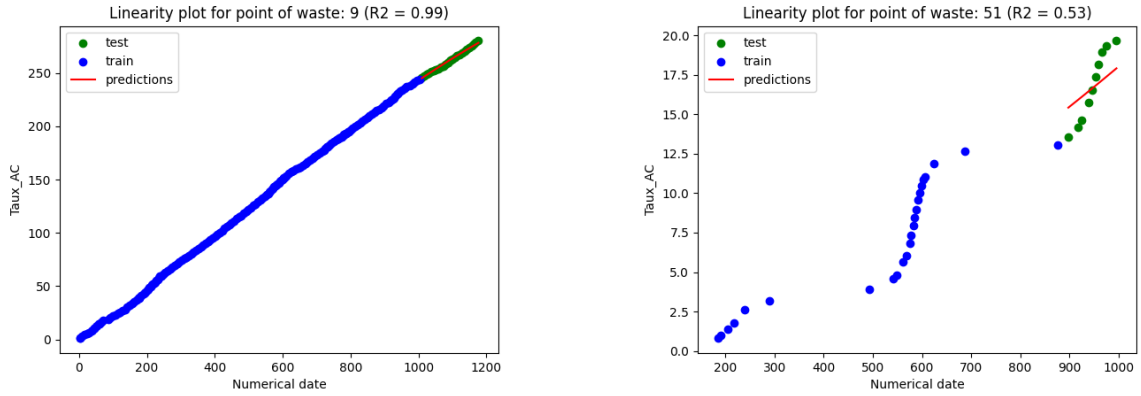


Figure 3.14: PR models (part 2)

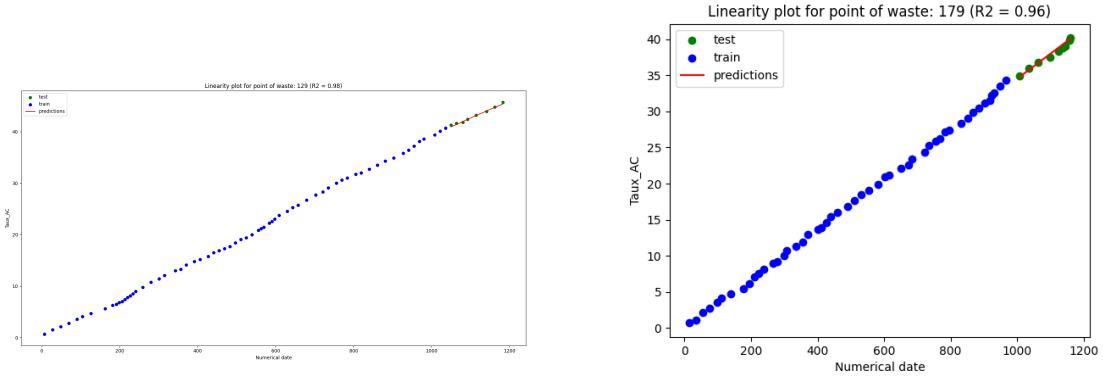


Figure 3.15: PR models (part 3)

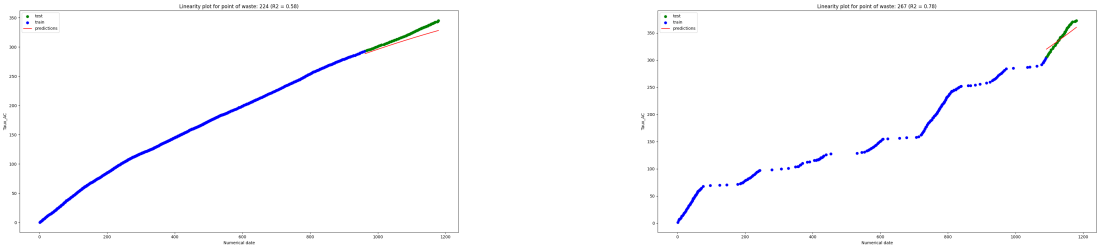


Figure 3.16: PR models (part 4)

General Results and Bloxplot

Let's analyze results and some statistics in the 274 points of waste. First we will show and interpret a boxplot with R^2 results in Figure 3.17.

In the boxplot it can be seen that the orange line (median) is around 0.94 and below 0.50 the values are outliers. Now let's see in how many points we get good results.

# Points s.t. $R^2 < 0.7$	$68 \longrightarrow 24.8\%$
# Points s.t. $R^2 \geq 0.7$	$206 \longrightarrow 75.2\%$

Table 3.13: General R^2 results **minimizing** *MSE* in PR

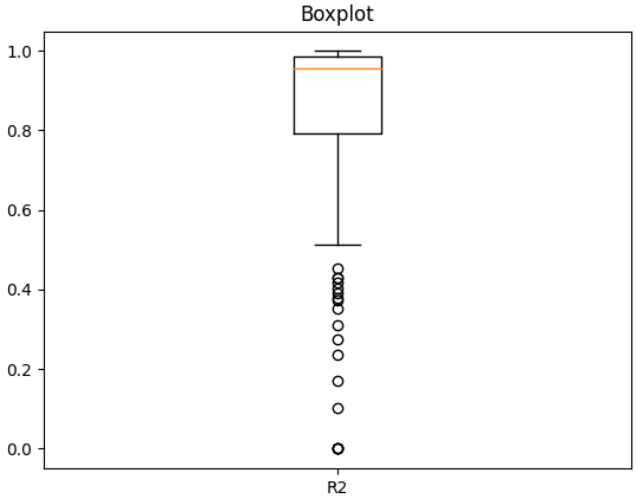


Figure 3.17: Boxplot R^2 results for PR

When comparing Table 3.4 in linear regression and Table 3.13 in polynomial regression, we can observe an improvement of the results as we expected.

General results (minimizing MSE) $\begin{cases} \textbf{Average } R^2 \textbf{ for all the points: } 0.76 \\ \textbf{Average } R^2 \textbf{ for points s.t. } R^2 \geq 0.7\textbf{: } 0.93 \end{cases}$

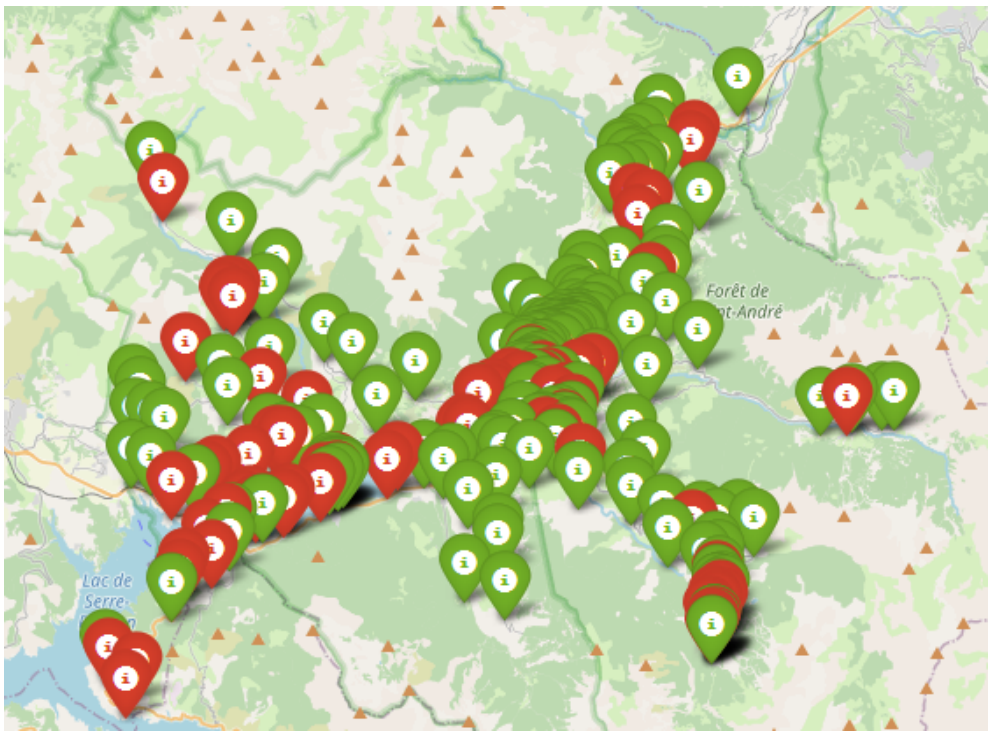


Figure 3.18: Map of Points with $R^2 \geq 0.7$ (green points) and $R^2 < 0.7$ (red points)

As previously mentioned, the focus is on minimizing the MSE since our main goal is to obtain better predictions. Let's examine the map of Figure 3.18. On this map which shows

good and bad model of polynomial regression we can observe that we have a higher number of models with $R^2 > 0.7$ than in linear regression.

Learning Curve: Overfitting Test

As it was mentioned in the State-of-the-Art, polynomial regression can lead to overfitting issues. Therefore, it was decided to observe learning curves for points 8, 9, 224, and 267 (chosen because their models have degrees 2 or 3 and because they have significantly different amounts of data).

These learning curves in Figure 3.19 will plot the percentage of training samples on the X-axis and show the variations in the coefficient of determination (R^2) and mean squared error (MSE) on the Y-axis. The curves will be plotted separately for the training set and the test set. This visual representation will help assess whether or not there is overfitting in our polynomial regression models.

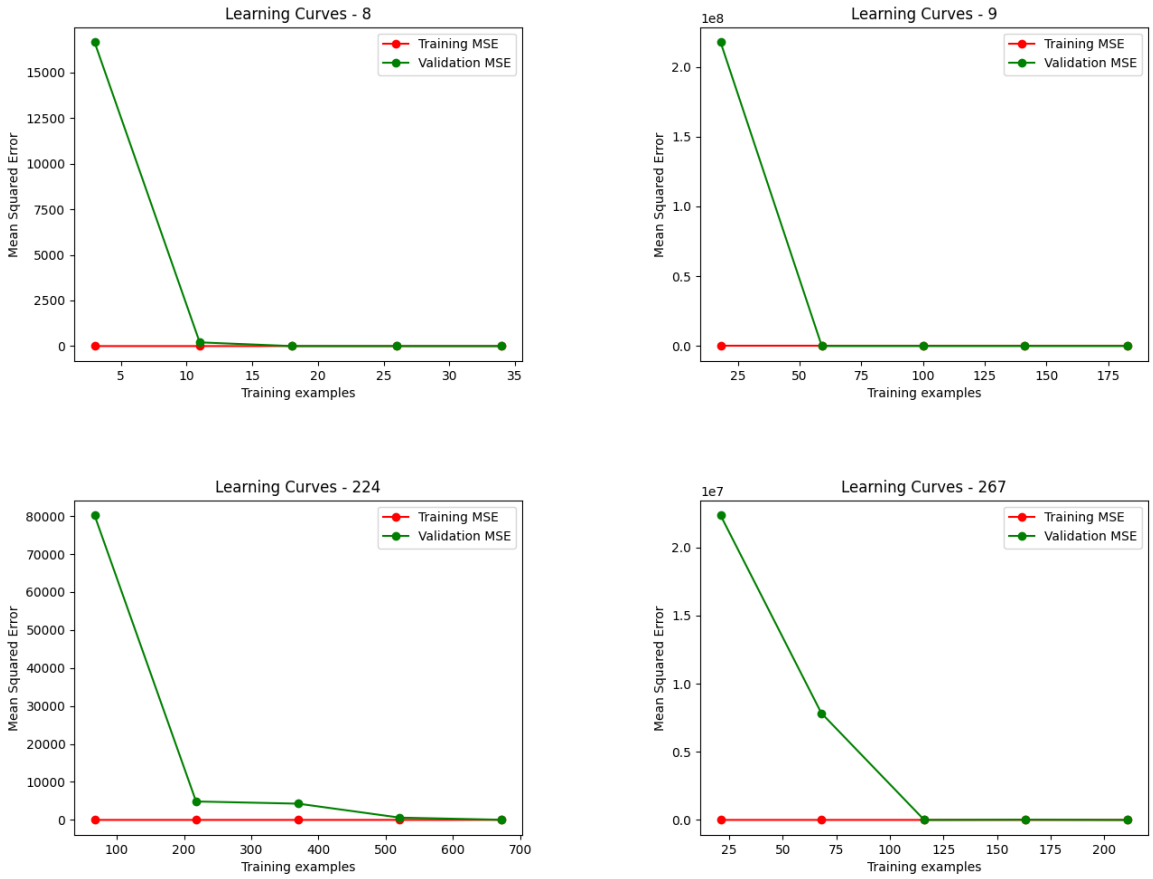


Figure 3.19: Learning curves to test possible overfitting

In conclusion, when analyzing the learning curves of the polynomial regression models, a consistent behavior was observed in the four cases. The MSE in the training set remained constant along the X-axis, indicating that the model is able to fit the training data well. On the other hand, the MSE in the validation set decreased significantly as the size of the training set increased, approaching the MSE of the training set. This is a positive indication that the model is generalizing well and does not exhibit overfitting. These results suggest that the polynomial regression models used are capable of capturing the underlying patterns in the data and adapting

to new observations.

Daily Filling Rates

Again, this section is the consequence of working with the variable 'Taux_AC', a variable that accumulates the waste assuming that it is not collected and, in this way, to be able to train the polynomial regressions.

The polynomial regressions will have the formulas $y = ax + b$ (degree 1), $y = ax^2 + bx + c$ (degree 2) or $y = ax^3 + bx^2 + cx + d$ (degree 3) and, then, for any day x ('Numerical_date') we can calculate the accumulated waste y ('Taux_AC'). To know the daily accumulated waste percentage, it would be enough to follow the same technique as in Table 3.6. Now we have the daily predictions for any day, which is fundamental to continue with the second problem of the project on the waste collection.

Observation: the filled volume $FV(d, p)$ for any day and point of waste is known and measured in litres.

— 4 —

Waste Collection Optimization Strategy

Once the daily filling rate predictions are generated at all points of waste and for each day, we move on to the second part of the project, the optimisation of waste collection planning strategy. This part will include: (1) a mathematical model for the clustering problem and heuristic approach to achieve a solution, (2) set of collection dates and heuristic for the choice of waste collection days for each point of waste and (3) CVRP problem for the routing and heuristic to solve this problem.

4.1 Mathematical Model for Clustering and Heuristic Approach

In this section, we describe how we mathematically model our clustering problem. Its objective consists in finding a set of clusters in terms of distance but the clusters must also balanced in terms of total volume in a specific period. As a basis we use the model given in [19] according to the relations set in Table 2.1. As discussed throughout the project, we work on a tactical level, i.e. medium time periods, and therefore the following mathematical model will be applied over a given period of time (set of consecutive days).

The given data looks as follows:

- $P = \{1, ..., n\}$: set of points of waste.
- $T = \{1, ..., d\}$: set of periods.
- $C = \{1, ..., k\}$: set of clusters.
- w_i^t: generated volume in litres of organic waste generated at each point of waste $i \in P$ during a specific period $l \in T$.
- $Dist_{ij}$: distance matrix between any two points of waste (i, j).
- Adj_{ij}: adjacency matrix that will not allow some points of waste to be in the same cluster because the distance exceeded is higher than a given value ρ. The binary variable Adj_{ij} can be defined as follows:

$$Adj_{ij} = \begin{cases} 1, & Dist_{ij} > \rho \\ 0, & \text{otherwise} \end{cases}$$

The decision variables of the base model are:

$$x_{ic}^{l} = \begin{cases} 1, \text{ if point of waste } i \in P \text{ is assigned to cluster } c \in C \text{ during period } l \in T \ , \\ 0, \text{ otherwise} \end{cases} \tag{4.1}$$

$$y_{ijc}^{l} = \begin{cases} 1, \text{ if points of waste } i, j \in P \text{ are assigned to the same cluster } c \in C \text{ during period } l \in T \\ 0, \text{ otherwise} \end{cases} \tag{4.2}$$

Now we can formulate our problem as:

$$Z = min\,(MaxVol^{l} - MinVol^{l}) \tag{4.3}$$

$$s.t.$$

$$\sum_{c \in C} x_{ic}^{l} = 1 \qquad \forall i \in P, l \in T, \tag{4.4}$$

$$\sum_{i \in P} x_{ic}^{l} \geq 1 \qquad \forall c \in C, l \in T, \tag{4.5}$$

$$y_{ijc}^{l} \geq x_{ic}^{l} + x_{jc}^{l} - 1 \qquad \forall i, j \in P, c \in C, l \in T, \tag{4.6}$$

$$V_{c}^{l} = \sum_{i \in P} w_{i}^{l} \times x_{ic}^{l} \qquad \forall c \in C, l \in T, \tag{4.7}$$

$$MaxVol^{l} \geq V_{c}^{l} \qquad \forall c \in C, l \in T, \tag{4.8}$$

$$MinVol^{l} \leq V_{c}^{l} \qquad \forall c \in C, l \in T, \tag{4.9}$$

$$x_{ic}^{l} \in \{0,1\} \qquad \forall i \in P, c \in C, l \in T, \tag{4.10}$$

$$y_{ijc}^{l} \in \{0,1\} \qquad \forall i, j \in P, c \in C, l \in T. \tag{4.11}$$

Here, the objective is to minimize the difference between the cluster with the highest volume and the cluster with the lowest volume in any period according to the objective of having balanced clusters in terms of volume of waste generated but also, taking distance into account. Constraint (4.4) assigns each point of waste i to 1 cluster in all the periods, (4.5) ensures that any cluster will be empty in any period, (4.6) ensures that if two points are incompatible according to the adjacency matrix Adj_{ij} will not be assigned to the same cluster in any period. (4.7) and (4.8) save the clusters with the highest and lowest volume respectively in each period and constraints (4.9) and (4.10) ensures binary variables for x_{ic}^{l} and y_{ijc}^{l}.

The following result is an approach to solve the above mathematical model in a specific period from 14/04/2023 to 16/06/2023 (9 weeks), which will follow two algorithms:

1. Agglomerative clustering algorithm: a purely geographical clustering will be carried out with the pseudocode of Algorithm 4 and making use of AgglomerativeClustering function from sklearn library in Python. The input data instance is:

   - Set of 274 points of waste in the area.
   - $Dist_{ij}$ distance matrix by road between any two points of waste (i, j), this was computed with an Google API.

- $k = 22$ number of clusters are desired. The reason is because 22 is the number of towns and cities we have in the area as we saw in the analysis. It is considered a reasonable number to split the total volume to collect in the selected period among all the points of waste, and then, in the second step of the approach it will be tried to merge these 22 into only 5 balanced clusters, they all will be collected every week in this period.

The output data is:

- Set C with 22 clusters.
- Generated volume per cluster V_c.

2. Algorithm to balance clusters: it will try to merge the clusters from the output of Algorithm 4 into l new balanced clusters ($l < k$) in terms of volume generated, Algorithm 5 will be applied. It consists in ordering the generated total volumes V_c of the input clusters in descending order, then add it them to the output cluster with the lowest final volume VN_d, always respecting the adjacency matrix. The adjacency matrix will have a 1 in (i, j) if the elements i and j cannot belong to the same output cluster by a distance greater than a parameter ρ. The input data instance is:

   - Set of 22 clusters C from 4.
   - Volume of the 22 clusters V_c from 4.
   - $l = 5$ clusters: 5 final clusters are desired, each cluster will be collected a fixed day of the week, e.g, cluster 1 will be collected on Mondays.

   The output data is:

   - Set of 5 final clusters D.
   - Volume generated for each cluster VN_d.

Algorithm 4 Pseudocode Agglomerative Clustering

Input: set of points of waste $1, 2, .., n$, distance matrix $Dist$ between all the points of waste and number of clusters desired k, generated volume per point of waste w_i.
Output: set of clusters C, generated volume cluster V_c.

$C \leftarrow \{\{1\}, \{2\}, ..., \{n\}\}$ // We initialize one cluster C_i per point of waste
while $|C| > k$ **do**
    $Dist(C_i^*, C_j^*) \leftarrow \arg\min C_i, C_j \in CDist(C_i, C_j)$
    $C \leftarrow C \backslash C_i$; $C \leftarrow C \backslash C_j$; $C \leftarrow C \cup \{C_i \cup C_j\}$
end while
for $c \in C$ **do**
    $V_c \leftarrow \sum_{i \in c} w_i$
end for
Return C, V_c

Now we apply this algorithm with the AgglomerativeClustering function from sklearn library and we can see the results in Figure 4.1.

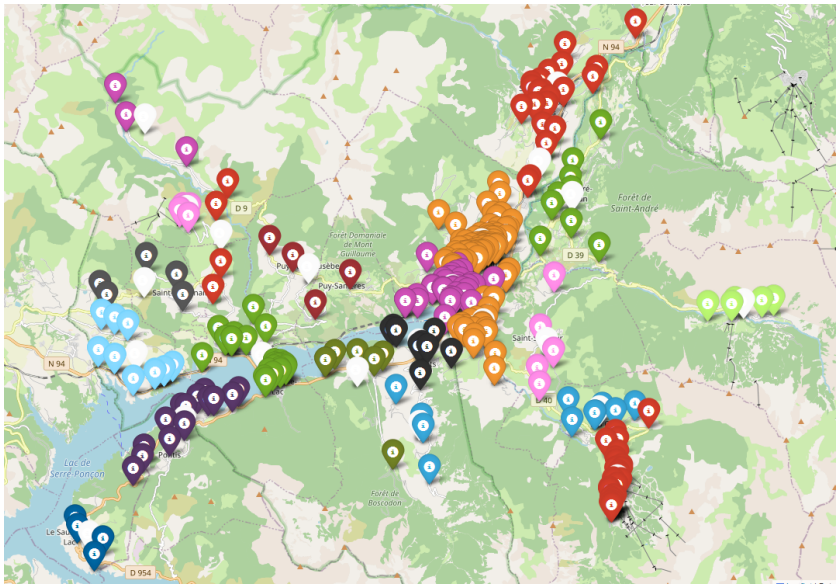


Figure 4.1: Agglomerative clustering results

As we said, the second approach consists in balancing the 22 clusters to obtain the final 5 using Algorithm 5. We can see the map with 5 colors joining the last clusters in Figure 4.2 (labels are the geographical centroids of agglomerative clustering result). Also we can see the results for each cluster in terms of % volume collected per final cluster.

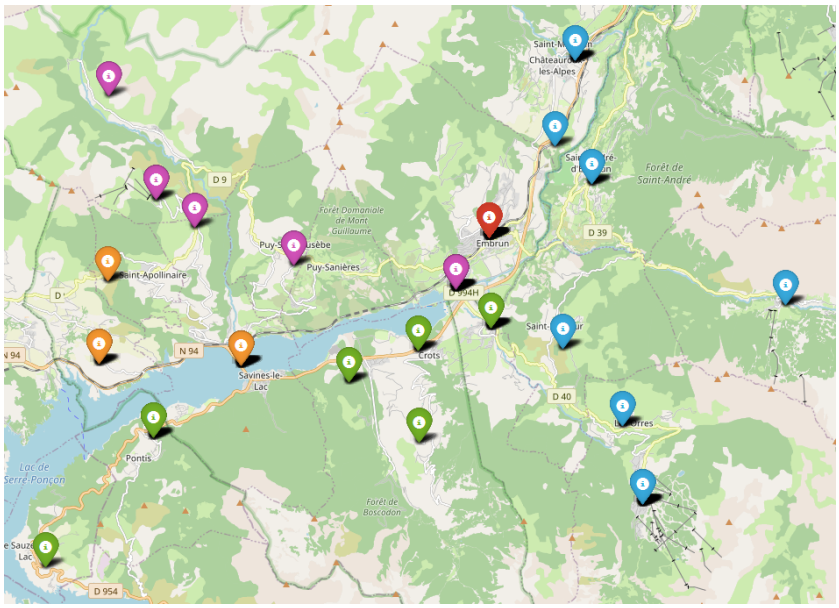


Figure 4.2: Balance clustering results

Cluster	Total % to be collected
0 (red)	29.05 %
1 (green)	17.52%
2 (blue)	19.61%
3 (purple)	17.61%
4 (orange)	16.21%

Table 4.1: % of total volume to be collected in spring off-season period 14th Apil - 16th June

According to the results in Table 4.1, clusters 1, 2, 3 and 4 can be considered balanced since the difference is not very significant. However, cluster 0 has a bigger difference, looking at Figure 4.2 is the red point and it has not been joint with any other cluster because Agglomerative Clustering already provided this result in consequence of the higher quantity of waste generated in the city of Embrun.

Algorithm 5 Pseudocode Create Balanced Clusters

Input: last set of clusters $1,2,..,k$ from Algorithm 4, generated volume per last clusters V_c, adjacency matrix between all the last clusters $AdjCluster$ and number of new clusters l.
Output: set of new clusters D, generated volume new cluster VN_d respecting adjacency matrix constraints.

$Pairs(index, V_c) \leftarrow OrderDesc\ V_c$ // Order V_c in descendent order in pairs with its index
$D \leftarrow \{\{\},\{\},...,\{\}\}$ // We initialize the set of new clusters, each empty set S_d
$VN_d \leftarrow [0,0,...,0]$ // We initialize the set of generated volumes in the new clusters
for $idx, num \in Pairs(index, V_c)$ **do**
  $D_{selected} \leftarrow None$ // Initialization selected cluster
  $MinSum \leftarrow +\infty$ // Initialization volume
  for $h, cluster \in D$ **do**
    $ValidAssign \leftarrow True$
    for $idx2 \in cluster$ **do**
      if $AdjCluster_{idx2j} = 1$ **then**
        $ValidAssign \leftarrow False$ // If these two elements are incompatible
      end if
    end for
    if $ValidAssign = True$ **then**
      if $VN_h \leq MinSum$ **then**
        $MinSum \leftarrow VN_h$ // Take smallest sum of VN_d
        $D_{selected} \leftarrow h$
      end if
    end if
    if $D_{selected}$ is not $None$ **then**
      $D_{selected} \leftarrow idx$ // If we find some cluster we add the index
      $VN_{D_{selected}} \leftarrow num$ // If we find some cluster we add the volume V_c
    end if
  end for
end for
Return D, VN_d

4.2 Heuristics for the Choice of Waste Collection Days

A heuristic was developed in Python that decides, once the collection days of each cluster are set, which points of waste should be visited based on their filling rate.

Assumptions

- Chosen period: 14th April 2023 - 16th June 2023, 9 weeks, 64 days.
- Date of collection $DateCol_{cluster}$: as we said, we have a set of 5 clusters as balanced as possible and we set the following collection dates if needed:
  - Cluster 0: since cluster of Embrun has almost 30% of total volume to be collected two weekly collections will be done.
  - Cluster 1: it will be collected on Mondays. In case of national holiday it will be collected on Tuesday, like 1st May 2023, 8th May 2023 and 29th May 2023.
  - Cluster 2: it will be collected on Mondays, same case for national holidays.
  - Cluster 3: it will be collected on Fridays.
  - Cluster 4: it will be collected on Fridays.
- Predictions for points of waste $Pred_{point}$ are chosen according to Algorithm 6.

Algorithm 6 How predictions are chosen for each point of waste

1: **for** *point in Points* **do**
2: $Pred_{point} \leftarrow MinMSE_{point}(TS, PR)$ // Predictions with smallest MSE are chosen
3: $Pred_{point}[0] \leftarrow Pred_{point}[0] + 0.5$ // We assume in the first day of period there was 50% accumulated rate in every point of waste
4: **end for**
5: Return $Pred_{point}$

- Maximum filling rate $MaxRate_{point}$: maximum rate at which each point of waste is collected, defined in 4.2.

Time Series	Polynomial Regression
80%	If $R^2 \in [0, 0.75]$ at 70%
	If $R^2 \in]0.75, 0.80]$ at 75%
	If $R^2 \in]0.80, 0.85]$ at 80%
	If $R^2 \in]0.85, 0.90]$ at 85%
	If $R^2 \in]0.90, 0.95]$ at 90%
	If $R^2 \in]0.95, 1]$ at 95%

Table 4.2: Maximum filling rate to collect per point of waste

Once the assumptions have been read the procedure is simple. For each cluster $c \in C$ and for each point of waste in that cluster $point \in c$, the predictions of the point $Pred_{point}$ will be used and, knowing on which dates the collections $DateCol_{cluter}$ were set, the parameter $MaxRate_{point}$ will not be exceed. In this way, the points to be collected and the amount of waste generated in liters will be obtained for each date. With this information, the next phase of the work, the *CVRP* problem and applied heuristic for solving, will be carried out.

4.3 Capacitated Vehicle Routing Problem

In section 4.2 it was possible to see the strategy used to know when to visit a point of waste and the volume in liters that will be collected. Hence, according to the State-of-the-Art we are faced with the Capacitated Vehicle Routing Problem (*CVRP*). We also know that this problem is *NP*-hard so it can only be solved for small instances, thus as we saw in the State-of-the-Art the Clark & Wright algorithm is a heuristic that can help us to find solutions close to the optimal ones. This algorithm will be applied for each set of points to be collected belonging to a cluster.

Input:

- Set of points of waste $P = \{1, ..., n\}$ to be collected.
- Depot, as it was said previously, this is the origin point where all the trucks depart and arrive, i,e, where the routes begin and end.
- Demand of the points of waste, i.e. volume generated (in litres) to be collected.
- Distance matrix $Dist_{ij}$, this measures the distance by road between any two points of waste (i, j) per route and it is computed with a Google API.
- Capacity of the truck: it is also measured in litres. According to data provided by the company, 1000 litres of organic waste in terms of volume are 125 kg in average. Therefore, as the truck has a capacity of 12.000 kg, the capacity of the truck is estimated to be 68.570 litres. This conversion is made since the data of the volumes generated at each point of waste are in litres and the same measure is needed as in the demands generated.

Output: final tour to be done, i.e. complete sequence of visits made by the truck and number of kilometers.

4.3.1 Results on Spring Off-season 2023 with Clusters

This section presents the results obtained for the spring off-season period from 14th April 2023 to 16th June. The full procedure: (1) predictions, (2) agglomerative clustering, (3) clustering to balance, (4) procedure for the choice of waste collection days and (5) Clark & Wright algorithm is set out in the Table 4.3.

Cluster	Distance performed (km)	Volume collected (l)	Ratio (l/km)
Cluster 0	607	2.876.300	4739
Cluster 1	1397	2.129.058	1524
Cluster 2	1693	2.035.192	1202
Cluster 3	1048	1.825.894	1742
Cluster 4	1249	1.516.079	1213
Average	5994	10.382.523	1732

Table 4.3: Results 14th April - 16th July 2023

4.3.2 Results on Spring Off-season 2023 without Clusters

This section presents the results obtained for the spring off-season period from 14th April 2023 to 16th June as in previous section but without the use of clusters. The followed procedure is:

(1) predictions, (2) procedure for the choice of waste collection days and (3) Clark & Wright algorithm. The results are presented in the Table 4.4.

Distance performed (km)	Volume collected (l)	Ratio (l/km)
7814	10.725.094	1372

Table 4.4: Results 14th April - 16th July 2023

Comparing the results with clustering in Table 4.3 and without clustering in Table 4.4, we can notice that the ratio is significantly better using clustering since we collected more litres in each kilometre done.

4.3.3 Results 2022 with real data

In Table 4.5 the real data provided by CCSP in the autumn off-season period from 1st September 2022 to 15th September are presented and, in addition, since the points of waste and the demands generated on each date are known, the following will be applied: (5) Clark & Wright algorithm to see if it is possible to improve the routes without applying any of the developed approach.

Results	Distance performed (km)	Volume collected (l)	Ratio (l/km)
Real data provided	7467	8.132.409	1089
Routes $C\&W$	5411	8.132.409	1503

Table 4.5: Results 1st September - 15th December 2023

In this table it can be seen that by simply applying the Clark & Wright heuristic to the demands provided by the company they would have collected the same volume by doing around 2000 kilometres less over a period of 3 months and a half.

— 5 —
Conclusion and Perspectives

By way of conclusion, we resume the different stages of our methodology introduced in section 1.4. For different stages, we draw up the results obtained.

1 & 6 With regard to the State of the Art, it was possible to identify the most generated waste among the different types: organic waste. Then, a deeper search was carried out on organic waste and it was discovered that the factor that most affects its generation is the population. It was also concluded that seasonal tourism has an important impact making its predictions more difficult.

Furthermore, articles on predictions of organic waste prediction (first problem to solve at tactical level) were read, here it is concluded: (1) there are hardly any predictive models with historical data and that most of them use socio-economic variables and (2) most studies are at strategic decision level instead of tactical level as was primarily sought. It is also concluded the most used technique is linear regression.

Finally, articles on waste collection process (second problem to solve at tactical level) were read. We have proposed an integrated approach to solve the problem in several steps: (1) an agglomerative clustering algorithm using road distances to partition the set of points of waste, (2) an algorithm to create a new and smaller set of clusters with balanced volumes, (3) compute the collection date collection for the whole set of points of waste and (4) compute the vehicle routing using Clark & Wright heuristic inside each volume balanced cluster.

2 & 3 & 6 Secondly, Machine Learning techniques were applied to predict the filling rate of the points of waste: linear regression, time series and polynomial regression (degrees 1, 2 and 3) are implemented. It is concluded that: (1) linear regression provides models with good accuracy at points where there is not much tourism and waste is accumulated very constantly, (2) it is not recommended to use time series with less than 100 data, as it may not be enough to train a model, in fact, 18% of our points of waste do not provide good results, (3) time series models provide good results at seasonal points of waste. Neural networks could be a technique applied as future research, but as read in an article it is recommended to have between 150 and 200 registers.

Regarding the accuracy of the models, the coefficient of determination R^2 (not in time series) and the Minimum Square Error (MSE) are measured. It should be noted that 82% of the time series models produce good results and 75.2% of the polynomial regression

models also, in general, time series provide better results but it depends on the behaviour of the different points of waste, usually time series provide better results in points with seasonality. Therefore, it is concluded that it is possible to make predictions with historical records and that it would be convenient to continue applying them and improving their accuracy.

4 & 6 Thirdly, and after applying the strategy to collect the organic waste (hierarchical clustering with distances, creating new clusters balancing the volume, fixing collection dates and heuristics to solve the routing problem), it is concluded that there are several ways to approach the resolution of this problem and that, as we will see in the next point, it produces good results. One of the research perspectives could be how to better fix the dates of collection in order to improve results.

5 & 6 Respecting to the results, it was computed a ratio (l/km), then, the higher this amount is, the better the results we obtain. We can see that in the data we got for autumn off-season period in 2022 they got 1089 l/km, however, if they would have applied Clark & Wright heuristics they could have got 1503 l/km, i.e. a significantly better ratio just by modifying the routes (here we do not take into account the predictions). However, this good result need to be checked and validated with CCSP to ensure that we did not miss to take some business constraints into account while solving with the heuristics.

In addition, applying our whole methodology (including predictions) on the spring off-season from 14th April 2023 - 16th June 2023, we get 1732 l/km, the best ratio among all. In summary, this implies that both the strategy pursued and the Clarke & Wright heuristic itself can produce better results than CCSP currently has.

One of the limitations we found is that UNICO France did not provide in time the actual route data for this period and this could have helped to compare the two global solutions at tactical level including waste predictions and collection process. It was intended to compare the waste predictions with the actual data and then, the collection process with the ratio and the kilometres done.

In general conclusion, we can provide an answer to the initial research question in section 1.3. It is interesting to apply Machine Learning techniques using historical data to predict the waste demand and design a waste collection planning to improve the current solutions. It is possible for the trucks of CCSP to run less kilometres: (1) just using a routing heuristic as we did and (2) collecting less the points, i.e. having good prediction models and wait a bit more time to collect them.

Appendix

References	Contributions	Data used	Decision level	Prediction Techniques	KPI's used and accuracy
[3]	Organic waste is the most significant among all the solid wastes. Introduction to seasonality in waste generation. Population is the most important factor in waste generation.	Socioeconomic factors	Strategic level	Linear regression	$R^2 = 0.55$ $R^2_{adj} = 0.51$ F-test significative
[5]	Tourism impact on waste generation. Seasonality, making waste generation prediction more difficult and subsequently waste collection.	Socioeconomic factors Geographical factors	Strategic level	Linear regression Spatial Autoregressive model (SAR)	$R^2 = 0.53$ (LR) $R^2 = 0.52$ (SAR)
[1]	Big review of 88 articles. Models applied on the articles. Identification on the main factors affecting waste generation.	Socioeconomic factors	Strategic level	Regression models, Artificial Neural Networks, Statistical Analysis, Descriptive Analysis, Time Series...	It is measured the significance of 16 variables in the different articles
[11]	Waste generation forecasting using historical data with a GBRT model in New York City.	Historical Data	Operational level	Gradient Boosting Regression Model	$R^2 = 0.88$ *RSME*
[9]	Linear regression article predicting waste collection using socioeconomic factors. Study the main Key Performance Indicators (KPI's) used in Linear Regression.	Socioeconomic factors	Strategic level	Linear regression	$R^2 = 0.82$ *MSE* *MBE* *MAPE* t-test
[18]	Presentation of five SARIMA models for waste collection and results. Data used to apply SARIMA models.	Historical data	Strategic level	SARIMA	*MSPE* *MRPE*
[17]	Article comparing some ARMA and ARIMA models for waste collection, including how parameter estimation is carried out. Revision on data used to apply ARMA and ARIMA models. Study the most commonly used KPI's to measure model accuracy.	Historical data	Strategic level	ARMA ARIMA	*MSE* *AIC* *BIC*
[16]	Article applying time series models for waste collection. Reviewing KPI's used to select the best model.	Historical data	Strategic level	ARIMA	*MAPE* *MAD* *MSE*

Table 5.1: Main characteristics of waste generation forecasting references

Acknowledgments

I would like to use these lines to thank those people who have helped me throughout this university master's degree.
First of all, I would like to thank Van Dat Cung for giving me the opportunity to work with him, for his support throughout the development of this work and for his confidence in assigning me this project.
To my parents, Julia and Antonio, whom I thank for all their support in making the decision to do this master's degree and for their psychological support during the year. I would also like to thank my sister Lydia for her support whenever I needed it and for helping me as much as she could.
Finally, I would like to thank my colleagues for the great work we have all done together during the first semester and for the general support, especially Vasilis, Benoist and Nicolas.
To all of you, thank you very much.

Bibliography

[1] Bruno Ribas Alzamora, Raphael Tobias de Vasconcelos Barros, Leise Kelli de Oliveira, and Sabrina Silveira Gonçalves. Forecasting and the influence of socioeconomic factors on municipal solid waste generation: A literature review. *Environmental development*, 44:100734–, 2022.

[2] Fernanda Medeiros Assef, Maria Teresinha Arns Steiner, and Edson Pinheiro de Lima. A review of clustering techniques for waste management. *Heliyon*, 8(1):e08784, 2022.

[3] Nilanthi J. G. J Bandara, J. Patrick A Hettiaratchi, S. C Wirasinghe, and Sumith Pilapiiya. Relation of waste generation and composition to socio-economic factors: a case study. *Environmental monitoring and assessment*, 135(1-3):31–39, 2007.

[4] Zuzana Borčinova. Two models of the capacitated vehicle routing problem. *Croatian Operational Research Review*, 8(2):463–469, 2017.

[5] Vincenzo Caponi. The economic and environmental effects of seasonality of tourism: A look at solid waste. *Ecological economics*, 192:107262–, 2022.

[6] D.M. Cocarta, E.C. Rada, M. Ragazzi, A. Badea, and T. Apostol. A contribution for a correct vision of health impact from municipal solid waste treatments. *Environmental Technology*, 30(9):963–968, 2009. PMID: 19803335.

[7] Waste Tech Engineering. Key components of your business' solid waste management system, 2022. 13/07/2023, https://wastech.com.au/wtnews/key-components-of-your-business-solid-waste-management-system/.

[8] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):1–34, 2012.

[9] Obiora B. Ezeudu, Chigbogu G. Ozoegwu, and Christian N. Madu. A statistical regression method for characterization of household solid waste: A case study of awka municipality in nigeria. *Recycling (Basel)*, 4(1):1–0, 2019.

[10] Argentina gobierno. Ministerio de ambiente y desarrollo sostenible. Etapas de la gestión integral de residuos sólidos urbanos. 12/07/2023, https://www.argentina.gob.ar/ambiente/control/rsu/etapas.

[11] Nicholas E. Johnson, Olga Ianiuk, Daniel Cazap, Linglan Liu, Daniel Starobin, Gregory Dobler, and Masoud Ghandehari. Patterns of waste generation: A gradient boosting model for short-term waste prediction in new york city. *Waste management (Elmsford)*, 62:3–11, 2017.

[12] Asiye Kurt, Mario Cortes-Cornax, Van-Dat Cung, Agnès Front, and Fabien Mangione. A Classification Tool for Circular Supply Chain Indicators. In Alexandre Dolgui, Alain Bernard, David Lemoine, Gregor von Cieminski, and David Romero, editors, *IFIP Advances in Information and Communication Technology*, volume AICT-634 of *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems*, pages 644–653, Nantes, France, September 2021. Springer International Publishing.

[13] Mathieu Lepot, Jean-Baptiste Aubin, and François H.L.R. Clemens. Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, 9(10), 2017.

[14] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[15] John A. Muckstadt. *Principles of Inventory Management When You Are Down to Four, Order More*. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 1st ed. 2010. edition, 2010.

[16] Amon Mwenda, Dmitry Kuznetsov, and Silas Mirau. Time series forecasting of solid waste generation in arusha city-tanzania. *Mathematical Theory and Modeling*, 4(8):29–39, 2014.

[17] Noryanti Nasir, S. Sarifah Radiah Shariff, Siti Sarah Januri, Faridah Zulkipli, and Zaitul Anna Melisa Md Yasin. Time series forecasting of solid waste generation in selected states in malaysia. *International journal of advanced and applied sciences*, 10(4):76–87, 2023.

[18] J. Navarro-Esbrı, E. Diamadopoulos, and D. Ginestar. Time series analysis and forecasting techniques for municipal solid waste management. *Resources, conservation and recycling*, 35(3):201–214, 2002.

[19] Joaquín Pérez-Ortega, Hilda Castillo-Zacatelco, Darnes Vilariño-Ayala, Adriana Mexicano-Santoyo, José Crispín Zavala-Díaz, Alicia Martínez-Rebollar, and Hugo Estrada-Esquivel. Una nueva estrategia heurística para el problema de bin packing. *Ingeniería, Investigación y Tecnología*, 17(2):155–168, 2016.

[20] M. Thürer, Y. H. Pan, T. Qu, H. Luo, C. D. Li, and G. Q. Huang. Internet of things (iot) driven kanban system for reverse logistics: solid waste collection. *Journal of intelligent manufacturing*, 30(7):2621–2630, 2019.

[21] Zeming Wei, Chufeng Liang, Hua Tang, et al. A cross-regional scheduling strategy of waste collection and transportation based on an improved hierarchical agglomerative clustering algorithm. *Computational Intelligence and Neuroscience*, 2022, 2022.

[22] Zaquin. Market overview. 12/07/2023, https://zaquin.com.my/alpha/market-overview/.