Forecasting yearly unplanned maintenance on underground waste containers based on maintenance history and demographics.

Computer Science Master thesis

Florian Ven

November 16, 2020

Contents

1 Abstract

Municipal solid waste is ubiquitous: every city, municipality, and country has to deal with waste generated by its inhabitants. A common collection method is to let citizens deposit their waste in large ($5m^3$) waste containers close to the home from which it is then collected. This thesis researches the unplanned (e.g. corrective) maintenance needs of such large underground waste containers and attempts to forecast those maintenance needs. Literature on the subjects of maintenance forecasting and municipal solid waste generation is studied to identify possible predictors. The identified predictors are validated using a linear regression model in a case-study of the municipality of Amsterdam. The case-study shows that asset age, specific asset types, and the number of assets are meaningful indicators of upcoming maintenance. The final predictive model has an expected error of 32% of the target variable. While the prediction error is too high for practical use, this thesis breaks ground on prediction on these specific types of assets and thoroughly documents the subject and open work for use in future work.

2 Introduction

For municipalities, it is important that municipal solid waste (MSW) is collected in a quick and efficient manner. Two common examples of MSW collection approaches are curbside collection, where the waste is collected directly from the curb near a residency, and making use of collection points, where garbage is deposited into centralized collection points by the citizens, and collected from these points by the municipality. Such centralized collection points can be implemented in the form of large waste containers near residencies. In the Netherlands, aside from curbside collection, it is common for municipalities to place $5m^3$ containers within walking distance (200m) of homes.

Just like most assets, these containers need to be maintained: they require preventive maintenance to ensure smooth functioning, require inspections to ensure they satisfy (legal) requirements, and require corrective maintenance when they unexpectedly break down. Assets that feature lots of moving parts may be monitored through sensors to preemptively predict exactly when a specific asset is at risk of failing, a common example of this is the analysis of vibrations in bearings [18]. However, garbage containers generally have few moving parts that can be actively monitored. Furthermore, placing sensors is not cost effective or overly complex due to a combination of factors such as rough handling conditions, being subjected to weather, and not consistently having access to a power supply. As such, real-time predictive maintenance on underground waste containers is not feasible. Instead of predictive maintenance, this study will attempt to forecast the yearly need for unplanned maintenance. Based on such a forecast, indicators can be identified and possible addressed.

In literature, forecasting maintenance is known as Prognostic Health Management (PHM). Generally, a Remaining Useful Life (RUL) of an asset is calculated/estimated, based on which estimates can be made on the required maintenance in the near future. Common methods of determining a RUL are real-time models using sensory input, e.g. [17], referred to as Predictive Maintenance (PdM), or statistical models, e.g. [14]. However, these methods either require sensor data or an extensive life-time history of (comparable) assets to build statistical models on. This study will add to literature by predicting the total required unplanned maintenance for a given timespan on a population underground waste containers based on recent maintenance history, asset properties, and external factors, such that sensor input or extensive life-cycle data is not required. The Dutch municipality of Amsterdam is used as a case study of the proposed methods.

The municipality of Amsterdam has been deploying underground waste containers since 2012, and has consolidated maintenance efforts and coordination of these containers since 2017. At the end of 2019 Amsterdam has 862 965 inhabitants [4] and roughly 13 000 underground waste containers, covering the majority of neighbourhoods. The distribution of inhabitants and containers over the city is discussed in Section 4 and Section 5 respectively. On these 13 000 containers, approximately 51 000 planned and 12 000 unplanned maintenance actions were scheduled in 2019. Assets and maintenance are administrated through the asset management application *Grybb*, created and maintained by the Dutch company *Curious Inc.* The situation in the municipality of Amsterdam is discussed in more detail in Section 4.

The municipality of Amsterdam has set the strategic goal of lowering the number of maintenance actions on the waste processing infrastructure. Since the planned issues consist mainly of (legally) required maintenance, and are thus hard to reduce, this reduction should be sought in unplanned maintenance. Curious Inc. wishes to extend their applications functionality with actionable (unplanned) maintenance forecasts. This research serves both goals in the following ways. First, by identifying meaningful factors to unplanned maintenance, the feasibility of lowering maintenance can be determined and implementation steps may be found. Secondly, the resulting predictive model can be used for maintenance forecasts in *Grybb*.

2.1 Problem statement

The field of PHM knows plenty of examples where maintenance on assets can be predicted on short term based on sensor input, or long term based on statistical models. These methods require sensor input or extensive history of the assets life-cycle to apply. Predictions on assets that are not equipped with sensors or do not yet have an extensive history to base statistical models on are not covered by existing literature. This thesis will attempt to fill this gap, specifically on the prediction of maintenance on underground waste containers. This prediction must be explainable to domain experts, such that the causes of maintenance may be addressed if possible. To guide the thesis, the following research questions are formulated.

> *Research question: How can unplanned maintenance on underground waste containers be estimated?*

To create a predictive model, two elements are of importance: what variables to consider and what type of model to use. As such, the following sub questions are posed.

> *Sub question 1: What factors can be relevant in predicting required maintenance on underground waste containers?*

Relevant factors must be clearly identified and it should be possible to explain such factors to domain experts, such that causality can be determined and optionally the causes addressed.

> *Sub question 2: What types of models are appropriate to model the required maintenance in such a way that relevant factors can be identified and explained to domain experts?*

To validate the identified variables and models, they will be used for a predictive model for unplanned maintenance on underground containers in the municipality of Amsterdam.

> *Sub question 3: To what extent can the chosen model predict the required unplanned maintenance for the municipality of Amsterdam?*

Having answered these questions, it should be clear if, given the imposed restrictions, prediction of maintenance on underground waste containers is feasible.

2.2 Methodology

To answer the formulated research questions, the design science research process (DSRP) model as discussed in [20] is used. It distinguishes six steps to designing a solution within information systems research: problem identification and motivation, solution objectives, design development, demonstration, evaluation, and communication. Problem identification, motivation, and solution objectives have been provided as part of this chapter.

Design and development will consist of two stages: a review of the state of the art in prediction of maintenance and MSW generation, and the design of a predictive model. A review of the state of the art should suggest what models commonly show success in related studies, and what input variables matter to maintenance and the domain of waste. A demonstration will then be implemented in the form of a case study in which the yearly unplanned maintenance for the municipality of Amsterdam is predicted. The application of the model in the context of the municipality of Amsterdam is further guided by the Knowledge Discovery in Databases (KDD) framework, which details the general process of distilling knowledge from raw data. Finally, the model is evaluated. This document constitutes the final step of the DSRP, in which the process, artifacts, and results are communicated.

The structure of this thesis is as follows: Section 3 discusses the state of the art of maintenance and MSW prediction and provides answers to sub questions 1 and 2 with regard to model and input variable selection. Section 4 details the current state of underground containers and related maintenance in the municipality of Amsterdam and Section 5 discusses the data that is available in the case study in detail. Section 6 then explains how the suggested model and variables are applied to the case study and Section 7 details the results of the case study implementation. Finally, all findings are discussed and concluded in Sections 8 and 9 respectively.

3 Literature review

To explore the state of the art related to maintenance on underground containers, literature in three areas is reviewed. Once done, this review should provide answers to sub questions 1 and 2, and the useful model(s) and input variable(s) should be clear. The following subjects are reviewed in the following subsections:

- **Maintenance prediction in public infrastructure.** Maintenance in the specific area of (underground) waste containers is, to the best of the authors knowledge, non-existent. However, the most relevant property of underground waste containers seems to be that they exist in the public domain and are actively used by citizens. These traits are common to public infrastructure, and related literature can therefor help place this research in context.
- **Municipal solid waste generation.** While not directly related to maintenance, much research has been performed on the generation of municipal solid waste by citizens. Patterns visible in the generation of waste may be transferable to maintenance on waste-processing assets.
- **Maintenance cost estimation.** The estimation of the total cost of maintenance is a subject that is also well-covered by existing literature. This literature gives a more complete overview of predicting maintenance without being bound to a specific domain.

Each section is concluded with the key points identified in literature for the current research.

3.1 Predicting maintenance in public infrastructure

?? Various works cover prediction of various maintenance needs in public infrastructure. While power networks, water mains, and roads are sufficiently covered in existing literature, smaller infrastructure such as garbage processing infrastructure seems to be an untouched topic. Some existing work on infrastructure will be discussed in an effort to identify common factors which may prove useful in maintenance prediction for garbage related infrastructure.

In [3] by Bessani et al., a statistical analysis is made in an effort to predict maintenance requirements on power substations in Brazil. Five categories of causes are identified: atmospheric, environmental, urban, operational needs, and equipment failure. For each category, the Kaplan-Meier estimator is used to estimate the repair times. It is concluded that atmospheric and environmental influences cause the most downtime. The authors state that the planning of maintenance capacity can be optimized by using atmospheric forecasts and maintenance can be prevented by keeping local vegetation pruned.

In [10] and [24], maintenance needs for the New York City power grid are studied. Gross et al. rank electrical feeders based on their susceptibility to failure in [10]. Rudin et al. expand on this by also ranking other components and manhole events, and estimating the mean time between failure of feeders in [24].

Gross et al. describe three types of attributes to be used for the proposed ranking: physical, electrical, and derived. Physical features are described by the components used, electrical are described by simulated and measured system load, while derived attributes are computed from formulas developed by domain experts. They compare three ranking algorithms: pairwise ranking algorithm RankBoost, a classification score based ranker using SVM, and Martingale Boosting. They show the SVM performs best while noting that in previous works it did not, showing that algorithm performance may vary within a single domain. Weather is found to cause concept drift: changing the distribution of failures during the monitoring period. This concept drift is compensated for using adaptive windowing techniques as opposed to static windowing.

Rudin et al. expand on feeder failure ranking by ranking other failing parts as well as including an absolute measure describing the probability of a feeder failure. Similar to the work of Gross et al., the authors use physical characteristics (i.e. parts used), date put into service, previous issues, previous power quality events, electrical characteristics, and real time electricity data to perform the ranking. As an absolute measure of risk, the mean time before failure is introduced. This measure is then used to estimate yearly unexpected outages. Success in predicting this value is shown using an SVM.

In [22], maintenance on a nuclear power plant is studied. It is shown that part failure can be modeled by a Weibull distribution. Subsequently, a Crow-AMSAA model is fit to the data, estimating the number of events in a given timespan. Then, a model is constructed to estimate the maintenance costs for specific parts. The model shows the following variables are of importance for work order cost: number of repairs, equipment level of risk to loss of power generation upon failure, equipment level of risk to damage frequency, parts being subject to additional quality control.

Apart from power grids, water distribution nets are subject of study. In [14], water pipe failures are predicted. Various types of pipes are examined, varying in material (steel, PVC) and thickness. Three statistical methods are compared: Cox-PHM, Weibull PHM, and Poisson method. Two meaningful factors are discussed that cause maintenance: (1) previous maintenance within 10 meters that has disturbed the soil and therefore caused a new failure, and (2) low number of users of specific parts of the system causing high fluctuation in pressure. Pipe length is identified as having marginal influence in probability of failure. The authors conclude that model performance varies between types of pipes, suggesting that different types of pipes exhibit different statistical properties.

Finally, road infrastructure is subject to forecasts of required maintenance. In [12], Karballaeezadeh et al. study a section of road in Iran in an attempt to predict the remaining service life of road sections. They employ SVM models with particle filtering to estimate the remaining service years of a road. It is concluded that major factors in an accurate estimation is the temperature and thickness of the road measured at set time intervals.

Conclusion

Various methods are identified in existing literature that seem suitable to the prediction maintenance for underground waste containers. Most notable are the statistical models such as the Weibull distribution and the related Crow-AMSAA model. The existing literature furthermore suggests that every part may exhibit it's own statistical properties, and that a prediction can benefit from modelling each part individually, instead of fitting a model to a generalized failure of the entire container.

3.2 Municipal solid waste generation prediction

While not much work has been done on the prediction of maintenance in the field of municipal solid waste specifically, a lot of work has been done on the prediction of the amount of municipal solid waste generated by a population. Two reviews on the field will be discussed. Both reviews give an overview of the methods used for MSW generation prediction. Finally, four papers using multiple regression analysis for waste generation prediction are discussed. Given the availability of presumably independent variables (i.e. container metadata and demographics) and the research question, the usage of multiple regression analyses are considered most relevant to the current work and are therefore covered in detail.

In [2] Beigl et al. name three groups of variables to predict waste generation: production related, consumption related, and disposal related. The first relates to how much goods are produced, and can in part be described by readily available monetary data such as waste generated per GDP unit, and price per product unit. Secondly, consumption related metrics describe the population that consumes products and thereby produces waste. Variables describing affluence are identified as important, such factors can be income, tenure

of properties, and population density. Apart from affluence related factors, variables describing the type of households show correlation with waste generation. Finally, disposal and collection related metrics can be used to identify the distribution of waste between fractions, these factors include for example fostered recycling activities, container size, and density of collection sites.

Furthermore, Beigl et al. note that models using multiple variables are complex due to interactions between variables, making model validation hard or impossible since it is hard to prove independence of variables and variance and error requirements. Therefore, bivariate analyses are common in predicting waste generation.

Kolekar et al. expand on the work of Beigl et al. in [15] by including more recent papers. Beigl et al. covered papers up to 2005, whereas Kolekar et al. cover papers published between 2006 and 2014. The authors confirm the continued use of income, employment, and urbanization as strong variables, and also shows that the factors seasonal variation, per capita municipal tax, wheather and temperature, and consumption of gas, water, and electricity show good results. Furthermore, level of education and age groups are noted as variables with the highest results in waste generation prediction. With regard to methods, Kolekar et al. show that single regression analysis is the most commonly used model. Artificial neural nets, multiple regression analyses, and fuzzy logic are also common.

Keser et al. [13] study the spatial dependency of variables in prediction of municipal solid waste generation in Turkey. The authors observe spatial dependency in the data: many factors exhibit a change in the east to west direction. This change is in in line with general wealth distribution over the subject area. The goal of the study is to show significance of this spatial variation and compensate for it in the prediction. Before analyzing spatial relations, the regular variables are checked for normality using the Kolmogorov–Smirnov test. Collinearity is between explaining variables is estimated by Pearson's r for bivariate collinearity and Variance Inflation Factor (VIF) for multivariate collinearity. Variables with a collinearity value of 0.4 and 4 respectively were eliminated. The authors consider four models: ordinary least squares regression (OLSR) with and without neigbouring areas, spatial autoregression (SAR), and geographically weighted regression (GWR). Spatial autoregression performs regression and takes a global spatial correlation into account, while geographically weighted regression allows for variation of spatial correlation between overarching areas in the considered data. The models are validated with a historic dataset. The OLSR and SAR models perform comparable, the GWR model shows that the importance of various indicators varies between different (types of) areas. Unemployment rate, temperature, higher education graduates ratio, and agricultural production value are listed as the main explanatory variables.

Chung [5] models the prediction of MSW generation as a time series in Hong Kong SAR. They present the use of an autoregressive model based on ARIMA and compare long term predictions (+-30 years ahead) to a pre-existing simple linear model on waste generation. The authors note that many studies only take factors into account that are positively correlated with waste generation such as population count. In contrast, the study also consider the number of housing estates that participate in waste source separation. The model is shown to track historic MSW generation numbers within a 95% confidence interval. The independent variables used in the final model are GDP per capita, population, and number of housing estates participating in source separation. The last variable shows a strong negative correlation.

In [16], the MSW generation of 542 municipalities in the province of Styria, Austria, are analyzed. A large part of these municipalities has less than 4000 inhabitants, 7% has more than 4000 inhabitants. Waste separation is reported to be highly developed in this area. A large number of demographic variables are considered, as well as 7 indicators related to waste management (i.e. type of collection). To reduce the number of variables used for the model, correlation was calculated for each variable with relation to MSW generation. Because not all variables are normally distributed, Spearman's rank correlation was used. Three criteria

were used to select variables: representation of the different groups of influencing factors, high correlation with MSW generation and low correlation with other variables, and data availability. This pre-selection yields 5 variables: household size, municipal tax per capita, difference in in- and outgoing commuters, percentage of residences with solid fuel heating systems, and overnight stays. Not-normally distributed variables were transformed to a normal distribution before being used in the model. Using the five pre-selected variables, multiple regression models were trained with 1-4 variables. MSW generation was weighed by population, to compensate for a hypothesized higher variance in lower-populated areas.

Conclusion

In conclusion, various demographic variables show promise in predicting the municipal solid waste generation. Factors such as household size, residency type, age groups, employment, gross domestic product (GDP), education, culture, geography, and climate can be seen to influence waste generation. It is possible that such indicators also translate to yearly required maintenance on waste containers either through the amount of usage of the containers, or the ways the containers are used by citizens. Furthermore, (spatial) autocorrelation is identified in existing works to exist in MSW generation. The use of linear models with multiple variables and compensation for (spatial) autocorrelation are common for studies that put emphasis on explainability of models.

The commonly used linear regression has several underlying assumptions which need to be tested before the model is applied, as encountered in [2, 13, 16]. The assumptions that must be satisfied before use of linear regression is valid, are (1) independence of explanatory variables, (2) constant variance and normality of errors (residuals), (3) linear relation between predictor and response variables. Lebersorger et al. [16] utilize the Glesjer test to test for the constant variance of residuals, which regresses the absolute residuals towards the absolute predicted values. Beigl et al. [2] note in their survey that not every paper validates these constraints, which holds true for [13] with regards to the constant variance of errors. Multicollinearity is calculated using variance inflation factor (VIF) by both [16] and [13].

3.3 Cost estimation

Finally, some papers reporting work in general cost estimation of maintenance are evaluated.

De Lucia et al. report success in predicting effort required of corrective maintenance in a software project in [6]. A multiple linear regression model was used to estimate the required effort of maintenance based on the system size and the number of tasks. Performance of the model increased when the types of the maintenance tasks were included. While sacrificing some accuracy, only features that were explainable were included in the model such that it is transparent to business managers.

In [23] a tool is presented to estimate operation and maintenance costs of off-shore wind farms. The tool uses an unspecified regression based on the annual failure frequency of components and estimated costs of these failures. The tool is validated in the field with success.

Edwards et al. [7] show a comparative analysis between a multiple regression and a multilayer perceptron model to predict the hourly cost of maintenance of excavator machines. They report good performance of both models while the multilayer perceptron performs slightly better. Moreover, important explainable features are presented: machine weight, company (attitude towards preventive maintenance), type of industry the machine is employed in, and type of machine.

Conclusion

In conclusion, it is observed that a general regression on the cost of maintenance can be viable. Details about the system being maintained combined with historic cost records can be used as good predictor variables which are transparent to the end users.

4 Case study background: municipality of Amsterdam

To test the models and variables that are identified from literature research on the prediction of maintenance on underground waste containers, the proposed model and variables are used on a case study. The subject of the case study is maintenance on underground waste containers in the municipality of Amsterdam, the capital city of The Netherlands. The case study is performed in three parts: first the situation regarding underground waste containers and related maintenance in Amsterdam is presented (this Section), then the data available as part of the case study is discussed (Section 5), and finally an experiment is laid out and performed to predict the required maintenance in Amsterdam on a yearly basis (Section 6).

In Amsterdam, the majority of municipal solid waste (MSW) is collected through the use of underground containers. These containers have an input mechanism allowing citizens to deposit garbage bags and store the garbage below street level. The containers are readily accessible to citizens, usually within 200 meters walking distance from their residence. The garbage is collected by the municipality from these containers, consolidating the effort of collection when compared to classic curbside collection where every household places their garbage in front of the house to be collected. Multiple containers are generally placed together in a "cluster" to facilitate separated collection of different waste streams (i.e. plastics, glass, etc.). Figure 1 shows a single container in its normal position, with only the insertion mechanism visible (1a), and lifted out of its well, in which the garbage storage is also visible (1b).

(a) Stationary in well

(b) Lifted out of well

Figure 1: Subject containers of this case study shown in its normal position (a), and lifted out of the well (b).

Amsterdam is divided into a hierarchy of three different administrative levels, from largest to smallest with their common Dutch name: districts/stadsdelen, wards/wijken, and neighbourhoods/buurten. Before 2017, municipal solid waste processing was managed by each individual district. Since 2017, management has been consolidated to be city-wide. The districts, wards, and neighbourhoods are shown in Figure 2 along with the population in each neighbourhood to give an overview of the size of the areas and their population density. Districts are annotated in the figure with their name and a letter code used by the

municipality. All districts, wards, and neighbourhoods have an identifying code. Districts are identified by single capital letters, wards add a two digit number to this letter, and neighbourhoods add a lowercase letter to this code. For example, neighbourhood A06j is a neighbourhood of ward A06, which is a part of district A. As of 2019, there are 8 districts, 99 wards, and 481 neighbourhood, however, since district "B Westpoort" is an industrial area and houses very few containers, it is generally ignored in this study.

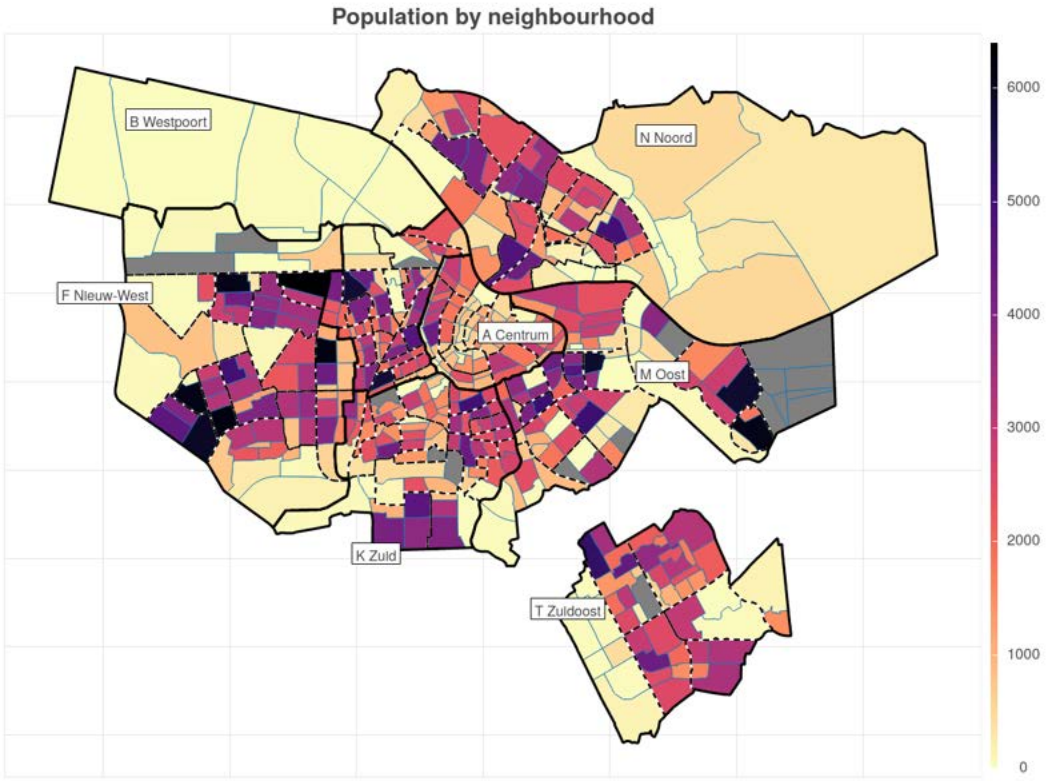


Figure 2: Population by neighbourhood. Districts are separated by black lines, wards by dashed lines, and neighbourhoods by blue lines.

4.1 Containers and wells

A container allows the citizen to deposit their garbage in a centralized place. It consists of five major parts, with exception of the electronics module, all items are shown in Figure 1:

- **Insertion mechanism/inlet**, referring to the entire encasing of the visible part of the container and specifically the mechanism that allows the citizen to insert garbage. Generally a revolving door that allows citizens to deposit 60L or 30L garbage bags. Due to the construction of the mechanism, there is no direct access to the garbage storage. There is an access door on the side of the house to allow mechanics access to the internals of this mechanism.
- **Garbage storage**. The core of the container where the garbage is stored. Typically a container has an internal storage of 5 m^3, but exceptions of 3 m^3 and 7 m^3 exist. The outside of the storage area is visible in Figure 1 (b).
- **Emptying doors**. Doors on the bottom of the container that allow the container to be emptied. The doors are controlled by the hoisting mechanism. The side of the emptying doors can be seen in Figure 1 (b)

(a) Partly opened safety floor

(b) Safety wall

Figure 3: Examples of subject wells, with their safety measures (partly) deployed, (a) shows a safety floor (deployed on the right halve, the left halve in a maintenance setting), while (b) shows a safety wall.

- **Hoisting mechanism**. A hook, rod, or similar construction on top of the container that an emptying or maintenance truck can latch onto to lift the container out of its well. It features a mechanism to release the emptying doors.
- **Pedestrian platform**. A platform around the inlet to cover the garbage storage and provide a tight and seamless connection to the surrounding street, over the edge of the well.
- **Electronics**. A mechanism to restrict and administrate citizens' ability to deposit garbage. May be used for usage-based billing of citizens (also known as DifTar: differential tariffs).

A container is placed in a well: a concrete support structure to allow the container to be easily lifted out of and lowered into the ground. The well features a safety mechanism which prevents direct access to the well whenever it is empty. The primary goal of this mechanism is to prevent people from falling into the well. The mechanism is implemented in one of two ways: a temporary floor is deployed by springs when the container leaves the well, or a temporary wall rises up around the edges of the well as the container leaves the well. Both mechanisms retract automatically as soon as the container enters the well. Figure 3 shows both mechanisms. Table 1 summarizes the parts of a container and well.

4.2 Container and well maintenance

The containers need to available to citizens as much as possible. Various types of maintenance play a role in ensuring maximum availability:

- Quarterly **cleaning** is scheduled to keep the container and well clean and remove any stray garbage from the well. Cleaning may also be scheduled incidentally if required.
- Yearly **inspections** and **preventive** maintenance ensure general safe function of the containers and prevent breakdown. The inspection is legally required and ensures the

Asset	Component	Function
Container	Inlet	Allow input of garbage bags
	Storage	Store the garbage
	Pedestrian platform	Allow people to walk on the container
	Emptying doors	Allow garbage to be released from the bottom of the container
	Hoisting Mechanism	Allow lifting the container and control the emptying doors
	Electronics	Mechanism to administrate citizens' access to dump garbage in the container
Well	Well	Allow the container to be placed underground
	Safety measure	Prevent direct access to the well while it contains no container.

Table 1: Summary of container and well components.

container does not pose a safety risk. Generally, preventive maintenance is scheduled shortly ahead of inspections to ensure the container passes inspection. During preventive maintenance, small defects are corrected and maintenance such as lubrication of moving parts is performed. If any defects are encountered that cannot be fixed immediately, followup maintenance is planned to correct them before inspection. Such followup maintenance is unexpected and hence categorized as unplannable maintenance in this study.

- When a container is unusable due to damage or a danger to the immediate vicinity, **corrective** maintenance is scheduled.

Maintenance tasks are prioritized based on their impact on safety and the ability to collect garbage. Four priority levels are used: critical, high, normal, low. Critical issues denote issues where either (1) the container poses a direct danger to the vicinity (e.g. sharp edges, danger of tripping) or (2) the container is unable to process garbage (e.g. it is full or the insertion mechanism is broken). High priority issues are issues where it is likely that either of the conditions of a critical issue will soon appear. Normal and low priority issues need to be solved but do not represent any immediate problems or chance thereof. Normal priority tickets generally are followup tickets to preventive maintenance were a defect has been detected that does not directly influence use or safety, but might if left unattended. Low priority issues generally are issues that are not of a pressing nature. An example of a typical low priority issue is an inspection: while it has to be performed eventually, they are planned far ahead and often have long deadlines of weeks or months.

Some of the common critical issues that present an immediate danger to the vicinity or prevent the container from accepting waste are:

- The container is (partly) raised outside of its well in rest, creating a dangerous sharp edge around the pedestrian platform. This can be caused by a damaged safety mechanism which is unable to be retracted, or an obstruction in the well such as leaked garbage, or water in the well causing the container to float.
- The safety measure cannot be deployed properly, posing a danger during emptying.
- The safety measure cannot be retracted properly, preventing the container from being returned to the well.
- The garbage inlet is damaged or blocked, preventing garbage from being input.

4.3 Maintenance administration

Every maintenance action is recorded in an application called *Grybb*, operated by *Curious Inc.*, a service provider of the municipality. For every action, a ticket is created to track the progress of the action. If the action requires a followup action, a new ticket is created with a reference to the previous ticket. A ticket records the progress of work, the related assets (container, well), the required action, parts used and work expended, priority, type of maintenance, and any additional comments made by reporter of the ticket, servicemen, and administrators. A more comprehensive overview of the data associated with a ticket is provided in Section 5.2. While all maintenance related activities are administrated in *Grybb*, other activities and functionalities such as emptying the containers and administrating electronic container access are administered in other systems.

A typical inspection cycle for a container may for example result in three tickets: first preventive maintenance is scheduled, during which a larger defect is detected requiring a separate action, and finally the inspection itself is performed. All tickets refer to their predecessor allowing to identify all related tickets. Critical issues often result in two tickets: one in which the immediate danger to the environment is quickly mitigated (e.g. the area is cordoned of), and one in which the underlying issue is resolved.

4.3.1 Unplannable maintenance

In this research, unplanned or unexpected maintenance is defined as all specific work that cannot be reasonably expected one year ahead. As such, in the context of Amsterdam, corrective maintenance falls under this definition, but also any issues that cannot be fixed during preventive maintenance. Preventive maintenance itself does not fall under this definition: every container is maintained once every year and one can reasonably assume a small and known set of parts and work to be used.

Additionally, some tickets are marked as being part of a ”project”, a specific task that is applied to many containers. Historic examples of such projects are to apply reference stickers on containers or replace parts with manufacturing defects. Such projects are not considered to be reasonably predictable. As such, project-related tickets are ignored in the study.

4.4 Ticket life-cycle

All tickets go through a certain process between creation and being closed. Figure 4 outlines the steps a ticket commonly goes through. A ticket is created by an end-user, generally the municipality using the container/well and is assigned to a service company which can either accept the ticket or reject it. Generally, a service company accepts a ticket. Only if another service company is better suited, or the service company has no capacity to address the issue it will reject the ticket and optionally forward it to another service company. As soon as the service company accepted the ticket, it should assess whether the expected costs of a ticket are high (>€250). If the expected costs are high, the customer has to approve the intended solution before-hand, otherwise the service company may continue without explicit approval. The ticket then goes through a series of stages within the service process, a serviceman is assigned to the ticket and the ticket is planned for a timeslot. The serviceman executes the work and solves the issue, after which the applied solution is approved by a service manager. Finally, the customer approves the applied solution and associated costs. During these four stages, a ticket may be prematurely closed if for example the problem appears non-existent (i.e. due to a false/wrong report), the problem has a different root cause then expected and the client must approve the expected costs, or any other edge case occurs. Generally, when a ticket is prematurely closed and there is still a problem, a followup ticket is scheduled.

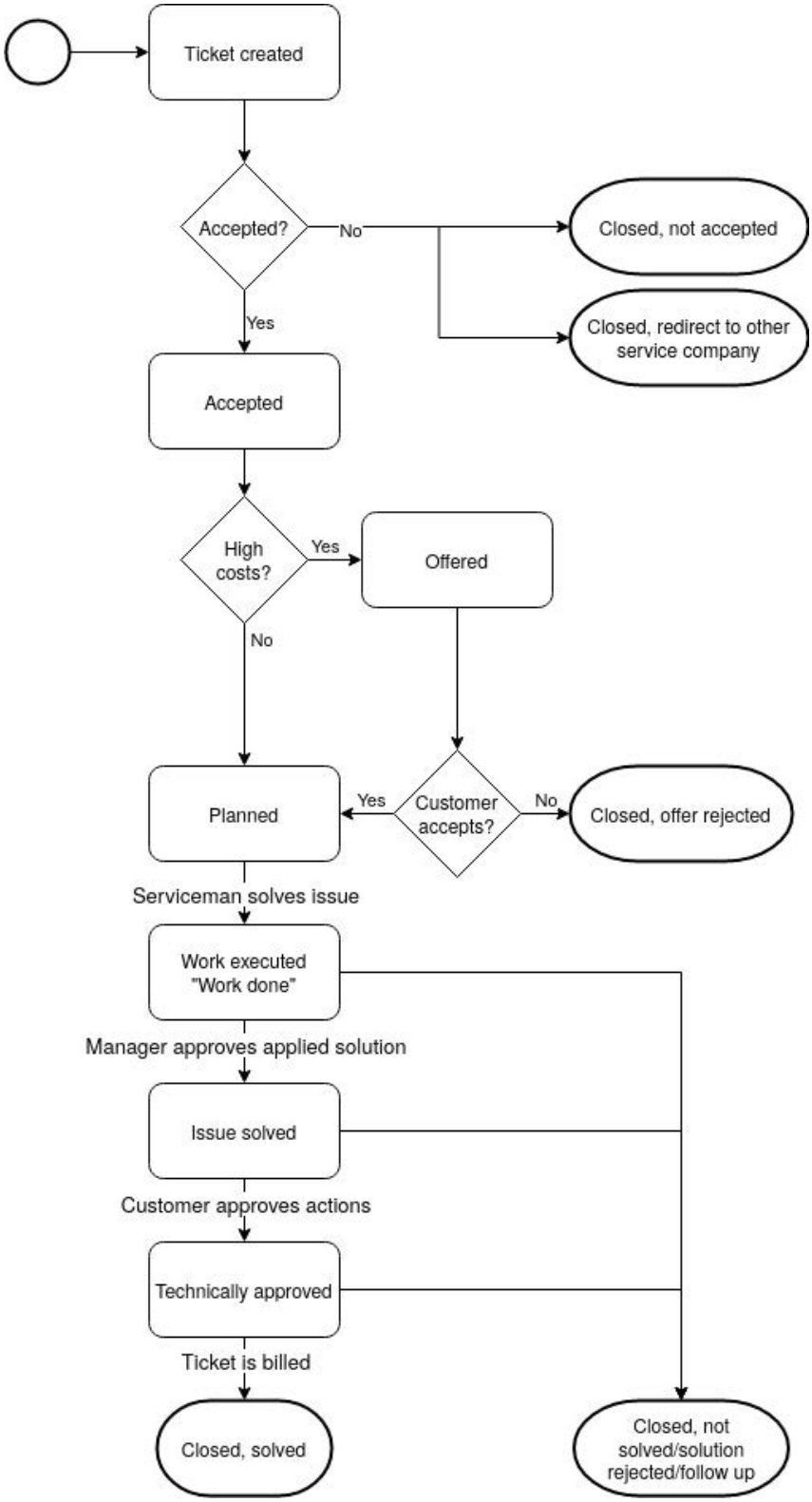


Figure 4: Overview of the states and transitions a ticket commonly goes through.

5 Case study: data

Relevant variables and models have been selected during literature research but not all data is available or relevant to the situation of the case study. This section identifies what data is available, why it is, or is not, used, and what data is unavailable.

The three available data sources that will be used for the prediction are as follows.

1. Containers and wells: detailing the assets that are maintained;
2. Maintenance: three years of maintenance records: 2017, 2018, 2019;
3. Demographics: demographic and geographic information of the municipality of Amsterdam.

Other data sources that are available but remain unused for the predictions are as follows. Section 8.1 discusses how these data sources may be used in the context of prediction maintenance on underground waste containers in future research.

1. Weather data: temperature and rain information is widely available. However, since the case study covers a relatively small area and the analyzed time-span is also relatively constrained (three years), weather is not expected to be discriminating enough.
2. Collection weights: data since 2017 is available that details when a certain fraction of waste from a "cluster" (see Section 4, first paragraph) is collected. It is impossible to consistently recover which exact container was emptied, and high inconsistency and error rate within the data has been encountered in previous research [9]. While an aggregation on neighbourhood or district level is feasible, the data data is left out due to its low quality and limited scope of this research.

The container and well, and maintenance history datasets are proprietary and based on an excerpt from the ticket management application *Grybb* made on August 5 2020. General demographics data is publicly available at Statistics Netherlands (CBS). For this study, a set that has been supplemented by Amsterdam is used which is publicly available [1]. In this research, the version of July 10 2020 is used. Both datasets are stored in a normalized format (in the context of database normalization), denormalization and transformation of the data is considered trivial and not discussed. Any changes to the contents of the data are discussed in Section 6.

The following subsections discuss each dataset in detail, Section 5.1 discusses the container and well data, Section 5.2 discusses maintenance records, and Section 5.3 demographics.

5.1 Containers and wells

Containers and wells have two properties in common: both have four dates that mark lifecycle events of the unit, and both can be described by a type detailing the make of the asset (i.e. parts used). Containers furthermore have a fraction associated with them, referring to either rest, plastics, bio, glass, paper, or textiles, and a well in which they are placed. Wells have a location in the form of longitude and latitude.

An asset can be marked inactive when it is no longer being used. This way any history pertaining to the asset is saved but the asset will not show up in operational reports. This generally happens when the asset is demolished if it has reached end of life.

A reference to all asset properties can be found in Table 2.

The dataset contains a total of 14 455 containers and 14 161 wells, while data of other municipalities is available, the data quality maintained by the municipality of Amsterdam is superior and makes the data eligible for analysis. Figure 5 shows the growth of containers and wells over recent years. It is clearly visible that the number of assets has grown steadily by approximately 1 000 containers and wells per year.

Field	Description	Example data
Container/well created at	Date of asset created in system	01-03-2016
Container/well delivery date	Date of asset delivered to customer	07-03-2016
Container/well placing date	Date of asset placed at location	10-03-2016
Container/well operational date	Date of asset entering service	10-03-2016
Container/well type	Make of the asset, detailing parts used	5m3 KHC Papier Amsterdam Standaard 1192*1192*2341
Container/well active	Whether the asset is in use	Yes/No
Container fraction	Type of waste handled by container	Glass
Container well	Well that holds the container	Reference to well by ID, e.g. 23412
Well location	Coordinates of well (latitude and longitude, 6 digit precision)	52.096487, 5.008091

Table 2: Summary of asset data available, with example data.

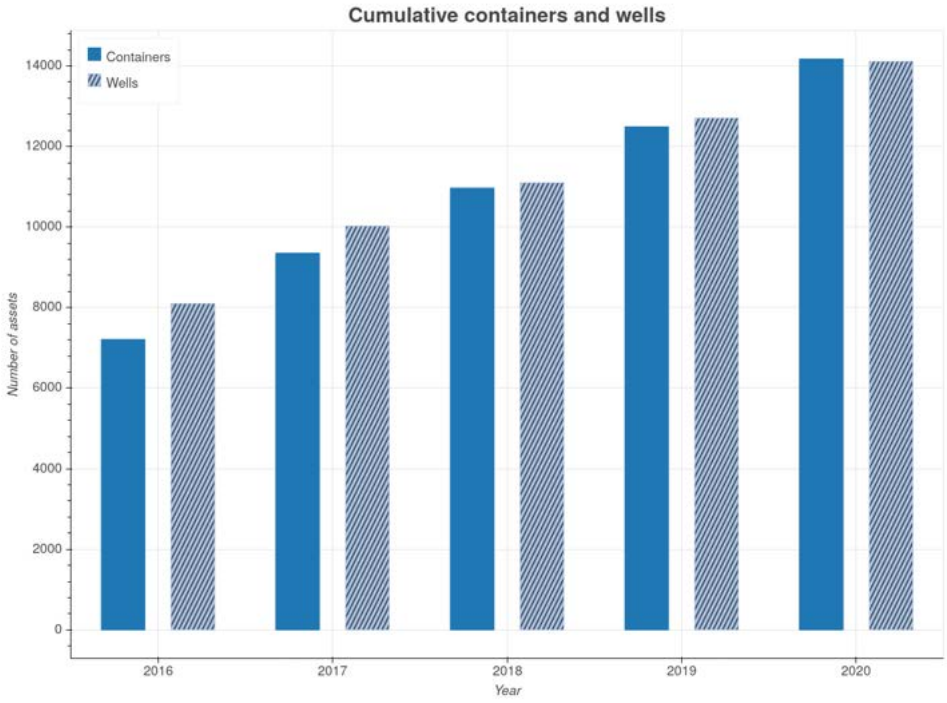


Figure 5: Bar chart showing the number of containers and wells in the municipality of Amsterdam per year. Data between the years 2001 and 2015 is omitted due to lack of relevance.

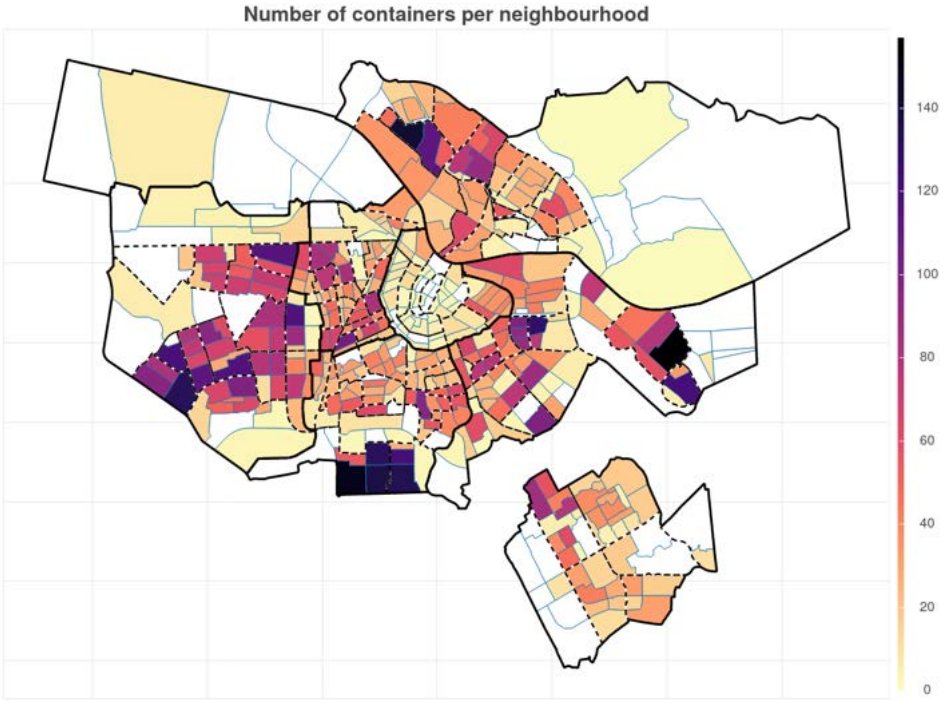


Figure 6: Heatmap of Amsterdam neighbourhoods showing how many containers are present in the neighbourhood. Districts are separated by black lines, wards by dashed lines, and neighbourhoods by blue lines.

Figure 6 shows a heat map that shows how many container each neighbourhood in Amsterdam has at the end of 2019. It shows similarity to the population density shown in Figure 2 and shows that some some residential areas have significantly more containers than others.

5.1.1 Container and well types

Each container and well is associated with a ”type” that provides information about the make and model of the asset. For every container type the insertion mechanism, hoisting mechanism, (empty) weight, and storage volume are administrated. Smaller details such as the form factor of the garbage storage are implicit to the model and not administrated explicitly. Well types are described by internal volume and dimensions, and the type of safety mechanism installed in the well. Table 3 shows a summary of the available data with examples.

Container and well type names are often overly specialized, making it non-trivial to extract common makes and models. The type names often include information not directly linked to the container type or details of the type that are also explicitly administrated. For example the container types ”5m3 KHC Papier Amsterdam Standaard 1380*1380*2690” and ”5m3 KHC Plastic Amsterdam Standaard 1380*1380*2690” contain: the volume of the type (5m3), the hoisting type (KHC refers to Kinshofer, a type of hoisting mechanism), the fraction (plastic), and the dimensions of the container. Volume and hoisting type are already separately administrated, while fraction is administered per container. The actual model of the container ”Amsterdam Standaard” is obscured by superfluous information in the name of the type.

In part due to the reasons explained above there are 210 different container types and 87 different well types in use. These are combinations of the previously explained properties.

Category	Field	Example data
Container type	Name	3TV1423R KH BEL 80 TROPL
	Hoisting mechanism	Kinshofer
	Insertion mechanism	Belfast
	Weight	550 (Kg)
	Volume	3 (m^3)
Well type	Name	1 delig 1670*1670*2750 KA-4 (Bammens, Amsterdam Standaard)
	Volume	5 (m^3)
	Dimensions	2750x1670x1670 (height, width, length, in mm)
	Safety type	Veiligheidsvloer zonder mangat

Table 3: Summary of asset type data available, with example data.

However, there are only 101 container types and 48 well types that are used more than 10 times. Figure 7 shows how many containers of each type exist at the end of 2019. It is clear from the figure that a few of the well and container types represent a large part of all the assets. Many (largely) unused container types have references to Amsterdams districts in the name, hinting that their lack of use may have been part of an administrative consolidation in which previously separately administrated container types were consolidated into centrally administrated ones.

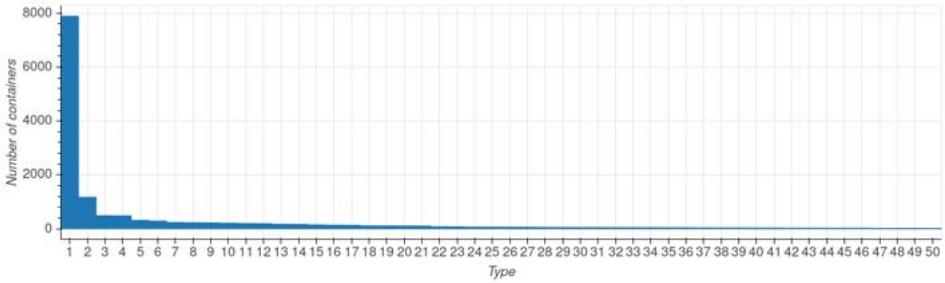


Figure 7: Number of containers of a given type for the 50 most occurring normalized container types, showing high usage of a very limited number of container types and a long tail of other types. Container type names are omitted for readability, the two most occurring types are "Kinshofer Amsterdam standard" and "Haaks metro".

5.2 Maintenance

Maintenance actions are recorded as tickets. A ticket signifies a single maintenance action taken on a container (and optionally the associated well). Table 4 shows a reference of all data available per ticket with example values and a short description of the field. Since *Grybb* is primarily an application that supports the daily work of service companies, the data contained in it is bound to contain a certain degree of noise. Not every edge case is covered by the application, and employees in the field may be forced to "misuse" a field to convey message that is otherwise not supported by the application. Alternatively, if a

situation is complex and not covered by the system, mechanics often fall back to a phone call to explain the situation, and put down a summary of the discussion in the ticket, which may not cover the full extent of the issue. Given the general scope of this study, this is not expected to have a significant impact on the experiment, but should be noted for future more in-depth studies.

Field	Description	Example data
ID	Unique identifier of the ticket	370048
Ticket type	e.g. corrective, preventive, ...	Corrective
Priority	e.g. emergency, high, medium, low	Emergency
Problem description	A free form text field describing extra details	Leegloper [1]
Problem modules	A selection of components of the container/well that are relevant to the problem	Other
State	Current state of the ticket, e.g. "new", "planned", "solved", ...	Approved by client
Created at	System time when the ticket was created	2019-12-16 07:41:24
Service company	Company assigned to solve the problem	ASW (Reiniging)
Ticket owner	Asset owner, responsible for creating the ticket	K Zuid [2]
Parent ticket	Reference to a preceding ticket	-
Work and parts	Consumed parts and spent time	Empty out well

Table 4: Summary of the relevant ticket data available, with example data.

To differentiate between the various types of action that can be taken, every ticket is assigned a category, various relevant ticket types are:

- Corrective
- Preventive
- Inspection
- Cleaning
- Project
- Repair as a result of preventive
- Repair as a result of inspection

A total of 19 types exist, ticket types irrelevant to this study have been omitted for brevity.

The *priority* of a ticket can either be emergency, high, medium, or low. Each priority has a specific reaction time associated with it which is governed by a Service License Agreement (SLA) between the ticket owner and service company. The *problem description* is a free text field that contains comments from the creator of the ticket or administrator. This field is generally used to specify the exact problem or reason of the problem. The *problem modules* field may refer to one or more components of the container or well that are related

[1] Rough translation: drained. Meaning the container contents have been accidentally dumped in the well.

[2] The name associated with city part Zuid in the application

to the issue. This field may help in strategic analysis of tickets, but is generally not specific enough to identify the exact problem, for which the problem description is used. Valid values of this field are the container and well components presented in Table 1.

The current state of the ticket shows what step in the process has been performed last, and which step is next. Generally, a ticket goes through the stages new, accepted by service company, planned, executed, closed, approved by customer, as discussed in Section 4.4. The time of each of these steps is logged for purposes of enforcing SLA and strategic analysis.

A ticket may serve as a followup to another ticket. For example the category "Repair as a result of preventive" is scheduled as followup to preventive maintenance if a defect was detected that could not be repaired at the time. In such case, the followup ticket will have a reference to the previous ticket so that it is clear that these tickets are related. This reference is referred to as the "parent ticket". It is only possible for a ticket to have a single parent, and having a parent is optional. However, every followup ticket itself can have a followup ticket again, allowing for ticket "chains" of arbitrary length. Such chains generally consist of 4 or less tickets; on a total of 133 864 chains (both plannable and non plannable), less than 500 consist of 5 or more tickets. The number of ticket chains with a given length are shown in Figure 8, it shows an roughly exponentially decreasing number of chains with increasing lengths (note the log scale of the number of chains).

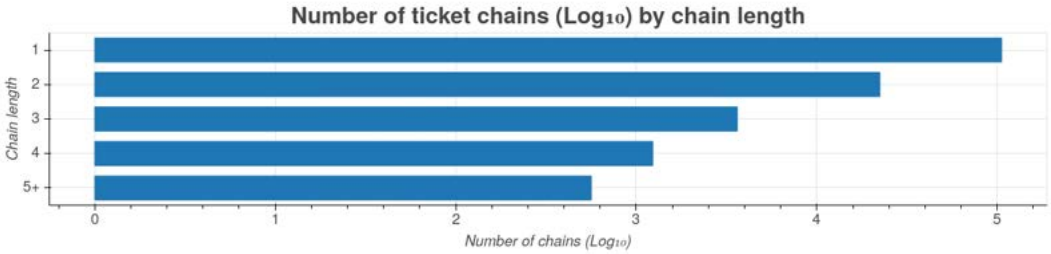


Figure 8: Number of ticket chains of a certain length in Log_{10} scale.

Work performed and parts used are often billed as a fixed item, for instance in the example presented in Table 4, "Empty out well" (last row) is a fixed work item with a fixed price. All parts have fixed prices. In the case where there is no specific work item to fit the work performed, a, standard work items for spent time (e.g. "15 minutes with/without truck") are available.

Between 2017 and 2019, 118 779 non plannable tickets were registered, Figure 9 shows the distribution of these tickets over the years and various categories. The number of tickets has rapidly increased over these years, almost doubling between 2017 and 2018 and more then doubling between 2018 and 2019. This trend is visible in the figure, however is not evenly distributed over the four major categories: periodic cleaning has relatively increased more when compared to the other 3 categories, the reason for this is unclear but is expected to be the result of increased adoption of the administrative system. Furthermore, the growth of tickets seems to be much higher than the growth in number of containers as shown in Figure 2. The cause of these increases is unknown but might be explained by factors considered in the case study experiment.

5.3 Demographics

Amsterdam publishes a general register with key figures of areas in the city with up to 788 variables per area known as the BBGA or Basis Bestand Gemeente Amsterdam (General Register Municipality Amsterdam) [1]. Areas are defined as districts, wards, neighbourhoods, and various alternative definitions. Districts are the largest areas, followed by wards and neighbourhoods. The 788 variables describe information on several themes, some of the useful themes for the purposes of this research are: population, age, diversity, activities,

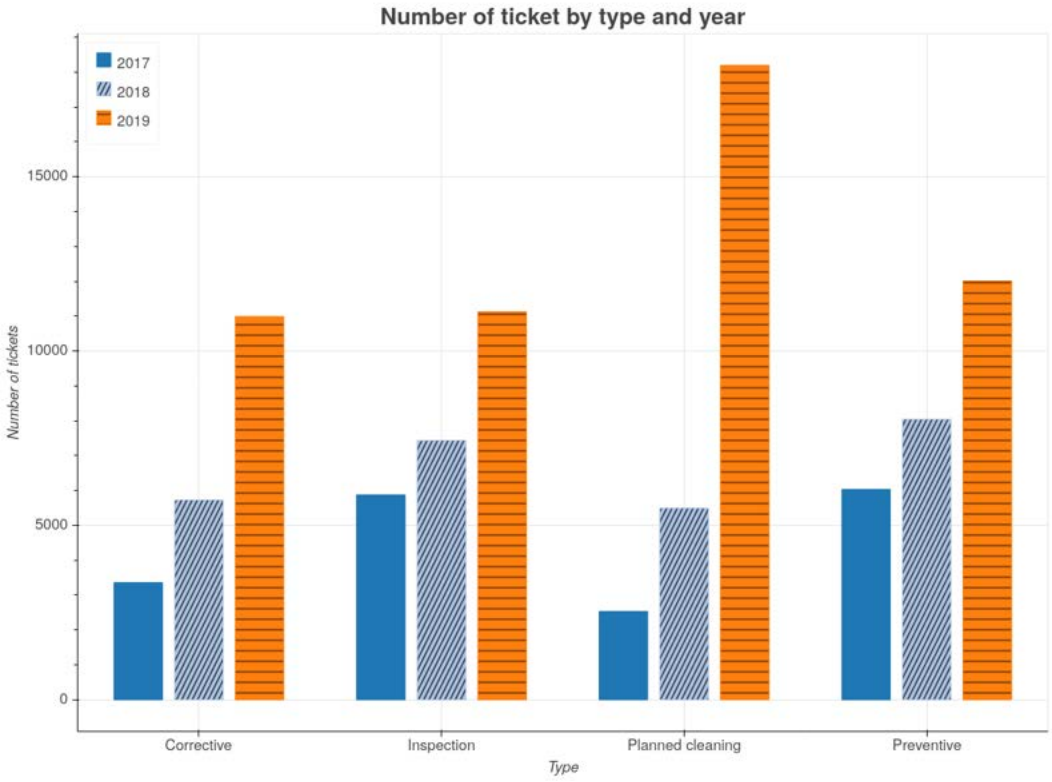


Figure 9: Distribution of all tickets in Amsterdam between 2017 and 2019 by category.

public space, education, work, income, and citizen participation. The case study experiment makes use of variables that report on the following subjects, based on the literature on predicting municipal solid waste: population, household details, income, residency type, and education.

A full list of used variables with examples is presented in Table 7, the category and original BBGA field name is shown along with a translation and explanation. An example of data of one specific variable in one year is shown in Table 5. It shows the spendable income in 2018 of two districts and the associated neighbourhoods. It shows reasonably consistency within neighbourhoods of a ward, and a general difference between wards; the spendable income in ward F86 is generally lower than that of ward K52, the spendable income of the associated neighbourhoods vary around the values for the wards.

Data is present for all districts, wards, and neighbourhoods for these variables with some exceptions. The percentage of missing data is detailed in Table 6 on the levels of ward and neighbourhood separately. Large amounts of missing data ($> 30\%$) in the income variables and education are caused by the fact that Statistics Netherlands (CBS) has not yet released these numbers for 2019 at the time of writing. Lower amounts of missing data ($\leq 5\%$) can be caused by either the sample size being too small to produce anonymous results, or the administrative area not existing at the time. This study assumes for simplicity sake that all administrative areas existing at the time of writing have not changed in the time considered.

[3]Calculation takes education and income in to account and produces a number between 2 and 10. Low scores are considered 2-4.

Area code	Spendable income
F86	35100
F86a	36900
F86b	33200
F86c	35500
F86d	42500
F86e	29000
F86f	25300

Area code	Spendable income
K52	48600
K52a	56500
K52b	45900
K52c	48600
K52d	49000
K52g	34800
K52h	65800

Table 5: Spendable income of two wards and accompanying neighbourhoods, showing the differences in demographics variables between administrative areas. Note that the first rows denote the wards themselves and subsequent rows the neighbourhoods.

	Missing data	
Category	Wards	Neighbourhoods
Population	1%	3%
Household	1%	6%
Income	34%	46%
Residency	1%	11%
Education	34%	43%

Table 6: Missing data per category on ward and neighbourhood levels.

Category	BBGA Field	Field description	Example data
-	-	Area code	A05c
-	-	Year	2018
Population	bevtotaal	Total population	2492
	bev0_4	Population aged 0-4 (incl.)	95
	bev5_9	Population aged 5-9 (incl.)	87
	...	...	...
	bev85_89	Population aged 85-89 (incl.)	23
Household	beveenouderhh_P	Single parent households %	116
	bevalleenhh_p	Single person households %	59%
	bevpaarzkindhh_p	Households without children %	21.7%
	bevpaarmkindhh_p	Households with children %	10.4%
	bevoverighh_p	Other households %	1.1%
Income	ihhink_gem	Average spendable income	40 200
	iinkq1_p	Household % in 1st income quantile	31%
	iinkq2_p	Household % in 2nd income quantile	21%
	iinkq3_p	Household % in 3rd income quantile	15%
	iinkq4_p	Household % in 4th income quantile	14%
	iinkq5_p	Household % in 5th income quantile	18%
Residency	wbezet	Average number of residency inhabitants	1.62
	wcorhuur_p	Residencies owned by housing corporation %	42.6%
	wparthuur_p	Residencies privately rented %	34%
	wkoop_p	Residencies privately owned %	23.4%
Education	bevopllaag_p	Population low education %	17%
	bevoplmid_p	Population medium education %	32%
	bevoplhoog_p	Population high education %	51%
	skses234_p	Population with low socio-economic score [3]	30%
	skses_gem	Average socio-economic score [3]	6.4

Table 7: Overview of demographics data used, with example data for neighbourhood "A05c" in the year 2018.

6 Case study: prediction

To test the models that were selected in the literature review, the available data is used to apply maintenance forecasting the case study: underground waste containers in Amsterdam. This section breaks down the process of building the model and discusses extracting *knowledge* from the data that may prove useful in satisfying Amsterdams goals: reducing the number of tickets.

To extract patterns from the available data, a process similar to the Knowledge Discover in Databases (KDD) method outlined in [8] is used. In KDD the final product gained after extracting patterns is termed *knowledge*, it represents the answers to the research questions in this study. First, data is selected from the raw sources. Secondly, the selected data is pre-processed to filter out errors in the data and fill missing data. Then, it is transformed into a format that allows for easy analysis. Finally it is mined for patterns to extract knowledge to answer the research questions. This process is outlined in Figure 10.

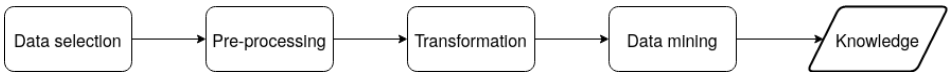


Figure 10: High level overview of the methodology, showing the process steps resulting in knowledge.

In this section, all process steps are discussed in detail. Before data selection, the target variable of the prediction is identified, such that the goal of the methodology is clear.

6.1 Target variable

The target of the prediction is the number of unplanned maintenance issues. In the case of Amsterdam and in the dataset, this value is represented by the number of unplannable tickets. Unplannable tickets are defined as those tickets that are of ticket type:

- Corrective
- Repair as a result of inspection
- Cleaning (as opposed to a ”Planned cleaning”, ”regular” cleaning is not planned)
- Repair as a result of preventive

The sum of all tickets that match this description is then the target variable. This sum can be calculated per year and area. Amsterdam is divided into administrative areas on three levels: districts, wards, and neighbourhoods. Any of these three area levels may be a suitable scope for the prediction. However, since there are only seven districts, predicting the target variable per district may lead to results that are not specific enough. Since demographics input data is only available per year, and there are three years of data available, a prediction horizon of a year is considered the most reasonable.

6.2 Data selection

To select data from the available data, information that may be relevant to the predicted variable is identified. For some data relevance is identified based on evidence from domain experts or based on literature. For other data the expected relevance is not grounded in existing research or expertise, in which case it is argued why the data is assumed to be relevant to the target variable. The selected information can be categorized in two categories: asset data and demographics. The selected data, that can subsequently be transformed into input variables to a predictive model is as follows:

1. Asset data:

   (a) **Asset count.** The number of assets is expected to influence the required maintenance; more assets leads to more (unplanned) maintenance.

   (b) **Asset age.** The asset age is shown by literature to give an indication of the "wear and tear" of an asset. Higher wear and tear is in turn expected to result in a higher likelihood of damage, causing more unplannable tickets.

   (c) **Asset type.** The asset type described the components of the asset. Literature has shown that various types of assets show different maintenance behaviour. Furthermore, domain experts have specifically indicated that certain asset types are more likely to suffer from unplannable maintenance. One example in this case is a container type that compresses garbage before storing it, these containers suffer from much more unexpected breakdowns than other types. It is assumed that similar trends may exist based on the example given. Certain types of assets or components may be more error-prone than other components resulting in more unplannable tickets than others. Asset types are identified by four types of data: the **container/well type** provides a single reference to the make and model of an asset while the associated **insertion type**, **hoisting type**, and **safety measure type** reference types of specific parts of the assets. Finally, to effectively capture the example of compression containers requiring more maintenance than other types, one extra feature is constructed from the asset type: the number of **compressing containers**. The feature is the sum of all containers of a type with the text "pers" (Dutch for "compression") in its name. These types are then discarded as separate asset type features.

   (d) **Waste fraction served.** An example has been presented by domain experts in which the fraction of the waste influences the way the container is emptied. Specifically, paper containers have been used in some cases to attempt to compress the waste in the truck it was emptied in with (possible) damages to the container (and truck) as a result. Other such cases may or may not exist and may be identified by this data.

2. Demographic data:

   (a) **Population count.** The number of inhabitants of an area may serve as a proxy to total usage of containers in an area. It is expected that heavily used containers are more likely to be the subject of unplanned maintenance. To quantify this, the population count per container can be calculated. Furthermore, it is identified in literature as a relevant factor in MSW generation.

   (b) **Household type**, **Income**, **Residency type**, **Education.** Demographic variables have been shown in literature (see Section 3) to influence the generation of waste and degree of separation into various waste streams. The amount of waste per stream generated may influence usage of the container and hence required maintenance.

This data must subsequently be pre-processed and transformed into usable input variables.

6.3 Pre-processing

To provide consistent input values that reflect reality, the data must first be filtered to remove errors or noise, and be corrected to fill missing values. It can then be transformed to be used as input variables in a predictive model. The filtering, correction, and transformation steps are as follows. Each step is discussed in further detail in the following sections.

1. Data filtering:
   (a) Wells that are used solely for administrative purposes are removed;
   (b) Plannable tickets and non-solved tickets are removed;
   (c) Project related tickets are removed.
2. Data correction:
   (a) Container names are stripped of extraneous details;
   (b) Missing asset lifecycle dates are filled;
   (c) Missing demographic values are linearly interpolated;
   (d) Chains of tickets are reduced to single tickets.
3. Data transformation:
   (a) Tickets are assigned the location of their subject well;
   (b) Ticket and asset creation dates and location are used to produce values per year/area combination.
   (c) Features with absolute values are scaled (0-1) within categories, such that it is possible to compare absolute values between categories at a glance. Percentage values are converted to a 0-1 scale if necessary.

6.3.1 Filtering

Two specific wells and their associated containers are removed from the dataset: the "Algemene order" well (ID 6271) and container (ID 6272), and the wells located at "Papaverweg 33" (IDs 59220 and 71935). The first well is used to book costs that cannot be otherwise administrated in the application. The second set of wells is located at a yard of the municipality and do not actually house containers. Containers are brought to the yard to be fixed or demolished and may remain here for unknown reasons. This action is purely administrative and not related to the actual functioning of containers in the municipality, as such they do not contribute to the dataset and are removed.

Since the only concern is to identify patterns in the non plannable tickets, all plannable tickets are removed from the dataset as defined in Section 6.1. Furthermore, tickets that are not solved are removed. Possible non-closed states may be that the ticket has been forwarded to another service company (and subsequently a new ticket is created), the ticket is intentionally not solved (for example because the price was too high, or the issue deemed irrelevant), or the ticket was never solved for reasons unknown.

Some of the tickets that are of a ticket type that is considered non-plannable are part of large projects. An effort is made to remove these types of tickets from the dataset since they do no represent incidents that requires acute maintenance. Instead they represent structural defects in past processes or manufacturing, or changes in policy. Two such projects are identified in the data: (1) (re)placing identification stickers on containers and (2) a systematic inspection of all components of a specific type. To remove ticket associated with these project, any ticket that contains the text "sticker", "referentie", "nummeren", or "ketting controle" is removed. The first three terms often identify projects to (re-)apply identification stickers, the last term is related to a systematic check of certain chains in wells. These terms are manually identified from observations of the data, no concrete methodology has been applied in identifying these terms.

6.3.2 Corrections

The names of container and well types are stripped of details that are already associated with the assets or their types in an effort to consistently reduce containers to their type. To build on the example of such data given in Section 5.1.1: after this normalization, all containers of the types "5m3 KHC Papier Amsterdam Standaard 1192*1192*2341" and "5m3 KHC Glas Amsterdam Standaard 1192*1192*2341" (difference in "Papier" and "Glas") can be matched to the type "Amsterdam Standaard" while the volume, hoisting type, fraction, and dimensions are still retained in the asset and type.

Assets record the dates at which they are declared operational, placed, delivered, and created in the system. However, the operational, placed, and delivered dates are not always filled in. Values are forward-filled in the following sequence: created in the system, delivered, placed, operational. Meaning that if a value is missing, the last known value is used. See Table 8 for a (fictional) example of this method. This method is judged to be sufficient by domain experts for the purposes of this study.

Field	Value
Operational date	n/a
Placed date	01-02-2017
Delivery date	n/a
Created date	01-01-2017

Field	Value
Operational date	**01-02-2017**
Placed date	01-02-2017
Delivery date	**01-01-2017**
Created date	01-01-2017

Table 8: Fictional example of forward filling of various lifecycle dates of assets. The table of the left shows the original data with missing fields, while the table on the right shows the corrected data with forward filled values in bold.

Missing demographics variables are interpolated by linear regression. The most notable missing values are in variables that are missing data for the entire year 2019 (as discussed in Section 5.3). Values for a certain variable in a given area often display a trend over the years which must be preserved. To that end a linear regression is applied to fill in the missing values based on available data. No extensive testing is performed on this linear interpolation, but domain experts agree that this method should be sufficient for the purposes of this research. Alternatives considered but not implemented are (1) forward filling values (basing missing values on the last known value), and (2) shifting values forwards such that observations for 2016 are used for 2017, observations from 2017 for 2018, and observations from 2018 for 2019. However these methods do not interpolate the trend of the data, or fail if more than 1 year of data is missing.

Since chains of tickets describe only a single problem and multiple tickets are only created for administrative purposes, the chains are squashed to single tickets. Figure 8 roughly shows how many chains of certain lengths exist. When only considering the non plannable tickets, 5 449 chains exist of length 2 or higher. These are reduced to single tickets by taking the values of the last ticket in the chain, the priority of the first ticket in the chain (this reflects the priority of the original issue), the creation date of the first ticket, and an aggregation of the ticket descriptions and problem modules.

6.3.3 Transformation

To count the number of tickets in an area, a ticket has to be assigned to an area. To this effect, the location of the subject asset of the ticket is assigned to the ticket. If the only subject is a container, the location of the well housing the container is used.

Up to this point, both tickets and assets are associated with an exact point in space in the form of a coordinate. However, all data must still be transformed such that every desired variable produces a single value for every area/year combination. Administrative

areas are described by bounded polygons, which can used to resolve a coordinate to the area. For tickets, the creation date is used to place the ticket in a year, and the ticket location is used to place the ticket in an administrative area. To count the existing containers and wells in an area, the operational date of the asset is used to place the asset in a year, and the well location to place the asset in an area. Counting the number of assets of a certain type or fraction is done in a similar fashion, but constrained to the type that is counted. To summarize the ages of asset, the descriptive statistics minimum, maximum, average, and standard deviation are calculated of the asset ages within a year/area combination.

A full overview of all available input variables (including the target variable "ticket count") with example values for a neighbourhood level prediction is shown in Appendix A.

Features with absolute values and the target variable are scaled on a scale of 0-1 so the features importance in the linear regression model can be better interpreted. Since absolute values within categories are expressed in the same unit (e.g. number of containers, age in days, population count, income), scaling values within categories accurately maintains the relative value of a feature. By scaling the feature, the linear coefficient does not need to correct for the order of magnitude of the feature, and can therefore be better compared between categories and prediction levels. An example of a difference in order of magnitude is between population (10^4) and average income (10^5). Percentage values are already ensured to be on a 0-1 scale.

6.4 Predictive models

Literature reports success with various models in the prediction of maintenance and MSW generation: statistical models such as Weibull and Poison distributions are used to predict the likelihood of failure within a given timespan, and regression models such as single- and multiple linear regression, support vector machine, and neural networks are used to predict the generated MSW of an area. The target model must be able to handle multiple input variables and clearly show the impact of these variables on their own. The best fit for these requirements in the multiple linear regression, which has shown success in predicting MSW generation. It provides an interpretation of the data that is easily interpreted by domain experts and clearly shows the importance of variables.

In addition to the multiple linear regression, a reference prediction is made using Random Forest regression. It may serve to identify the prediction potential by using non-linear relations in the data. However, depending on the decision tree size it may be more complex to explain the model to domain experts and the model is more likely to overfit than the linear regression. Due to these reasons, the main focus of the research will be predicting the target variable using a linear regression.

The models are generated using the LinearRegression and RandomForestRegressor modules of the Python programming language library scikit-learn [19] version 0.23.2. The linear regression is implemented using Ordinary Least Squares (OLS). The random forest model has plenty of hyper-parameters to influence the way the model functions. To prevent overfitting, the max depth of the random forest is limited to 4 levels, and the minimal samples of a leaf is set at 3.5% of the sample size, and the number of estimators in the model is set to 30. These values have been determined using an exhaustive grid search for values between 1 and 10, 0.01% and 10%, and 10 and 150 respectively on the complete ward dataset.

6.5 Experiment: ticket prediction

To identify meaningful factors that may predict the number of unplannable tickets in an area in a given year, the discussed models are used to predict the number of tickets based on the identified input variables. A baseline model is constructed that predicts the target variable with only the number of containers as input using linear regression. This is an intuitive model that translates to a certain percentage of assets having a number of issues per year. Figure 11 shows the number of tickets and number of assets in several selected

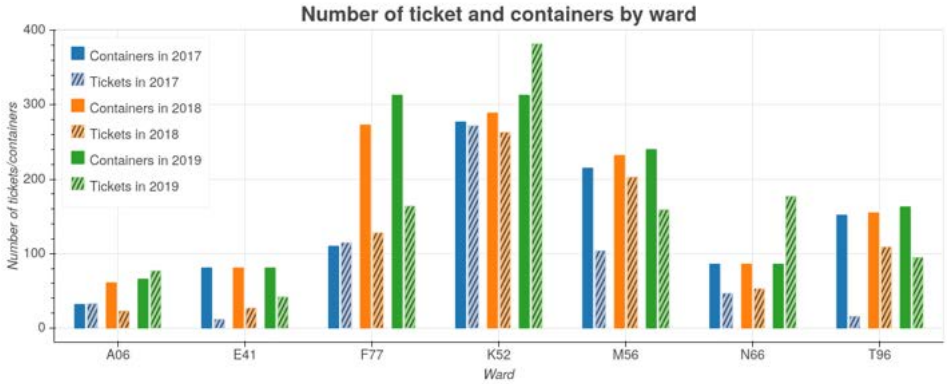


Figure 11: Number of assets and number of tickets for selected wards, showing the relation between number of assets and tickets.

wards. It can be gathered from the figure that wards with higher number of assets generally have more tickets than wards with lower numbers of assets. It is also clear from the figure that this relation does not always hold, nor is the general number of tickets per container always the same.

Following the construction of a baseline model, some steps are taken to arrive at a final model. The steps are designed to reduce the high number of features to an explainable and good performing set.

1. For each category of predictors:
   (a) Collinear predictors are eliminated to satisfy model assumptions, the elimination process is described in Section 6.5.1.
   (b) A linear model is constructed using the remaining predictors of the category and those of the baseline.
   (c) If the performance of the model is better than the performance of the baseline in terms of R^2_{adj}, the category of predictors is marked ”meaningful”, i.e. it is meaningful to the target variable.
2. A feature set of all remaining features of the meaningful categories is created, multi-collinear features are again eliminated using the process discussed in Section 6.5.1.
3. A final linear and random forest model is created using the remaining feature set.

This process ensures that the features are not strongly dependent of each other. A high degree of multicollinearity **within** the categories is to be expected since they describe similar concept, while a medium degree of multicollinearity **between** categories is expected, since many social factors still correlate with each other, for example wealth and education [21]. Among collinear features, the features with the highest predictive value are kept.

The final models with the remaining features are validated and used for final results. Based on these results, manual alterations can be made to the model to further optimize it in terms of simplicity and performance.

6.5.1 Discarding collinear features

To eliminate multi-collinear features from a given feature set, an iterative process is used. Given a set of features, the variance inflation factor (VIF) is calculated for each feature. For all features with a VIF higher than the threshold of 4, a linear model is constructed to predicts the target variable using only the specific feature. The feature model with the

lowest R^2 is eliminated, and the process repeated. The process stops when there are no features with a VIF over 4, or when there is a single feature left. This process ensures features are not multi-collinear while retaining the most valuable features

Once this process is completed, a large part of the features are eliminated, but features may still be present that only marginally improve the results of the model. Such features can be removed manually. By removing features that provide only a small improvements to the performance of the model, the model is improved in terms of explainability and generalizability.

6.6 Model evaluation

Model evaluation consists of two steps:

1. The use of the model is **validated**: the linear regression model makes assumptions about the data, if these assumptions are not valid, results of the model may be overly optimistic or pessimistic;

2. The **performance** of the model is evaluated so it is clear how well the model predicts the target variable and explains the data.

To properly evaluate the model, the dataset has to be split into a training dataset and an validation dataset. The model is trained on the former and is evaluated using the latter. Care must be taken to prevent dependency between these two sets. In the case of predicting maintenance occurrences for city areas it is reasonable to assume that a given area performs similar over time. To prevent the model from learning the characteristics of specific areas and reliably evaluate the model, the training and validation sets must be representative for the entire dataset. Since the dataset exists of data over three years, it is possible to split the dataset on these years. Two thirds is then used for training purposes, while the final third is used for validation. Three-fold cross-validation is then applied: training and validating the model with each possible combination. Evaluation can then be performed over the three folds.

6.6.1 Model assumptions

Familiarity with the linear regression model is assumed, however the required assumptions for the model to be valid are discussed in order to clearly discuss the required steps in model validation.

In general, linear regression models build on several assumptions in order for the expected value of the predicted variable to be reliable. Given a linear model of the form $y = X\beta + \epsilon$, these assumptions are [11]:

1. **Weak exogeneity**. A linear regression is assumed to model causation of the predictor variables x and error term ϵ on predicted variable Y. Therefore, predicted variable Y must not cause X. Also, since ϵ in part causes Y, ϵ must also not cause X. By assuming exogeneity of X, we assume that the predictor variables x are fixed values imposed from outside the model, and not caused by the model itself.

   This assumption also means that X is assumed to be free of any (measurement) errors. Only the error term ϵ accounts for any errors in the dependent variable Y.

2. **Linearity.** The predicted variable must be a linear combination of parameters and predicting variables. However, since it is possible to transform the predicting variables in any way before applying the regression, this assumption only restricts the model parameters β.

3. **Homoscedasticity.** The variance of the errors in the predicted variable is constant over the entire range for predictor variables. This assumption is for example violated when the variance of the errors for large predicted values is larger than that of small predicted values.

   If this assumption is violated and the variance of errors varies based on the predictor variables' values, the variance can not be accurately modeled for the entire range of the data. This can result in under- or overestimating the importance of a residual since it is unclear if the residual can be caused by the variation or not.

4. **Independence of errors.** The errors of the predicted variable should be uncorrelated with each other. An example of such dependence in this context is spatial autocorrelation: errors in a certain area may be higher than other areas. If the errors are not random, the model does not fully capture all relations.

5. **Lack of perfect multicollinearity.** Predictor variables should not be perfectly correlated with each other. For example, this can happen if a variable and a linear transformation of this variable are present in the dataset.

6.6.2 Model validation

To confirm the assumptions made by the linear regression model, the following actions are taken once a model has been constructed:

1. For each predictor variable: reason that the predicted variable is unlikely to cause an influence in the predictor.

2. Plot the residual against the location and time based data. If a pattern emerges, the errors are likely (spatially) autocorrelated, which means they are not independent.

3. Plot the residuals against the predictors to identify if the variance of the errors is constant.

4. Test multicollinearity using Variance Inflation Factors (VIF). Literature shows a common upper threshold of 4 or 5. If the value is above this threshold, serious multicollinearity is present.

Furthermore, normality of errors is specifically desired when employing the common Ordinary Least Squares (OLS) linear regression. Normality of errors can be tested using a QQ plot. The random forest regression makes little assumptions about the data, however normality of errors is desired because it implies the errors can not be predicted any more by available data.

6.6.3 Evaluation criteria

Several common measures exist to evaluate model performance:

1. Mean error (ME) is calculated by taking the mean of the residuals. It can show whether the model systematically over- or underpredicts values, i.e. has a bias. It places no emphasis on outliers which may skew the result.

2. Mean absolute error (MAE) is comparable to ME, however it takes the mean of the absolute values of the residuals. It can therefore show how large the average error is.

3. Root mean square error (RMSE) is calculated by taking the square of the residuals and then taking the root of the mean of that. By squaring the residuals, it is more sensitive to outliers than ME or MEA and does not indicate the direction of the error.

4. Coefficient of determination, denoted by R^2, expresses the percentage of variance in the predicted variable that is explained by the linear model. When using multiple variables, the adjusted R-squared $\bar{R}^2$ is recommended, R^2 and $\bar{R}^2$ are formulated as follows:

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

The ME identifies a systematic bias in the model, while the MAE and RMSE give an indication as to the size of the errors. R^2_{adj} shows how much of the variance in the data is identified by the model. All metrics carry value for judging the performance of the produced models. These evaluation criteria can be used for both linear regression as well as random forest regression since they only deal with the prediction and target value.

6.7 Feature importance

Finally, to determine which factors are important to the prediction, the feature importance is determined for all features of the final model. In linear regression, the features are assumed to be independent and the impact of the feature on the prediction can be determined by taking the coefficient of the feature, or β_i. A large absolute value of a coefficient implies that the feature has a large impact on the target variable. Feature importance for the random forest can be measured using the Gini index.

7 Case study: results

In this section, the results of data pre-processing, the experiment, model validation, and feature importance are presented. With regards to pre-processing, notable changes in size of the dataset and contents of the data are reported. Construction of the final model and its features and performance are discussed, model validation is performed and feature importance reported.

7.1 Data pre-processing

Data pre-processing has lowered the number of tickets and changed the values in the data. This subsection presents the impact of the pre-processing.

7.1.1 Filtering

Filtering has decreased the number of tickets per year and category. While removing the non-plannable tickets is part of the filtering process, number of tickets after filtering are also reported for the non-plannable tickets for reference.

After applying filtering, the number of tickets for analysis is reduced from 178 554 (all tickets created by Amsterdam on assets owned by Amsterdam created in 2017, 2018, or 2019) to 114 558, of which 32 446 are non plannable. Table 9 shows a breakdown of the number of tickets per year and category before and after processing. Around 40 000 tickets are discarded due to sharing a single root ticket of which roughly 6 000 tickets are non plannable. Furthermore, approximately 5 500 tickets are removed due to being likely part of a project of which 2 500 are non plannable.

	2017		2018		2019	
Category	Before	After	Before	After	Before	After
Corrective	3 632	2 898	6 180	4 539	11 653	9 297
Cleaning	206	197	2 712	2 590	7 244	7 182
Repair after preventive	578	298	3 577	1 964	1 813	954
Repair after inspection	1 810	1 057	704	378	2 128	1092
Other (plannable)	23 510	14 752	34 341	21 078	69 466	46 282
Total non-plannable	6 226	4 450	13 173	9 471	22 838	18 525
Total	38 736	19 202	47 514	30 549	92 304	64 807

Table 9: Number of tickets per category and year before and after pre-processing.

A trend of increasing number of tickets over the years is visible both before and after filtering.

7.1.2 Data correction

The number of distinct container types is reduced from 211 to 155 by normalizing container types names. Of the available 28 000 assets (wells and containers), 16 000 are missing either the operational, placement, or delivery date which is subsequently forward filled. As an indication of the impact of this imputation: for those assets that do have the operational date and created date filled (process wise these dates are the last and first dates recorded respectively), the 90^{th} percentile of the difference in these dates is 223 days. This means that the difference between these dates is generally quite large, and forward filling these dates may have a large impact as an asset is likely to be placed in the wrong year.

The number of missing demographic values is discussed in Section 5.3. With exception of one ward in the ward dataset and 47 neighbourhoods in the neighbourhood dataset, all missing values are amended through the use of linear regression. Ward IJburg Oost (M50) has too much missing data to interpolate missing values. Furthermore, 3 years of data for 46 neighbourhoods and 1 year for the remaining neighbourhood can not be interpolated. The result is that 3 records in the ward dataset and 139 records in the neighbourhood dataset have to be discarded due to missing data.

7.1.3 Data transformation

Data is transformed to have one value for each feature per year and area combination. Two variants are produced: one on the resolution of neighbourhoods and another on the resolution of wards. The former exists of 1 193 datapoints, while the latter exists of 292 datapoints. An example of a full record of the neighbourhood variant is shown in Appendix A. The data consists of 227 features, the distribution of features over categories is shown in Table 10 in the column "Initial". The other columns show how many features in these categories remain after applying two iterations of feature selection as discussed in the methodology, these values are further discussed in the next subsection. It should be noted about the feature set that the container, insertion, and hoisting type features are very sparse arrays, given the fact that most areas have a very low number of different types of containers.

		Ward		Neighbourhood	
Category	Initial	Intermediary	Final	Intermediary	Final
Existing containers	1	1	1	1	1
Asset meta information	4	3	2	3	2
Asset type	175	-	-	-	-
Compression containers	1	1	1	1	1
Hoisting type	5	5	4	5	4
Insertion type	15	-	-	-	-
Waste fraction	7	-	-	-	-
Demographics	39	10	-	1	-
Population	19	1	-	1	-
Household	5	-	-	-	-
Income	6	1	-	-	-
Residency	4	3	-	-	-
Education	5	2	-	-	-

Table 10: Number of features in each category in the initial dataset, after intermediary elimination of features based on multi-collinearity within categories, and after elimination of features for the final model based on multi-collinearity between categories.

7.2 Experimental results

After data pre-processing, the experiment is performed: the feature set is narrowed down based on the predictive power and multi-colinearity of features, a model is created and its performance evaluated. The experimental results consist of the following parts, all parts are discussed in the following sections:

1. Baseline model performance: linear and random forest;
2. Meaningful category selection;

3. Final model performance and validation;

4. Final model feature set and importance.

7.2.1 Baseline model

The baseline model is constructed to predict the number of tickets solely based on the number of containers in an area. Model performance for the baseline and final model for both ward- and neighbourhood level prediction are shown in Table 12. These results are discussed in results discussed in Section 7.2.3. In both the ward and neighbourhood models, the linear model outperforms the random forest model.

7.2.2 Meaningful categories

The number of features has been heavily reduced by eliminating multi-collinear features within categories. As shown in Table 10 after initial collinearity elimination and removing categories that do not improve the baseline model, the total number of remaining features is 20 for ward level models, and 11 for neighbourhood level models. These features are distributed over the following categories that have shown to improve the baseline model: asset age information, hoisting type, population, income, residency, education.

Table 11 details the performance of the intermediary models that combine the baseline model with all non-collinear features of each category. A notable result is the excessively bad scores of the model that has been augmented with the type features. This likely a result of the high number of features combined with the sparseness of these features, even after eliminating multi-collinear columns within the category, it still holds 55 features. The model is expected to overfit during training because of the high number of very specific features.

The variables are combined into a single set per prediction level. After eliminating multi-collinear variables within these sets, all demographic variables are eliminated and the number of variables in the remaining categories is reduced. Only one difference remains between the ward and neighbourhood models: while the ward-based model kept the minimum and *average* asset age, the neighbourhood-based model kept the minimum and *maximum* asset age. The final feature sets for the ward- and neighbourhood-based models are shown in Table 14 along with the importance of each feature. Feature importance is further discussed in Section 7.2.4.

7.2.3 Model evaluation

The final models produced by the proposed methodology are those with the non-collinear feature sets as described in Table 14. This subsection discusses the performance and validity of the produced models. The scaled results of the final models are presented alongside their baseline equivalents in Table 12, which allows for comparison of ward and neighbourhood performance. The absolute results of the final models are shown in Table 13, which allow for interpreting the errors in terms of number of tickets. The final models are subsequently manually improved, the results of which are shown in Table 15 (scaled performance), Table 16 (absolute performance), and Table 17. The steps taken to improve the models and discussion of their results is discussed in Section 7.2.5. The results show that the final models outperform their baseline equivalents, but still show a relatively low R^2_{adj} value, and relatively high errors.

To validate the models, the steps outlined in Section 6.6.2 are performed.

1.
   - It is unlikely that the number of tickets has a direct effect on the number of containers with a given hoisting type; no decisions to place containers with specific hoisting types have been known to be made based on the number of tickets.

Prediction level	Category	ME	MAE	RMSE	R^2_{adj}
Ward	**Age**	0.003	0.072	0.118	0.407
	Compression	0.001	0.078	0.125	0.340
	Hoisting type	0.001	0.078	0.125	0.322
	Population	0.000	0.080	0.128	0.299
	Income	0.001	0.082	0.130	0.281
	Education	0.001	0.082	0.130	0.277
	Housing	0.000	0.083	0.132	0.260
	Insertion type	0.004	0.084	0.130	0.240
	Fraction	0.004	0.086	0.134	0.230
	Type	6.943e+10	6.943e+10	3.884e+11	-7.919e+24
Neighbourhood	**Hoisting type**	0.001	0.065	0.099	0.317
	Compression	0.001	0.066	0.101	0.294
	Age	0.000	0.066	0.101	0.284
	Income	0.000	0.070	0.105	0.238
	Education	0.000	0.070	0.105	0.236
	Population	-0.001	0.070	0.105	0.236
	Insertion type	0.003	0.068	0.105	0.228
	Housing	0.000	0.070	0.105	0.227
	Fraction	0.003	0.070	0.107	0.205
	Type	-1.485e+09	1.485e+09	1.208e+10	-1.106e+22

Table 11: Performance of intermediary linear models that add all features of a category to the baseline features. Categories are ordered by descending R^2_{adj} score, such that the best performing categories per level are at the top. Categories in bold are considered meaningful features.

- The number of tickets is unlikely to influence the minimum, average, and maximum asset ages. High number of tickets may lead to high number of asset replacements, lowering the ages of assets. However, this would require a high number of severe tickets (requiring replacement of the asset) and no such large amounts of severe tickets has been observed in the data.
- The high number of tickets may influence the number of compression containers. It is a known fact that compressing containers have a much higher need for maintenance than non-compressing containers. As such, high number of tickets have lead to a stop in acquiring such containers. The impact of this decision is however not expected to be noticeably present in the data given the time of this decision and the long life-span of existing containers.

2. The residuals of the predictions are plotted against the district the predicted ward belongs to, and against the years they were made in, in Figure 12. Residual plots for the neighbourhood predictions are similar to those of the ward predictions and are therefore not shown. The residuals per year show a clear upward trend for both linear regression and random forest regression, implying that the number of tickets has increased in a way that the models/data cannot explain. The residuals per district show a reasonably independent spread with the clear exception of the linear regressions in district "N" (Noord). This exception may be explained by the fact that Noord is the only district with high numbers of compression containers and is therefore significantly different from other districts. This is a non-linear relation the linear model can clearly not explain. The random forest model is, however, able to account for this phenomenon, as can be observed from Figure 12 (d), where the residuals for district Noord are also evenly spread around 0.

Model	Type	ME	MAE	RMSE	R^2_{adj}
Ward baseline	Linear	0.001	0.082	0.130	0.284
	Random forest	0.002	0.085	0.143	0.139
Ward final	Linear	0.003	0.072	0.115	0.426
	Random forest	0.016	0.072	0.119	0.385
Neighbourhood baseline	Linear	0.001	0.070	0.105	0.240
	Random forest	0.000	0.070	0.110	0.162
Neighbourhood final	Linear	0.001	0.061	0.097	0.345
	Random forest	0.004	0.062	0.102	0.273

Table 12: Performance metrics of all produced models, with exception of intermediary models used for feature selection. Note that all features and the target variable are scaled between 0-1, meaning that the absolute performance metrics do not give an indication in number of tickets.

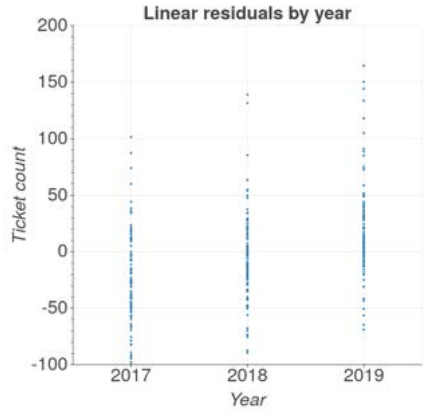
Model	Type	ME	MAE	RMSE	R^2_{adj}
Ward baseline	Linear	0.443	45.286	71.657	0.284
	Random forest	0.378	47.130	78.433	0.142
Ward final	Linear	1.683	39.547	63.766	0.429
	Random forest	6.835	40.629	67.785	0.343
Neighbourhood baseline	Linear	0.112	14.307	21.447	0.240
	Random forest	0.139	14.298	22.488	0.165
Neighbourhood final	Linear	0.117	12.556	19.852	0.345
	Random forest	1.021	12.875	21.223	0.251

Table 13: Performance metrics of all produced models, with exception of intermediary models used for feature selection. Error values are absolute, allowing them to be interpreted as number of tickets.

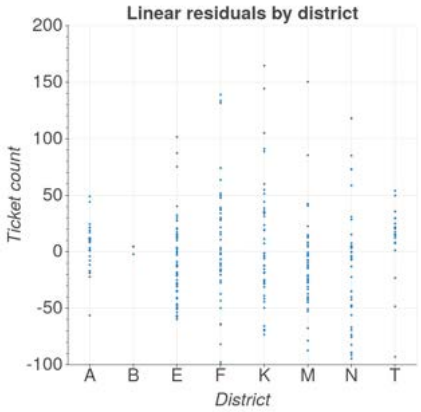
3. Plotting the residuals per feature shows one clear issue: the residual plots for hoisting types show that the range of these features is very limited, consisting of 3-10 unique values with the majority of points being 0. An example of this observation is shown in Figure 13 (a) showing the residuals of hoisting type "3 haken" for neighbourhood level prediction. While this phenomenom is less pronounced for ward level predictions shown in Figure 13 (b), there is still a noticibly high number of "0" values. Given the limited range of values, it seems the features are not informative to the model.

4. VIF scores are presented in Table 14, they show that the VIF scores of all features are below the threshold of 4, this is a natural result of the chosen methodology in which all features over the threshold of 4 are discarded.

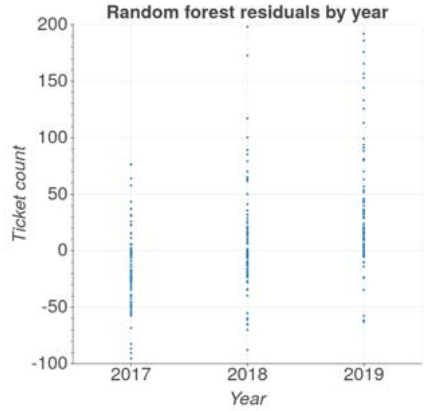
Errors are not normally distributed, showing less than expected errors in the lower positive residual ranges (0-150) and more than expected errors in the higher positive range (150-500), see Figure 14. A possible explanation of this observation is that the variance in the number of tickets increases with the number of tickets, making the higher ranges less predictable than the lower ranges.



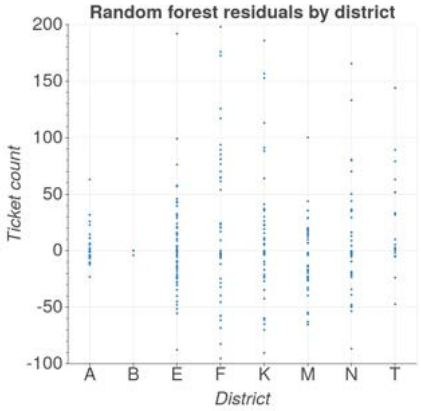
(a) Residuals by year - Linear regression



(b) Residuals by district - Linear regression



(c) Residuals by year - Random forest regression



(d) Residuals by district - Random forest regression

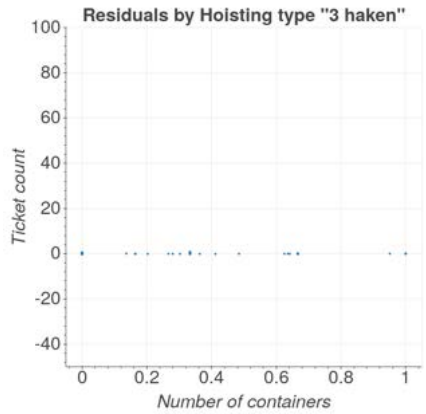
Figure 12: Residual plots for the linear model predicting number of tickets in wards for (a) years and (b) districts and for the random forest model for (c) years and (d) districts.

7.2.4 Feature importance

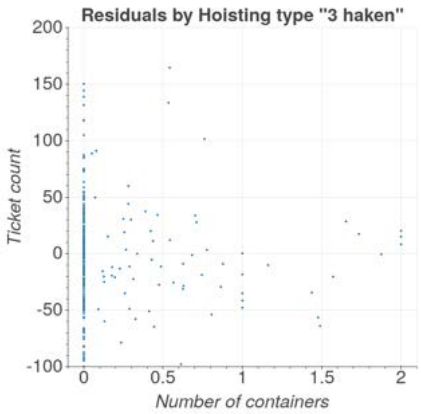
The feature importance of the final models used in predicting the number of tickets for wards and neighbourhoods are shown in Table 14. From the tables, it is clear that both the linear model as well as the random forest model identify the number of existing containers, the number of compression containers, and the lowest and average asset ages as important factors. The hoisting types ”1 haak” is identified by the random forest model as reasonably important, but in the linear model, the feature is not less important.

7.2.5 Manual improvements

Based on the determined feature importance, features can be pruned that do not add significant value to the model. All hoisting type features and the lowest asset age have relatively low coefficients and/or Gini indexes. If these features are removed from the model, the predictive capabilities of the linear models stays roughly the same, while the model is conceptually simpler. The performance of the final models with reduced features is shown in Table 15 (scaled) and Table 16 (unscaled), compared to the original models with extended features. Feature importance is reported in Table 17.

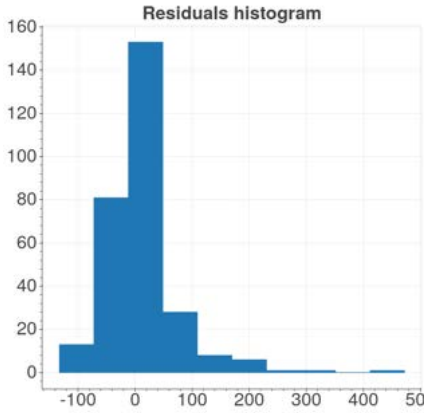


(a) Neighbourhood level prediction

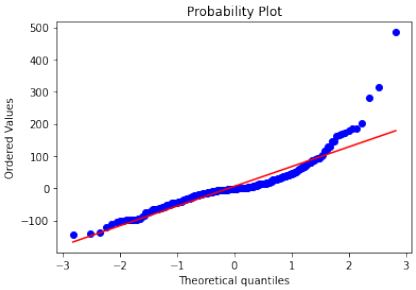


(b) Ward level prediction

Figure 13: Residual plots for hoisting type ”3 haken” for neighbourhood and ward level predictions, showing the low number of unique values and high number of ”0” values.



(a) Error histogram - Random forest



(b) QQ-plot of residuals - Random forest

Figure 14: Error histogram and QQ-plot of the prediction of number of tickets for wards. The linear regression model plots look similar.

The final manually improved models show that for ward level predictions more variance in the data is explained when compared to neighbourhood level predictions. This is likely due to a reduction in noise when comparing ward level data to neighbourhood level data. However, the (scaled) MAE and RMSE scores of the ward models are higher than those of the neighbourhood models, implying that the ward model perform worse. However, the difference is small, and the reason for this difference may also lay in the scaling method: large outliers in one data level and small outliers in another may skew the scales unrealistically. Based on the MAE scores, one can reasonably expect an error of 40 tickets on ward level predictions and 12 on neighbourhood level predictions. RMSE values of approximately 65 and 20 show that the errors are prone to having outliers. Given an average number of tickets of 206 and 50 respectively, these error rates are reasonable, but seem too large for practical use.

Prediction level	Feature	Coefficient β	Gini index	VIF
Ward	Hoisting type "1 haak"	0.062	0.124	1.709
	Hoisting type "3 haken"	0.044	0.023	1.168
	Hoisting type "anders"	-0.101	0.003	1.079
	Hoisting type "geen"	-0.092	0.045	2.029
	Compression containers	0.188	0.170	1.172
	Existing containers	0.387	0.350	2.550
	Lowest asset age	-0.111	0.104	1.883
	Average asset age	0.262	0.182	2.918
Neighbourhood	Hoisting type "1 haak"	0.268	0.259	1.458
	Hoisting type "3 haken"	0.038	0.016	1.116
	Hoisting type "anders"	-0.020	0.001	1.032
	Hoisting type "geen"	-0.042	0.027	1.804
	Lowest asset age	0.004	0.152	1.564
	Highest container age	0.067	0.138	2.236
	Existing containers	0.265	0.269	2.669
	Compression containers	0.204	0.139	1.233

Table 14: Features in the final prediction models. Feature importance for the linear model (Coefficient β) and random forest (Gini index) are shown.

	Type	ME	MAE	RMSE	R^2_{adj}
Ward final	Linear	0.003	0.072	0.115	0.426
	Random forest	0.016	0.072	0.119	0.385
Ward final - reduced features	Linear	0.001	0.072	0.116	0.429
	Random forest	0.004	0.073	0.122	0.370
Neighbourhood final	Linear	0.001	0.061	0.097	0.345
	Random forest	0.004	0.062	0.102	0.273
Neighbourhood final - reduced features	Linear	0.001	0.060	0.094	0.381
	Random forest	0.001	0.063	0.102	0.276

Table 15: Scaled performance metrics of the final model with reduced features, compared to the original.

	Type	ME	MAE	RMSE	R^2_{adj}
Ward final	Linear	1.683	39.547	63.347	0.426
	Random forest	6.835	40.629	67.785	0.343
Ward final - reduced features	Linear	0.826	39.536	63.766	0.429
	Random forest	2.078	40.743	66.917	0.371
Neighbourhood final	Linear	0.117	12.556	19.852	0.345
	Random forest	1.021	12.875	21.223	0.251
Neighbourhood final - reduced features	Linear	0.173	12.244	19.337	0.381
	Random forest	0.279	12.785	20.838	0.281

Table 16: Unscaled performance metrics of the final model with reduced features, compared to the original.

Prediction level	Feature	Coefficient β	Gini index	VIF
Ward	Average asset age	0.259	0.335	1.709
	Compression containers	0.202	0.184	1.162
	Existing containers	0.411	0.480	1.602
Neighbourhood	Average asset age	0.386	0.392	1.496
	Compression containers	0.234	0.177	1.136
	Existing containers	0.340	0.437	1.459

Table 17: Feature importance of the final models with reduced features.

8 Discussion

In this section, the literature study, case study experiment, and their results will be discussed. An analysis is made of the weaknesses and strengths to this study, what new questions and areas of research this study uncovered, and what threats exist that could invalidate the results of this study.

The literature study has identified that maintenance prognostics generally rely on the age of an asset, and that multiple instances of a model are created to take into account different circumstances such as different cities, different asset types, and different time spans. These models are often of a statistical nature, in which the likelihood of an asset requiring maintenance is a function of time or usage, but exceptions exist. Using this likelihood of is also known as condition based maintenance. The chosen approach for the case study experiment, training and applying a linear regression model for a specific area, is in line with these common approaches. The used model, linear regression, was chosen to support more input variables than common statistical methods and to allow for easy determination of feature importance. However, since the results identify the age of an asset as the primary factor, a statistical model may have sufficed.

While the literature study does identify the taken approach as a commonly successful one, it lacks quantitative support for the decision to apply these methods to the case study. This study would benefit from a more structured literature study in which the use of this approach is identified and the performance of different types of models, scenarios, and used variables are reported for a higher number of studies than is currently the case, in a more structured manner.

The literature study furthermore identifies relevant information from the domain of MSW prediction. Even though the prediction of waste generation is not directly transferable to the prediction of maintenance, it was expected to assist maintenance predictions by providing (latent) information on the usage of the container, which can subsequently impact maintenance requirements. However, the experiment has shown that this impact is either not present or negligible. Future research into the topic of this thesis should therefore treat this problem more as a maintenance-oriented problem, and less as a socially-oriented one.

The case study describes three parts: the background of underground waste container maintenance in Amsterdam, the available data, and the experiment. While not a primary goal of the thesis, the gathered background information and data description is expected to be of great value to the stakeholders in this research: the municipality of Amsterdam and Curious Inc. The consolidation of information and complete presentation of current processes and data can serve as a reference for both parties to understand each others processes for future collaboration.

The experiment is grounded in existing work and leads to the clear conclusion that the only relevant factors for underground container maintenance in the municipality of Amsterdam are related to asset age. An issue left unexplained by the experiment is the apparent yearly increase in maintenance. An increase which is not (linearly) correlated with the increase in number of assets as one might expect. This open point is further discussed as future work in Subsection 8.1.

Some problems in the maintenance data have become apparent that are expected to have an impact on the results. The most concrete issue with the data is that historic locations of a container are not recorded. If a container is moved from one location to another, all references of the container existing at the previous location are lost. Moreover, since the maintenance history of the container is moved along with it, the new area inherits all past maintenance of the container. The exact impact of this deficiency is not clear, however domain experts indicate that moving a container between neighbourhoods seldom happens. Additionally, various aspects of the data reflect another issue: the current data set and maintenance administration application is very much optimized for operational support, but not so much for strategic analysis. Indicators of this include the needlessly

high number of container types (which may work well in practice, but make aggregation hard), the previously mentioned lack of history of a containers placement (which makes reliable historic analysis questionable), and lack of administration of certain fields (most notably placing/delivery/operational dates). Detailed causes/damaged parts for historic maintenance actions are also hard to deduce, since only the *problem module* is consistently identified and the free form text description is hard to parse. It should be noted that since 2019, the municipality has started to consistently administrate the cause of the problem (i.e. misuse by garbage collector, misuse by citizen, fire, vandalism, etc.) which should assist in future research. Based on the issues encountered in this thesis, the municipality is advised to also create an exhaustive list of common problems and resolutions (i.e. "inlet stuck due to damage", "repair inlet" or "replace inlet" respectively) such that these may be easily identified in future research and analysis on specific problems can be performed, without the need to rely on automated text parsing.

Some demographics data has been interpolated based on previous years, while assumed to be accurate enough for the purposes of this study, it must be noted as a possible threat as it may have caused compounding errors. For a deeper analysis of this approach and its advantages and disadvantages, the reader is referred back to Section 6.3.2.

8.1 Future work

This subsection will discuss various new avenues of research that have surfaced during the writing of this thesis. Such avenues may serve to deepen this research, revisit assumptions made in this thesis, or address new issues that have presented themselves.

The following questions have surfaced during this study that may be answered in subsequent research.

A trend of rising numbers of tickets over time that is not proportional to the increase in number of assets has been shown, as is well illustrated in Table 9 and Figure 9. This trend can be explained by the increasing age of the assets as the years increase. However, the research has also shown that while age is indeed an important factor to estimate the yearly maintenance of a container, it has also shown that it is not a particularly strong predictor. As such, it is expected that other causes exist that cause the number of tickets to increase over the years such as an increased willingness to use *Grybb* to report and administrate issues.

The various districts have been shown to behave differently when it concerns the number of tickets as also clear from Figure 11. Such a difference could be easily explained given the fact that the districts where in control of waste processing themselves until end 2016, with varying policies and processes as a result. While the processes have been consolidated city-wide at the start of 2017, old behaviour of users in the various districts is expected to still be present, and have an impact on the number of tickets generated by a district. As such a relation can hardly be represented by a linear regression model while still being generalizable, the discrepancies between districts may be addressed by creating a model per district. Such models should then be created with the feature set that is concluded by city-wide experiments to keep the models generalizable.

Apart from optimizing models to fit various districts, the maintenance policies of these districts could be analyzed for similarities and differences based on the results found in this study. The results of the study may be used to help calculate the ideal age to replace a container; at which point the cost of the generated tickets break even with replacement costs and subsequent reduced maintenance.

To deepen this specific research topic, the following opportunities have shown themselves.

The garbage collection data that has been ignored by this study may be re-examined at a later time when flaws are hopefully resolved. Furthermore, given that the flaws of the data are known and consistent, it may be possible to formulate some coarse features with

it. Specifically, while a roughly estimated 20% of all garbage collections is missing, as long as this missing data is spread evenly over districts it may be useful for the purposes of this particular subject. However, determining the extent of the data reliability is a matter of research on its own.

A comparative study between municipalities may be performed to (1) validate the results of this study and (2) generalize the results, given that the municipality has a similar setup and workflow regarding underground waste containers. Furthermore, if such a municipality shows a significantly different weather pattern than the municipality of Amsterdam, such a study may also study the effects of weather on maintenance load.

Opportunities specific to the maintenance data concern processing the free text description of a ticket and refining the exact issues that are the cause of a ticket. It seems feasible that the free text can be mined to include extra features with fine-grained descriptions of the problems. Furthermore, an attempt may be made to specialize prediction models on very specific problems, such as predicting only the number of broken hoisting rods. Ticket description mining may assist in narrowing tickets down to a suitable level.

While useful to know, predicting the number of tickets in a year is not the most important target variable when considering the use case of a yearly budget planning. The most important variable to predict could be the total cost, or materials and man-hours required. These numbers can, in part, be retrieved from the available data but it is unknown how consistent and reliable these records are. If performed, a regression can be made directly on the total expected cost. Alternatively, the expected cost/time given a type of ticket can first be calculated and then combined with a prediction of the how often such a ticket occurs to arrive at a total estimate.

Furthermore, this study performed a forecast of maintenance and not predictive maintenance due to the lack of real-time data. Given the results of the experiment which seem to point towards age as being the primary explaining factor of maintenance, it seems irrelevant at this time to invest in gaining real-time data of containers and their usage. Further research into more specific cases of damage may find that they do require sensor data. For example, if the causes of damage to the inlet is further analyzed, information on how often the inlet mechanism is opened and how much garbage is deposited might be benificial.

Some assumptions that are made in this research may be separate subjects of study.

Most notable in this category is the prediction of demographic variables: is such a prediction feasible and accurate, and how should it be executed? Additionally, the validation of the predictive model assumes that the performance of an administrative area is independent from the performance of previous year. However, cases are imaginable in which this is not true: for example when a neighbourhood has an excessive number of tickets in one year, special care may be taken during the next year to keep the number of tickets low. The development of a single area over multiple years may be studied to confirm or disprove such assumptions.

Finally, this study has only considered unplanned tickets to satisfy one of the goals of the municipality of Amsterdam: lowering the overall number of tickets. The reasoning was that lowering the number of unplanned tickets is more feasible than lowering the number of planned tickets because planned tickets are generally (legally) required to be performed. However, periodic cleaning is part of these planned tickets and performed quarterly per container. A feasible approach to lowering the number of planned tickets may be deciding whether or not a container should actually be cleaned quarterly, or if a lower frequency is also acceptable.

8.2 Research sidetracks

Aside from the research presented in this thesis, two other research directions were explored but not thoroughly followed due to lack of results or promise. This section makes note of those sidetracks in a short, less formal way such that this knowledge is not lost. Two

subjects were explored: predictive maintenance on underground waste containers using collection data, and cost prediction of yearly maintenance.

Application of predictive maintenance was attempted based on waste collection data. Domain experts have indicated that certain types of damages are a direct result of rough handling during collection (or maintenance) of a container. For example, when the container is not lifted out of its well perpendicular to street level, horizontal forces are exerted on the hoisting mechanism, which is primarily designed to handle vertical forces, causing the hoisting rods to bend occasionally.

During garbage collection, the weight of garbage in the container is measured using sensors. The container is lifted out of it's well, kept hanging stationary in the air to perform a measurement, emptied, weighted again, and lowered back into the well. The expectation was that anomalies might present themselves that would correlate to specific cases of damage such as bent hoisting rods. However, such a pattern did not become visible and the research was not pursued further.

Secondly, an optimistic goal of this thesis was to predict the expected cost of maintenance for a year. The chosen approach was to first predict the number of maintenance issues and then predict the costs based on the expected number of tickets. The primary issue encountered in this approach was that the prediction of number of tickets is not reliable enough to base another prediction on. Furthermore, estimating the cost of specific issues (i.e. a bent hoisting rod, broken emptying doors) proved a challenge and may be a subject of research on its own. As such this part of the research was removed from scope. The main challenge in estimating costs of specific issues is the absence of detailed and structured data with regards to the exact damaged parts, causes, and actions taken. Such data is often generally described by structured data, but the specifics are contained in free text fields. As such, future research in this direction might start with text mining of free text fields in tickets.

9 Conclusion

In this thesis, a model is created to predict the required yearly maintenance on underground waste containers for the municipality of Amsterdam, the Netherlands. The goals of the study are to identify relevant factors to the yearly required maintenance such that they may be addressed if possible, and the maintenance lowered. The construction of the model is inspired by existing literature in the areas of Prognostics Health Management (PHM) and Municipal Solid Waste (MSW) generation. Subsequently, the model adapted based on local domain expert input and available information and applied to the case study. To conclude the thesis, the research questions are answered.

> *Sub question 1: What factors can be relevant in predicting required maintenance on underground waste containers?*

Literature on PHM suggests use of asset age, weather conditions and location of the asset. Literature on MSW shows demographics play a role in MSW generation and names factors such as household size, residency type, age groups, employment, gross domestic product, education, culture, geography, and climate.

> *Sub question 2: What types of models are appropriate to model the required maintenance in such a way that relevant factors can be identified and explained to domain experts?*

In the field of PHM, mostly statistical models are used such as Weibull distributions, Kaplan-Meier estimators, and Cox Proportional Hazard Model. Such models generally only take one input parameters and produce a likelihood of the asset failing in a given time span. To compensate for different operating circumstances of assets (weather, environment, etc.), separate models are constructed for different operating contexts. Both a Weibull distribution and a Kaplan-Meier estimator have limited complexity and can be explained to stakeholders. Predictions on MSW generation are commonly performed using a wide array of data driven models such as linear regression, support vector machines, and neural networks. Such studies often note that a linear model is preferred given that it is easy to explain to (non-technical) stakeholders.

> *Sub question 3: To what extent can the chosen model predict the required unplanned maintenance for the municipality of Amsterdam?*

A linear regression model is created to predict the number of unplanned maintenance occurrences on underground waste containers in the municipality of Amsterdam. An iterative approach is used to discard (multi-)colinear or irrelevant features. The final model has a mean absolute error of 40 tickets, on an average value of 69 tickets per district per year. As such, roughly speaking, one can expect a 60% error on the prediction. While this final model performs better than an intuitive baseline prediction, its predictions are hardly actionable in a practical context. The iterative feature elimination has discarded all features except the number of containers, age-related features, and asset type features. The latter, however, have been shown to not be meaningful predictors, and have been discarded manually.

> *Research question: How can unplanned maintenance on underground waste containers be estimated?*

The primary answer this thesis gives to this question is that a solution to the problem should be sought within the domain of PHM and maintenance. Factors that have shown to be relevant to waste generation have little to no impact on the number of unplanned maintenance occurrences. Furthermore, maintenance prediction using a linear model has

been shown to be feasible, but the prediction is overall not accurate enough to be useful in practice. Prediction using a linear model does show promise, and, supported by findings in this thesis, future work may be able to arrive at a model suitable for use in practice.

The main contribution of this research is the exploration of maintenance forecasts on underground waste containers. The approach of maintenance forecasting in domains similar to underground waste containers has been tested against a use case and found to be applicable. During implementation of the use case, a detailed breakdown of available data has been made exposing its strengths and flaws which may serve to improve the gathering of data for the sake of future research. Furthermore, future work has been suggested to further delve into maintenance forecasts for underground waste containers in the municipality of Amsterdam, which is expected to be transferable to different Dutch municipalities, lowering the future cost and effort required to keep garbage disposal facilities available for citizens.

A Full example record

Below is a full example of a record on neighbourhood level (as opposed to ward level) used to train and validate the predictive models. This record is from the full dataset and has not been checked for multi-collinearity.

Feature name	Value
n_code	A00d
year	2019
year_num	2019
d_code	A00
cp_code	A
existing_containers	2
ticket_count	1
meta_age_min_container_age	-256
meta_age_average_container_age	-25
meta_age_max_container_age	93
meta_age_stddev_container_age	123.107
type_afvalboei_bovengrondse_container_blauwgeel...	0
type_algemene_order	0
type_asw_bos_lommer_belfast_khc_rest	0
type_asw_khdc_glas_kh	0
type_asw_khdc_papier_kh	0
type_asw_khdc_rest	0
type_asw_khdc_rest_kh	0
type_asw_rubens_rest_kh	0
type_asw_vconsyst_metro_rest_kh	0
type_bauer_rest_bovengrondse	0
type_bg_brood_rolcontainer	0
type_bos_en_lommer_ondergrondse_rest	0
type_bovengronds_cushion_bwaste	0
type_bovengrondse_collector_bammens_glas	0
type_bovengrondse_collector_bammens_papier	0
type_bovengrondse_collector_bammens_rest	0
type_bovengrondse_glas	0
type_bovengrondse_rest	0
type_bovengronds_papier	0
type_bovengronds_rest	0
type_bovengronds_textiel_symphanie	0
type_centrum_amsterdam_evo_glas	0
type_centrum_amsterdam_evo_papier_kh	0
type_centrum_evo_l_glas	0
type_centrum_evo_l_papier	0
type_centrum_evo_l_rest	0
type_centrum_glas_kh_evol	0
type_centrum_papier_kh_evoll	0
type_centrum_rest_evo_kh	0
type_centrum_stadsdeel_evo_papier_kh	0
type_centrum_stadsdeel_evo_rest_kh	0
type_centrum_stadsdeel_glas	0
type_centrum_stadsdeel_glas_kh	0
type_centrum_stadsdeel_glas_kh_eigen_zuil	0
type_centrum_stadsdeel_papier_kh_wolff	0

type_engels_upperground	0
type_gft_kh_oc_vconsyst	0
type_glasbont_kh_semibg_bauer	0
type_glas_wit_hkz_oc_vconsyst	0
type_haaks_glas_bammens_downcost	0
type_haaks_glas_metro	0
type_haaks_papier_metro	0
type_haaks_plastic_metro	0
type_haaks_rest_metro	0
type_hms_rest	0
type_khc_glas_amsterdam_standaard	0
type_khc_glas_amsterdam_vconsyst	0
type_khc_glas_bg	0.888889
type_khc_kartonklep_amsterdam_standaard	0
type_khc_papier_amsterdam_standaard	0
type_khc_papier_amsterdam_standaard_kk	0
type_khc_papier_amsterdam_vconsyst	0
type_khc_papier_bg	1
type_khc_plastic_amsterdam_standaard	0
type_khc_plastic_amsterdam_vconsyst	0
type_khc_plastic_bg	0
type_khc_rest_amsterdam_standaard	0
type_khc_rest_amsterdam_vconsyst	0
type_khc_rest_belfast	0
type_khc_rest_bg	0
type_khc_rest_semi	0
type_khc_textiel_amsterdam_standaard	0
type_kh_glas_bammens_downcost	0
type_kh_glas_bammens_inwerpzuil	0
type_kh_glas_icova	0
type_kh_glas_rub_tr_verstel	0
type_kh_metro	0
type_kh_papier_bammens_downcost	0
type_kh_papier_icova	0
type_kh_papier_metro	0
type_kh_rest_bammens_downcost	0
type_kh_rest_bammens_inwerpzuil	0
type_kh_rest_metro	0
type_kikker_rest	0
type_kunststof_plastic_kh_bg_mcb	0
type_mcb_bovengrondse_glas	0
type_mcb_bovengrondse_papier	0
type_mcb_bovengrondse_plastic	0
type_mcb_bovengrondse_rest	0
type_nieuwe_west_belfast_glas	0
type_nieuwe_west_belfast_rest	0
type_nieuw_west_belfast_papier	0
type_nieuwwest_glas_kh_tr_opl_inst	0
type_nieuw_west_kuub_ru_ki_pap	0
type_nieuw_west_papier	0
type_nieuwwest_papier	0
type_nieuw_west_rest	0
type_nieuw_west_rest_belfast	0

type_nieuwwest_rest_rubens	0
type_nieuwwest_sv_rest_tr_opl_inst	0
type_oost_belfast_ed_kh_db_rest	0
type_oost_stadsdeel_bwaste_glas	0
type_oost_stadsdeel_bwaste_papier	0
type_oost_stadsdeel_bwaste_rest	0
type_oost_stadsdeel_dc_glas_h	0
type_oost_stadsdeel_dc_papier_h	0
type_oost_stadsdeel_dc_rest_h	0
type_oost_stadsdeel_ijburg_papier_kh	0
type_oost_stadsdeelijburg_rest_rubens_kh	0
type_oost_stadsdeel_ijburg_rubens_kh_glas	0
type_oost_stadsdeel_ijburg_rubens_rest_kh	0
type_oost_stadsdeel_khdc_glas	0
type_oost_stadsdeel_khdc_papier	0
type_oost_stadsdeel_khdc_rest	0
type_oost_stadsdeel_rubens_glas_kh	0
type_oost_stadsdeel_rubens_papier_kh	0
type_oost_stadsdeel_vconsyst_glas	0
type_oost_stadsdeel_vconsyst_glas_ral	0
type_oost_stadsdeel_vconsyst_kunststof	0
type_oost_stadsdeel_vconsyst_papier	0
type_oost_stadsdeel_vconsyst_papier_ral	0
type_oost_stadsdeel_vconsyst_rest	0
type_oost_stadsdeel_vconsyst_rest_grijsblauw_ral	0
type_oost_stadsdeel_vconsyst_rest_ral	0
type_oost_stadsdeel_vconsyst_textiel	0
type_papier_hkz_oc_vconsyst	0
type_papier_kh_semibg_bauer	0
type_pers_rest_icova_metro_klem	0
type_plastic_khc_pers_amsterdam_standaard_sidcon	0
type_rest_hk_oc_vconsyst	0
type_rest_hkz_oc_vconsyst	0
type_rest_khc_pers_amsterdam_standaard_sidcon	0
type_rest_kh_oc_vconsyst	0
type_rhino_pers_rest_belfast_kh	0
type_rolcontainer_kunststof_rest	0
type_rolcontainer_staal_rest	0
type_semi_ondergrondse_oc_asw	0
type_sia_melding	0
type_sidcon_pers_rest	0
type_sulo_classic_ii	0
type_svg_hkpf_evol_tropl	0
type_svg_hkz_bel_pap_tropl	0
type_textiel_hkz_oc_vconsyst	0
type_tvg_hkz_bel_pap_tropl	0
type_tvg_hkz_bel_tropl	0
type_tvg_kh_bel_tropl	0
type_tvg_khc_utr_tropl	0
type_tv_khfc_snaas_ams_rest_trinl	0
type_tvr_g_kh_bel_gl_tropl_isol	0
type_tvr_g_kh_bel_pap_tropl	0
type_tvr_g_kh_bel_tropl	0

type_tvri_hz_metro_rest_tr	0
type_tvr_kh_bel_gl_tropl	0
type_tvr_kh_bel_pap_tropl	0
type_tvr_kh_bel_pap_tropl_g	0
type_tvr_kh_bel_tropl	0
type_tvr_kh_bel_tropl_g	0
type_tvrkliko_bel_tropl	0
type_utrecht_rhino_perscontainer_rest	0
type_west_stadsdeel_vconsyst_glas	0
type_west_stadsdeel_vconsyst_papier	0
type_west_stadsdeel_vconsyst_rest	0
type_zuidoost_stadsdeel_kliko_rest	0
type_zuid_oost_stadsdeel_metro_glas	0
type_zuidoost_sulo_rest	0
bbga_bevtotaal	352
bbga_beveenouderhh	8
bbga_bevalleenhh_p	58
bbga_bevpaarzkindhh_p	29.2
bbga_bevpaarmkindhh_p	5.3
bbga_bevoverighh_p	4
bbga_wbezet	1.35
bbga_ihhink_gem	45455.6
bbga_iinkq1_p	23.9643
bbga_iinkq2_p	20.6786
bbga_iinkq3_p	22.4643
bbga_iinkq4_p	13.5714
bbga_iinkq5_p	19.25
bbga_wcorhuur_p	5.4
bbga_wparthuur_p	67.3
bbga_wkoop_p	27.3
bbga_bevopllaag_p	5.5
bbga_bevoplmid_p	27.5
bbga_bevoplhoog_p	67.5
bbga_skses234_p	17
bbga_skses_gem	7.33333
bbga_bev0_4	12
bbga_bev5_9	4
bbga_bev10_14	6
bbga_bev15_19	4
bbga_bev20_24	53
bbga_bev25_29	74
bbga_bev30_34	51
bbga_bev35_39	39
bbga_bev40_44	19
bbga_bev45_49	16
bbga_bev50_54	19
bbga_bev55_59	16
bbga_bev60_64	13
bbga_bev65_69	13
bbga_bev70_74	5
bbga_bev75_79	5
bbga_bev80_84	2
bbga_bev85_89	1

hoisting_type_1_haak	0
hoisting_type_3_haken	0
hoisting_type_anders	0
hoisting_type_geen	0
hoisting_type_kinshofer	1.88889
insertion_type_amsterdam	0
insertion_type_amsterdam_standaard	0
insertion_type_anders	0
insertion_type_belfast	0
insertion_type_bovengrondse	1.88889
insertion_type_broodbak	0
insertion_type_bwaste	0
insertion_type_down_cost	0
insertion_type_evolution	0
insertion_type_geen_inwerpzuil	0
insertion_type_kikker	0
insertion_type_metro	0
insertion_type_rubens	0
insertion_type_type_2002	0
insertion_type_v_consyst	0
pers_containers	0
fractie_brood	0
fractie_gft	0
fractie_glas	0.888889
fractie_papier	1
fractie_plastic	0
fractie_rest	0
fractie_textiel	0

References

[1] Gemeente Amsterdam, Onderzoek, Informatie en Statistiek . Basisbestand Gebieden Amsterdam . https://data.amsterdam.nl/datasets/G5JpqNbhweXZSw/basisbestand-gebieden-amsterdam-bbga/, november 2019. Accessed: 2019-02-18.

[2] Beigl, P., Lebersorger, S., and Salhofer, S. Modelling municipal solid waste generation: A review. *Waste management 28*, 1 (2008), 200–214.

[3] Bessani, M., Fanucchi, R. Z., Achcar, J. A., and Maciel, C. D. A statistical analysis and modeling of repair data from a brazilian power distribution system. In *2016 17th International Conference on Harmonics and Quality of Power (ICHQP)* (2016), IEEE, pp. 473–477.

[4] CBS (Statistics Netherlands). Cbs statline. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37230ned/table?fromstatweb, 2019. Accessed: 2020-09-10.

[5] Chung, S. S. Projecting municipal solid waste: The case of hong kong sar. *Resources, Conservation and Recycling 54*, 11 (2010), 759–768.

[6] De Lucia, A., Pompella, E., and Stefanucci, S. Assessing effort estimation models for corrective maintenance through empirical studies. *Information and Software Technology 47*, 1 (2005), 3–15.

[7] Edwards, D. J., Holt, G. D., and Harris, F. C. A comparative analysis between the multilayer perceptron "neural network" and multiple regression analysis for predicting construction plant maintenance costs. *Journal of Quality in Maintenance Engineering* (2000).

[8] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM 39*, 11 (1996), 27–34.

[9] Finnegan, M. An evaluation on multi-site municipal solid waste generation forecasting using deep learning andarima approaches, 2020.

[10] Gross, P., Salleb-Aouissi, A., Dutta, H., and Boulanger, A. Ranking electrical feeders of the new york power grid. In *2009 International Conference on Machine Learning and Applications* (2009), IEEE, pp. 359–365.

[11] Hayashi, F. *Econometrics*. Princeton University Press, 2000.

[12] Karballaeezadeh, N., Mohammadzadeh S, D., Shamshirband, S., Hajikhodaverdikhan, P., Mosavi, A., and Chau, K.-w. Prediction of remaining service life of pavement using an optimized support vector machine (case study of semnan–firuzkuh road). *Engineering Applications of Computational Fluid Mechanics 13*, 1 (2019), 188–198.

[13] Keser, S., Duzgun, S., and Aksoy, A. Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in turkey. *Waste management 32*, 3 (2012), 359–371.

[14] Kimutai, E., Betrie, G., Brander, R., Sadiq, R., and Tesfamariam, S. Comparison of statistical models for predicting pipe failures: Illustrative example with the city of calgary water main failure. *Journal of Pipeline Systems Engineering and Practice 6*, 4 (2015), 04015005.

[15] Kolekar, K., Hazra, T., and Chakrabarty, S. A review on prediction of municipal solid waste generation models. *Procedia Environmental Sciences 35* (2016), 238–244.

[16] Lebersorger, S., and Beigl, P. Municipal solid waste generation in municipalities: Quantifying impacts of household structure, commercial waste and domestic fuel. *Waste management 31*, 9-10 (2011), 1907–1915.

[17] Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., and Hampapur, A. Improving rail network velocity: A machine learning approach to predictive maintenance. *Transportation Research Part C: Emerging Technologies 45* (2014), 17–26.

[18] Orhan, S., Aktürk, N., and Celik, V. Vibration monitoring for defect diagnosis of rolling element bearings as a predictive maintenance tool: Comprehensive case studies. *Ndt & E International 39*, 4 (2006), 293–298.

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[20] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. A design science research methodology for information systems research. *Journal of management information systems 24*, 3 (2007), 45–77.

[21] Pfeffer, F. T. Growing wealth gaps in education. *Demography 55*, 3 (2018), 1033–1068.

[22] Popova, E., Yu, W., Kee, E., Sun, A., Richards, D., and Grantom, R. Basic factors to forecast maintenance cost and failure processes for nuclear power plants. *Nuclear Engineering and Design 236*, 14-16 (2006), 1641–1647.

[23] Rademakers, L., Braam, H., Obdam, T., Frohböse, P., and Kruse, N. Tools for estimating operation and maintenance costs of offshore wind farms: state of the art. In *Proc. of EWEC* (2008), Citeseer.

[24] Rudin, C., Waltz, D., Anderson, R. N., Boulanger, A., Salleb-Aouissi, A., Chow, M., Dutta, H., Gross, P. N., Huang, B., Ierome, S., et al. Machine learning for the new york city power grid. *IEEE transactions on pattern analysis and machine intelligence 34*, 2 (2011), 328–345.