FHV
Vorarlberg University
of Applied Sciences

Classification of GPS Track Data Using AI Methods

A Case Study of Waste Collection Vehicles

Bachelor thesis
for obtaining the academic degree

Bachelor of Science in Engineering (BSc)

Vorarlberg University of Applied Sciences
Computer Science - Software and Information Engineering

Supervised by
Dipl.-Ing. Dr. techn. Ralph Hoch

Submitted by
Matthias Hefel
Dornbirn, May 2025

Dedication

Dedicated to my younger self, who found a fascination in software from a young age and decided to follow his dreams!

And to my parents, who wholeheartedly supported me throughout this journey.

Thank you.

Kurzreferat

Klassifizierung von GPS-Spurdaten mit Unterstützung von KI-Methoden am Beispiel von Abfallsammelfahrzeugen

In der Abfallwirtschaft ist die strategische Tourenplanung ein wichtiger Prozess, in dem durch optimale Gebietsaufteilung eine maximal effiziente Fuhrparkauslastung bei möglichst geringen Kosten ermittelt wird. Dies geschieht in Entsorgungsbetrieben sowohl für bestehende Auftragsgebiete, als auch bei der Kalkulation von neuen Ausschreibungen. Vor Allem bei Regionen, in denen keine Erfahrungswerte vorliegen müssen für eine robuste Tourenplanung zahlreiche unscharfe Annahmen getroffen und manchmal auch Schätzungen vorgenommen werden. Um diese Unsicherheiten durch die Analyse von geographischen Strukturen zur verringern soll eine Technologie in die bestehende Tourenplanungssoftware der Firma integriert werden, die folgende Aufgabenstellung automatisiert lösen kann: Anhand von bestehenden GPS-Aufzeichnungen sollen strukturelle Eigenschaften der jeweilige Sammelgebiete numerisch bewertet und klassifiziert werden. Gleichermaßen sollen anhand von geographischen (und möglichst frei verfügbaren Strukturdaten) aus noch unbekannten Gebieten erhoben werden können um diese auf die selbe Art und Weise klassifizieren zu können. Dadurch entsteht einerseits eine Referenzdatenmenge (von bestehenden Sammeltouren) und eine Vergleichsdatenmenge (aus den neuen Ausschreibungsgebieten). Dort wo die Klassifizierungsdaten übereinstimmen, kann davon ausgegangen werden, dass die planungsrelevanten Kennzahlen aus bestehenden Auftragsgebieten ohne gewagte Annahmen einfach übernommen werden können. Die Klassifizierung von GPS-Daten und geographischen Strukturdaten soll mit Hilfe von künstlicher Intelligenz automatisiert erstellt werden können. Auch die Überlegung, welche geographischen Strukturdaten denn überhaupt aussagekräftig sind um einen Vergleich anzustreben, sollen ggf. mit Hilfe von KI Technologien erfolgen.

Das Ziel der praktischen Arbeit ist es einen Sandbox-Service zu implementieren, der von der bestehenden Software der *infeo GmbH* aufgerufen und mit Daten befüllt werden kann um so "auf Knopfdruck" Klassifizierungen und Vergleiche von GPS-Daten und Ausschreibungs-Strukturdaten zu erstellen. Die Anwender:innen haben dadurch die Möglichkeit für neue Ausschreibungen entsprechend passende Planungsparameter aus ihren bestehenden Auftragsgebieten zu berechnen und somit die Unsicherheiten bei der Ausschreibungskalkulation deutlich zu reduzieren.

GPS-Datenklassifizierung, Abfallwirtschaft, Künstliche Intelligenz, Geografische Datenanalyse, Maschinelles Lernen, Automatisierung

Abstract

Classification of GPS Track Data Using AI Methods: A Case Study of Waste Collection Vehicles

In waste management, strategic route planning is a crucial process where optimal fleet utilization is determined through the efficient division of service areas, with the goal of minimizing costs. This process is applied by waste disposal companies both for existing service areas and when calculating bids for new tenders. Especially in regions where there is no prior experience, numerous uncertain assumptions and estimates must be made for robust route planning. To reduce these uncertainties through the analysis of geographical structures, a technology will be integrated into the company's existing route planning software, which can automatically solve the following task: Based on existing GPS records, the structural characteristics of the respective collection areas should be numerically evaluated and classified. Additionally, geographical structural data (preferably from freely available sources) from unknown areas should be collected and classified in the same way. This approach will create both a reference data set (from existing collection routes) and a comparison data set (from new tender areas). Where the classification data match, it can be assumed that planning-relevant parameters from existing service areas can be applied to the new areas without risky assumptions. The classification of GPS data and geographical structural data should be automated using artificial intelligence. Furthermore, the consideration of which geographical structural data are meaningful for comparison should, if necessary, also be supported by AI technologies.

The practical goal of this work is to implement a sandbox service that can be called and populated with data by the existing software of *infeo GmbH*, enabling the creation of classifications and comparisons of GPS data and tender structural data "at the push of a button." This will provide users with the ability to calculate appropriate planning parameters from their existing service areas for new tenders, thereby significantly reducing uncertainties in bid calculations.

GPS Data Classification, Waste Management, Artificial Intelligence, Geographic Data Analysis, Machine Learning, Automation

Contents

List of Figures **8**

List of Tables **9**

List of Abbreviations **10**

1 Introduction **11**

1.1 Motivation 11
1.2 Problem Statement 12
1.3 Solution Approach and Goal 12

2 Background and Related Work **13**

2.1 Technical Background 13
2.1.1 Feature Engineering for Geographic data 13
2.1.2 Clustering Algorithms 13
2.1.2.1 Density-Based Clustering with DBSCAN 13
2.1.2.2 k-means Clustering 14
2.1.2.3 Combining DBSCAN and k-means 15
2.2 Related Work 15
2.2.1 Transportation Mode Prediction with Feature Engineering 16
2.2.2 Clustering of GPS Data using Distance-based Features . 16

3 Problem Definition and Solution Approach **18**

3.1 Big Picture 18
3.2 Description of the Dataset 20
3.2.1 Overview 20
3.2.2 Source and Collection Method 20
3.2.3 Structure of the Data 20
3.2.4 Size and Coverage 22
3.2.5 Data compression 22
3.2.6 Limitations 22
3.3 Dataset Analysis 23
3.3.1 Sample Analysis 23
3.4 Solution Approach 27
3.4.1 Exploratory Data Analysis 28

3.4.2 Feature Extraction from Sample Dataset 28
3.4.2.1 Features Available in Trackings 28
3.4.2.2 Features Extracted from Waypoints 29
3.4.3 Filtering and Preprocessing the Full Dataset 29
3.4.4 Clustering and Pattern Discovery 29
3.4.5 Training a Classification Model 30
3.4.6 Comparison with OSM Geofences 30
3.4.7 API Integration 30
3.5 Structure of the Work 30

4 Implementation **31**
4.1 Implementation of the Big Picture 31
4.2 Data Handling and Preprocessing 32
4.2.1 Loading and Structuring the Dataset 32
4.2.2 Filtering Faulty Trackings 32
4.3 Feature Extraction 34
4.3.1 Available Metadata Features 34
4.3.2 Computed Features from Waypoints 34
4.3.2.0.1 Number of points 34
4.3.2.0.2 Bounding box area 34
4.3.2.0.3 Point density 34
4.3.2.0.4 Average segment distance 34
4.3.2.0.5 Number of stops 35
4.3.3 Feature Storage and Output 35
4.4 Clustering Implementation 35
4.4.1 DBSCAN for Outlier Removal 36
4.4.2 K-Means++ for Category Discovery 37
4.5 Classification Model 38
4.5.1 Model Selection and Training 38
4.5.2 Feature Importance Analysis 38
4.5.3 Saving and Exporting the Classifier 38
4.6 Integration with existing systems 38

5 Evaluation and Discussion **39**
5.1 Definition of the data sets used for the evaluation 39
5.2 Evaluation of the results 39
5.3 Reflection on the results 39

6 Conclusion **40**
6.1 Future Directions 40
6.2 Limitations 40

Literaturverzeichnis 41

Affidavit 43

List of Figures

2.1 DBSCAN: Core points and border points [7] 14
2.2 DBSCAN: (a) Density-reachability and (b) density-connectivity [7] 14
2.3 Comparison: k-means, DBSCAN clustering algorithms [8] 15
2.4 Steps of the framework for predicting of transportation modes proposed by Etemad et al. [9] 16
2.5 Steps of the framework proposed by Koh et al. [10] 17

3.1 High-level overview of training the GPS track classifier 19
3.2 Waypoint curvature calculation used for compression by *infeo GmbH* 22
3.3 Mapgrid for selected sample trackings 24
3.4 Pairplot of selected GPS route features grouped by area label . 25
3.5 Correlation Matrix of selected GPS route features grouped by area label 26
3.6 Boxplot of point density grouped by area label 27

4.1 Big picture implementation pipeline 31
4.2 Histogram 36
4.3 Correlation matrix 36
4.4 DBSCAN 37
4.5 K-means++ clustering 37

List of Tables

3.1 Structure of a Tracking Entry 21
3.2 Structure of a GPS Point Entry 21

4.1 Tracking reduction per filtering step 34
4.2 Overview of extracted features per tracking. 35

List of Abbreviations

GPS Global Positioning System

AI Artificial Intelligence

ML Machine Learning

API Application Programming Interface

CSV Comma-Seperated Values

DACH Germany, Austria and Switzerland

DBSCAN Density-Based Algorithm for Discovering Clutsers in Large Spatial Databases with Noise

1 Introduction

"The world's most valuable resource is no longer oil, but data" [1]

In today's digital age, where electronic devices are a part of everyones daily lives, increasing amounts of data are being generated every day, and this trend shows no signs of slowing down. [2] With this increase in data, businesses ranging across all industries recognize the importance of leveraging it for decision-making and operational efficiency. This has lead to a growing demand for technolgies that can gather insights from data and integrate seamlessly into strategic processes.

One industry in which data-driven decision-making is becoming increasingly important is the waste management industry.

1.1 Motivation

> "The Europe Waste Management Market size was valued at USD 116.21 billion in 2023, and is predicted to reach USD 169.37 billion by 2030, at a CAGR of 4.5% from 2024 to 2030."[3]

This growth reflects the increasing need for efficient and sustainable waste management practices, in response to rising waste volumes across residential, commercial, and industrial sectors.

> "The rapid population growth, urbanisation and economic development over the last decades has led to an increase in the generation of solid waste across the world. Providing high-quality waste management services is crucial to safeguard public health and protect the environment, but also to support resource efficiency, climate change mitigation and job creation." [4]

The waste management sector is considered critical infrastructure and is under growing pressure to adapt to personnel shortages, blackouts, and rising costs. In many cases, the only viable solution is to optimize and automate operational processes. Digital transformation plays a key role in making day-to-day operations significantly more robust and efficient.[5]

This is the mission of *infeo GmbH*, the software company for which this thesis is being conducted:

> “Die Abfallwirtschaft ist ein systemkritischer Wirtschaftsbereich und zählt zur kritischen Infrastruktur. Die Branche muss sich vor Personalausfällen, Blackouts und Kostensteigerungen schützen. Das gelingt oft nur durch Automatisierung und Optimierung von Prozessen, die das Tagesgeschäft maßgeblich robuster und effektiver machen. ”[5]

This thesis supports the mission of *infeo GmbH* to revolutionize digital processes in the waste management sector by implementing AI-powered tools to automatically classify and compare GPS track data and geographical structural data. The resulting service enables waste management companies to efficiently gather planning-relevant parameters from existing service areas and apply them to new regions, reducing uncertainty, increasing planning accuracy, and accelerating decision-making.

1.2 Problem Statement

Companies operating in the waste collection business have trouble calculating accurate bids for new service areas when expaning their field of business. They often have to make assumptions and rough estimates on several paremters concerning the operation cost in new service areas. A data driven estimation can help create more accurate and less risky assessments for unknown collection locations. This can help reduce uncertainties and improve the accuracy of bid calculations.

1.3 Solution Approach and Goal

Therefore, the goal of this work is to reduce uncertainty for waste collection companies expanding their business to new service areas using existing GPS tracking data. The approach involves classifying the GPS tracking data from known service areas into four categories ranging from urban to rural. AI methods are then used to cluster the data into these categories and to train a classifier able to automate the categorization. Furthermore, algorithms and AI techniques to compare patterns found in the exisiting data with public geographic data from OpenStreetMap. This functionality is then bundled into a ready to use API service. The endgoal is to provide an API which suggests planning parameters for a given input geofence.

2 Background and Related Work

2.1 Technical Background

2.1.1 Feature Engineering for Geographic data

Geospatial data is not inherently suitable for machine learning algorithms, which is why the extraction of features for each tracking is crucial for a performant and accurate classifier. Extraction of features for GPS data points consisting of latitude, longitude and time

2.1.2 Clustering Algorithms

> "Clustering is a useful tool in data science. It is a method for finding cluster structure in a data set that is characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters."[6]

2.1.2.1 Density-Based Clustering with DBSCAN

DBSCAN (Density-Based Algorithm for Discovering Clutsers in Large Spatial Databases with Noise) was introduced by Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiowei Xu as an algorithm to identify clusters of arbitrary shape in spatial databases and distinguish noise. It relies on the notion of density-reachability among points based on a distance threshold *Eps* and a minimum number of points *MinPts*.[7]

Let D be a database of points and let *dist(p, q)* be the distance between points p and q (typically Euclidean).

- The **Eps-neighborhood** of a point p is defined as:

$$N_{Eps}(p) = \{q \in D \mid \mathit{dist}(p,\ q) \leq \mathit{Eps}\}$$

- A point p is a **core point** if:

$$|N_{Eps}(p)| \geq \mathit{MinPts}$$

- A point q is **directly density-reachable** from p if:

$$q \in N_{Eps}(p) \textit{ and } p \textit{ is a core point}$$

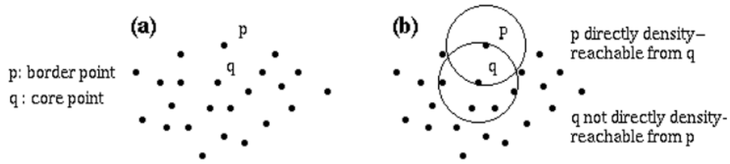

Figure 2.1: DBSCAN: Core points and border points [7]

- A point q is **density-reachable** from p if there is a chain of points $p_1, p_2, ..., p_n$ such that $p_1 = p$, $p_n = q$, and each p_{i+1} is directly density-reachable from p_i.
- Two points p and q are **density-connected** if there exists a point o such that both p and q are density-reachable from o.

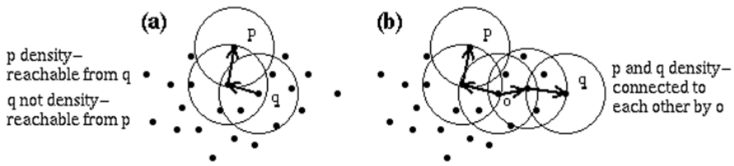

Figure 2.2: DBSCAN: (a) Density-reachability and (b) density-connectivity [7]

This concept allows DBSCAN to detect clusters without requiring the number of clusters as input, unlike other clustering algorithms such as K-Means and CLARANS, making it effective for outlier detection. [7]

2.1.2.2 k-means Clustering

K-Means is the first unsupervised algorithm used to cluster data in an euclidean space and is still widely used to partition a dataset $\{x_1, x_2, \ldots, x_n\}$ of size n into $k \leq n$ sets $S = \{S_1, S_2, \ldots, S_k\}$ [6]

The goal to minimize the within-cluster sum of squares (WCSS) can be shown with following function:

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

The implementation of the algorithm uses an iterative refinenement technique, also known as Lloyd's Algorithm. An initial set of k means is selected, randomly or are more sophisticatedly initialzed in adaptations such as k-means++. Then the algorithm repeats the following two steps until the assignments do not change and k-means therefore converges:

1. **Assignment:** Assign each data point to the cluster with the least squared Euclidean distance (mean):

$$S_i = \{x_j : \|x_j - \mu_i\|^2 \leq \|x_j - \mu_l\|^2, \forall\, 1 \leq l \leq k\}$$

2. **Update:** Update the means of data points assigned to each cluster:

$$\mu_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

2.1.2.3 Combining DBSCAN and k-means

DBSCAN can be used to identify anomalies by dividing the data set into two clusters. Cluster 0 with the valid data and cluster -1 with anomalies. Using the output of DBSCAN and removing the detected anomaly cluster -1 as the input for k-means, a more effective partitioning of the data into k number of groups can be achieved. Without DBSCAN, k-means clustering will likely assign one cluster to outliers of faulty data.

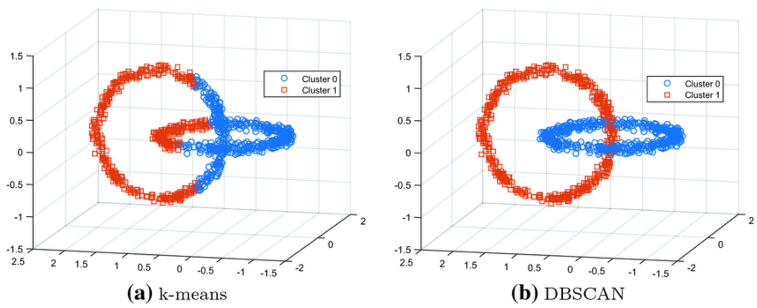

Figure 2.3: Comparison: k-means, DBSCAN clustering algorithms [8]

2.2 Related Work

GPS-Data analysis and classificatino has been studied in various sutdies and projects, such as indetifying transportation modes or movement behavior. This section highlights related approaches to the classification and comparison of GPS data are relevant to the problem of this thesis.

2.2.1 Transportation Mode Prediction with Feature Engineering

Etemad, Soared and Matwin [9] propose a five step framework to predict the underlying transportation mode from GPS traces. The steps of this framewok include in order the preparation of data, point feature extraction (e.g., distance, speed, bearing rate and its derivatitves), trajectory feature extraction (e.g., min, max, mean and std., percentiles), noise removal and normalization.

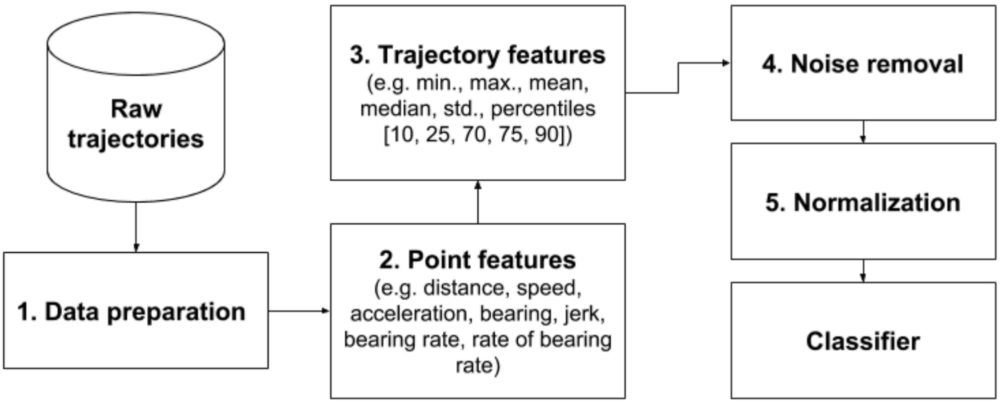

Figure 2.4: Steps of the framework for predicting of transportation modes proposed by Etemad et al. [9]

Their framework achieves competitive results with the classifier scoring 96.5% accuracy and a f1 score of 96.3%. The study shows the importance of noise reduction and feature engineering when working with GPS trace data. [9]

While their work focuses on the classification of transportation modes (e.g., walking, bike, car, etc.), the techniques used for extracting features and noise reduction, are also applicable for the structural classification of waste collection routes, which is the focus of this thesis.

2.2.2 Clustering of GPS Data using Distance-based Features

Koh et al. [10] present a framework to cluster users based on their movement patterns captured via an application on their mobile phones. The authors proposes a new metric, *Daily Characteristic Dinstance (DCD)*, which is used to create a fair comparison between working and nonworking users on on workdays and offdays and extract features in combination with *Origin-Destination (OD)* matrix features.

The features derived from the DCD are combined with OD matrix features and are used in a k-means clustering algorithm to group the users into three behavior groups. The study analyses the resulting clusters with two newly proposed metrics: *User Commonality* (percenctage of users that visited each point of interest (POI) category) and *Average Frequency* (average percentage of trips to each POI category). [10]

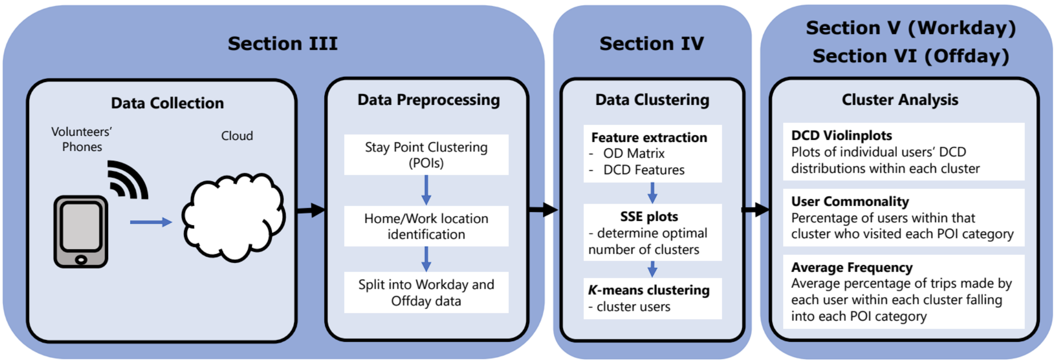

Figure 2.5: Steps of the framework proposed by Koh et al. [10]

The work of Koh et al. shows that meaningful patterns can be extracted from GPS data using unsupervised clustering methods. This approach is highly relevent for this thesis, as it demonstrates that structural characteristics can be derived from engineered features and effectively used for clustering and classification. Although the papers focus is on gps traces of individual people and captures data with mobile phones, the underlying methods can be applied for identifying patterns and simlaritiess in GPS track data collected by waste collection vehicles.

3 Problem Definition and Solution Approach

3.1 Big Picture

GPS trackings are collected by waste collection vehicles during real-world operation in the DACH region. These raw trackings are compressed by *infeo GmbH* to reduce storage size while keeping necessary route information, and then saved in a centralized database. The data provided is a subset approved by *infeo GmbH* for analysis in this thesis.

The aim of this thesis is to develop a pipeline that can autoamtically classify GPS trackings to four categories (e.g., rural, town, suburban and urban).

The proposed solution follows a high-level workflow with the following steps:

1. **Data Preperation:** Data is cleaned and filtered to remove invalid or incomplete trackings.
2. **Feature Extraction:** Meaningful features such as point density and bounding box area are calculated for each tracking.
3. **Outlier Detection:** Secondary detection of invalid trackings and trackings that significantly deviate from expected patterns are removed using density-based clustering (DBSCAN).
4. **Clustering:** The cleaned tracking data is grouped into four clusters based on similarity using K-Means.
5. **Classification:** A Classifier is trained to generalize the clustering results and allow automated classification of new trackings.

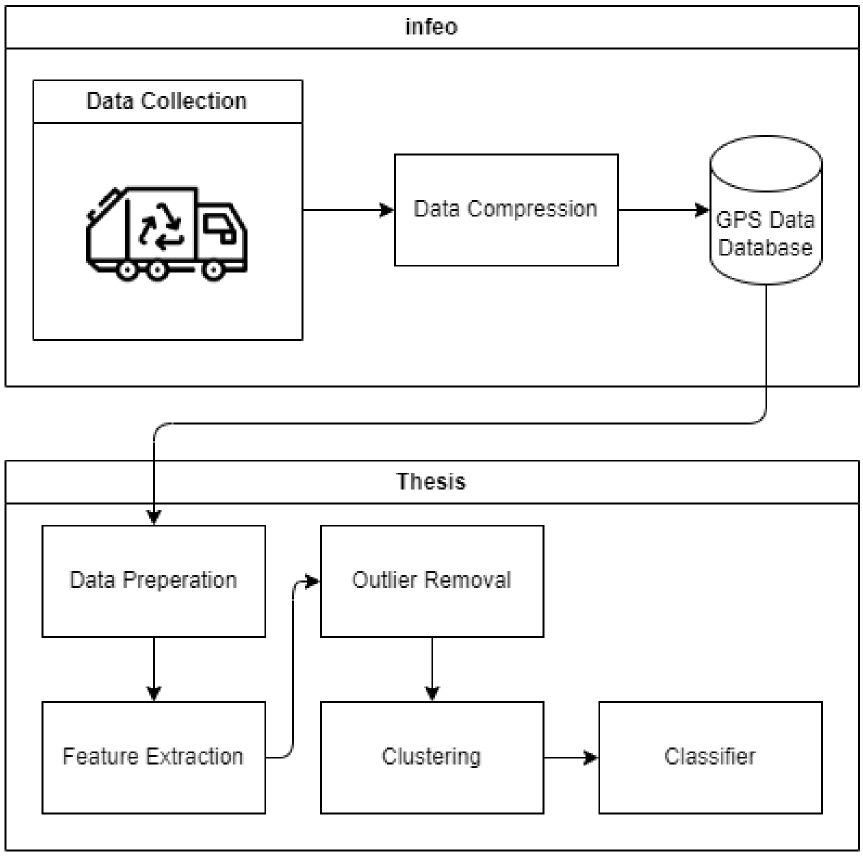

Figure 3.1: High-level overview of training the GPS track classifier

3.2 Description of the Dataset

Gaining an understanding of the provided data structure is crucial for identifying the available information and its limitations. This analysis enables the derivation of meaningful features and which types of information can be extracted from the dataset, and which cannot.

3.2.1 Overview

The dataset used is a collection of GPS tracking data collected by wastecollection vehicles from various wastecollection businesses and provided by *infeo GmbH*. It represents real-world data collected during regular wastecollection operation in the DACH region.

3.2.2 Source and Collection Method

The data was obtained by the onboard tracking systems installed by *infeo GmbH*, which collects GPS coordinates in regular intervals during regular operation. Each tracking represents a complete wastecollection route taken and includes metadata aswell as a list of GPS coordinates.

3.2.3 Structure of the Data

Each dataset entry represents a single recorded route refered to as *tracking* and contains metadata aswell as a time ordered list of gps coordinates.

Each tracking contains the following fields:

Table 3.1: Structure of a Tracking Entry

Field	Type	Description
`id`	Integer	Unique identifier of the tracking entry.
`name`	String	Name of the tracking (randomized for anonymization) identification.
`description`	String	Route metadata, often includes internal codes.
`recorded`	DateTime	Start date and time of the tracking.
`length`	Float	Total length of the route in kilometers.
`duration`	Integer	Total duration of the tracking.
`vehicleId`	Integer / Null	ID of the vehicle (nullified for anonymization).
`tourId`	Integer / Null	ID of the associated tour (nullified for anonymization).
`isExported`	Boolean	Flag indicating if the tracking was exported.
`editState`	Integer	Edit state used by the system.

Each GPS point contains the following fields:

Table 3.2: Structure of a GPS Point Entry

Field	Type	Description
`id`	Integer	Unique identifier of the GPS point.
`time`	DateTime	Timestamp of when the point was recorded.
`latitude`	Float	Latitude coordinate.
`longitude`	Float	Longitude coordinate.
`speed`	Float	Instantaneous speed at the time (in km/h).
`heading`	Float	Direction of movement in degrees.
`sequence`	Integer	Position of the point in the tracking sequence.
`metaTag`	Integer	Custom metadata tag.
`metaValue`	String	Value associated with the metadata tag.
`pointBaseType`	Integer	Internal point type used by the system.

3.2.4 Size and Coverage

The dataset consists of 91,452 individual trackings with a combined number of 123,785,460 waypoints. Trackings were recorded by waste collection vehicles operating in the DACH region.

3.2.5 Data compression

The dataset used in this thesis is pre-compressed internaly by *infeo GmbH* while keeping essential route information [11]:

1. **Delete same waypoints:** Consecutive waypoints with identical GPS position are deleted.

2. **Delete waypoints with distance smaller than (x):** Delete all Consecutive waypoints with a distance to each other of less than (x) meters.

3. **Delete waypoints with path curvature smaller than (x)^s:** For every group of three consecutive waypoints A, B, and C, the middle point B is deleted if all the following criteria are met:
    a) The distance $\overline{AB}$ is less than a specified threshold (e.g., 100 meters), **or** $\overline{BC}$ is less than the threshold.
    b) The distance $\overline{AC}$ is less than a specified threshold (e.g., 1000 meters).
    c) The change in heading α between segments $\overline{AB}$ and $\overline{BC}$ is smaller than a defined angle (e.g., 10°).

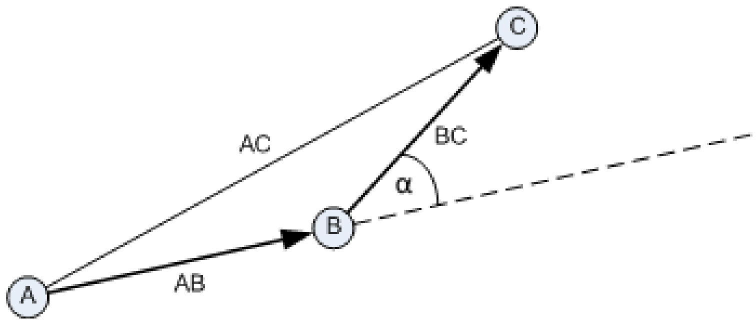

Figure 3.2: Waypoint curvature calculation used for compression by *infeo GmbH*

[11]

3.2.6 Limitations

Missing Values: GPS gaps etc, useless trackings etc.

With a lot of data, also comes a lot of unfurbished and faulty data. Through communication with *infeo GmbH* and through visual examination, it is clear that many trackings are not collected properly. Since the tracking devices installed on the garbage collecting vehicles are operated partly manually, some trackings show GPS tracks only on the parking facility and not in operation, while others show that the tracking device being turned on over night, while no operation are taking place either.

3.3 Dataset Analysis

Formatvorlage für den Fließtext.

3.3.1 Sample Analysis

A small, manually selected sample of 8 tracking routes was selected for initial exploratory data analysis. Each route was inspected on the AWM-Map-Tool and then categorized into on of the four area type labels: RURAL, SUBURBAN, TOWN or URBAN. Each Label is represented by 2 tracking routes in the sample data to ensure a balanced representation.

Feature extraction was performed to calculate route-level metrics such as length, duration, bounding box area, point density, number of stops and average distance between points.

The goal of this sample is to explore patterns, validate assumptions, and identify features useful for future automatic classification.

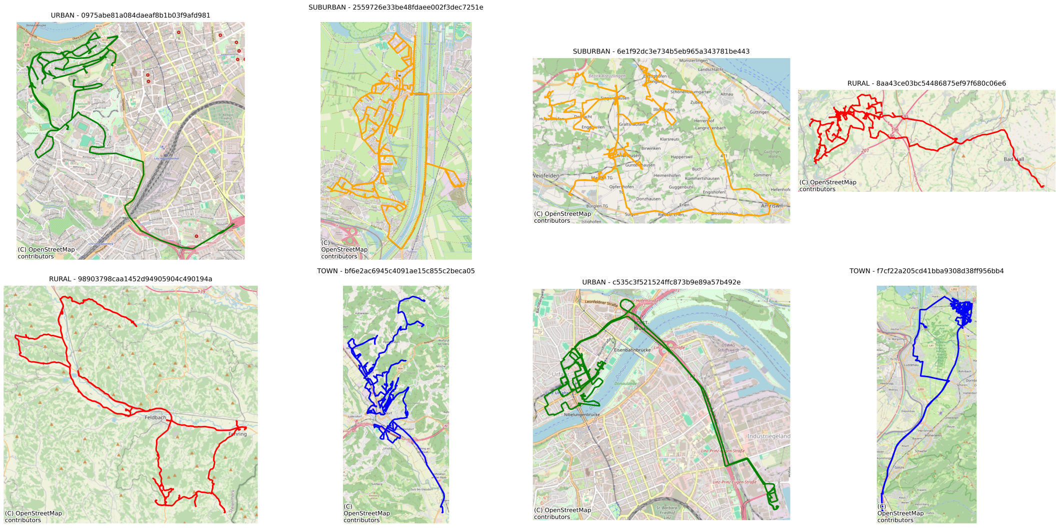

Figure 3.3: Mapgrid for selected sample trackings

The mapgrid shows spatial layout of each tracking route overlayed on a visual map. Each subplot corresponds to a single tracking and is colorcoded with the assigned lable (RURAL=Red, SUBURBAN=Orange, TOWN=Blue, URBAN=Green). This visual representation clearly shows the difference in route shapes and sizes between the four different area types.

Urban:

- Routes are geographically compact and highly localized.
- Movement appears dense with short travel distances between stops.
- Often confined to a small cluster of city blocks or neighborhoods.

Town:

- Coverage is slightly more dispersed than urban routes.
- Still relatively compact but less tightly packed.
- Serves a central area and nearby residential surroundings.

Suburban:

- Routes extend farther and cover wider areas than town routes.
- Show transitional behavior between urban and rural structures.
- Less dense stop distribution, indicating more spaced-out residential zones.

Rural:

- Routes are long and span large geographical areas.
- Stops are widely spaced, often connecting small, isolated settlements.
- The shape and path vary significantly, often following main roads between distant collection points.

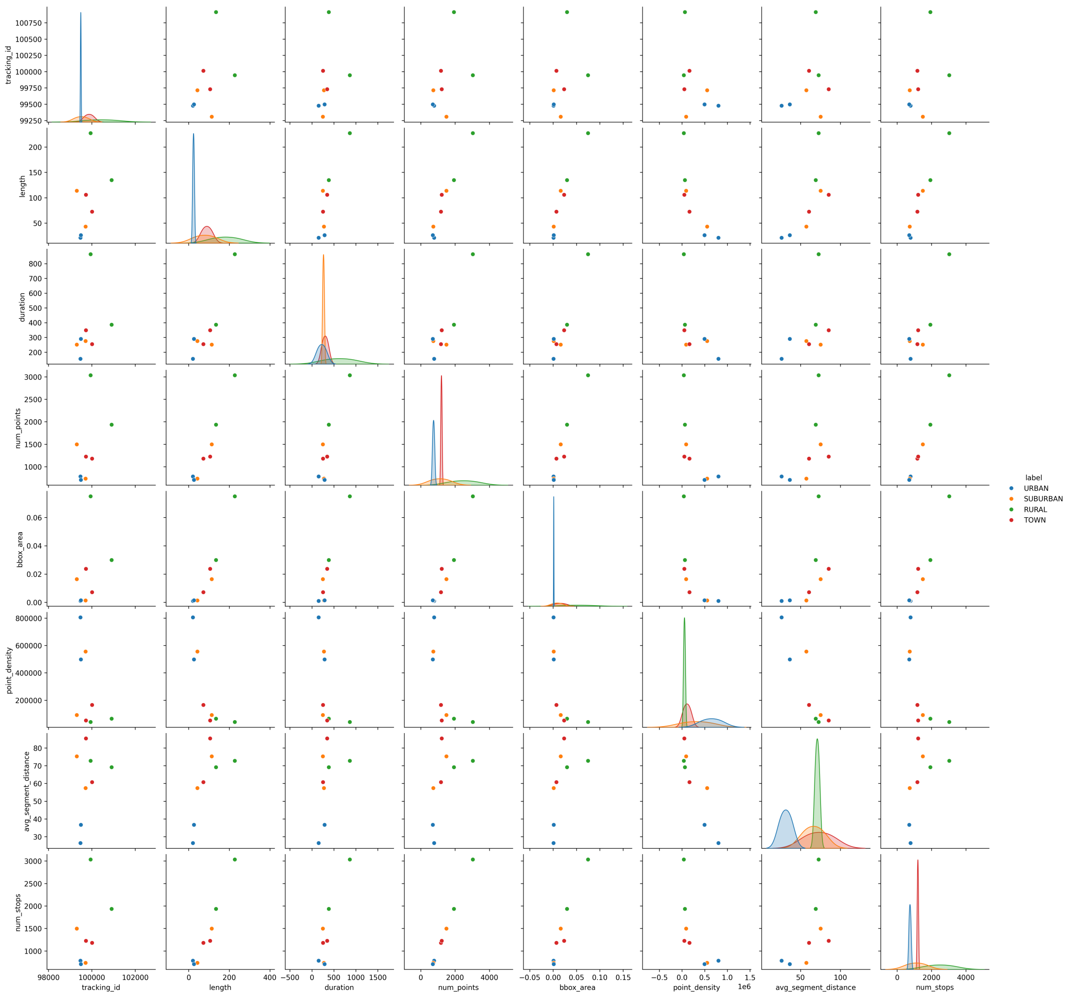

Figure 3.4: Pairplot of selected GPS route features grouped by area label

The pairplot compares the relationships between the extracted features (eg., length, duration, number of points, bounding box area, point density, average segmengt distance and number of stops). Rural and urban trackings show a clear difference in multiple features. Point density and average segment distance are especially good discriminators. Suburban and town trackings show more variance and occasionally overlap with eachother, giving a not so clear distinction.

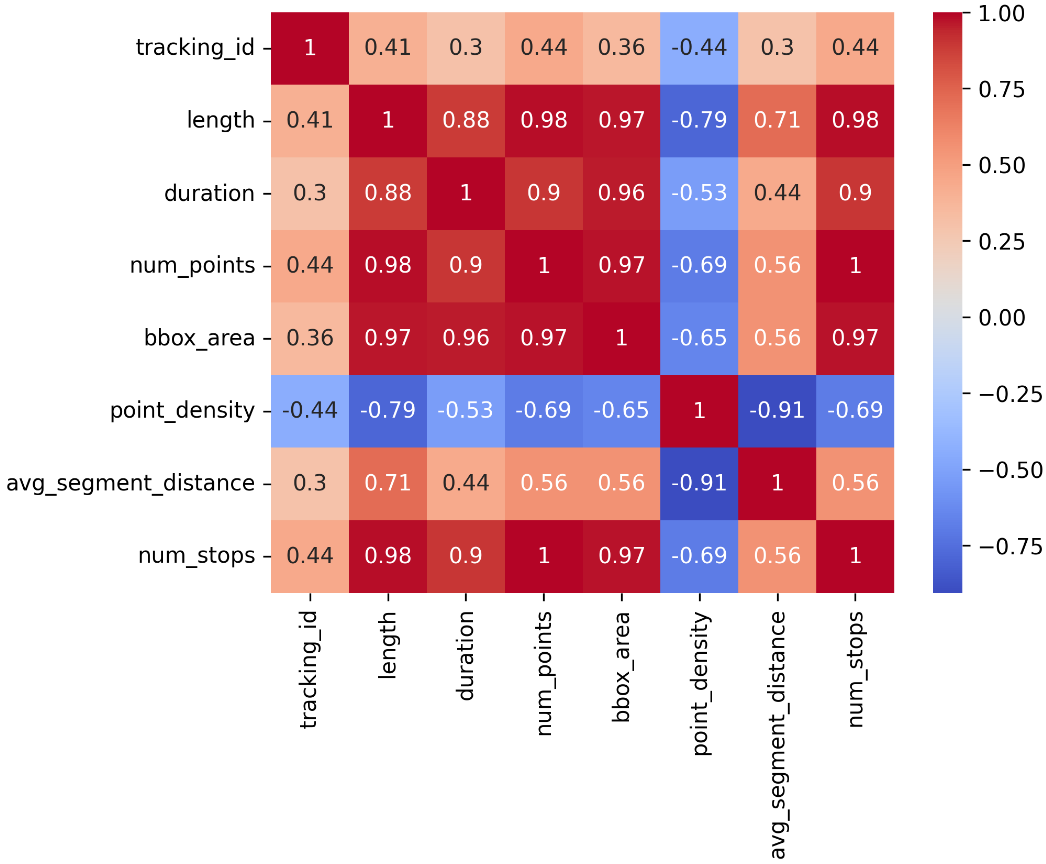

Figure 3.5: Correlation Matrix of selected GPS route features grouped by area label

The Correlation Matrix highlights the strong correlations between features in the selected sample. A high posive correlation between length, duration and number of points can be observerd. Additionally the point density has a strong negative correlation with the tracking length, number of points and average segment distance.

This correlation is to be expected and confirms the consistency of the data, since the points (GPS coordinates) are recorded in uniform intervals. This plot also indicates the possible redundance of features such as number of points and number of stops.

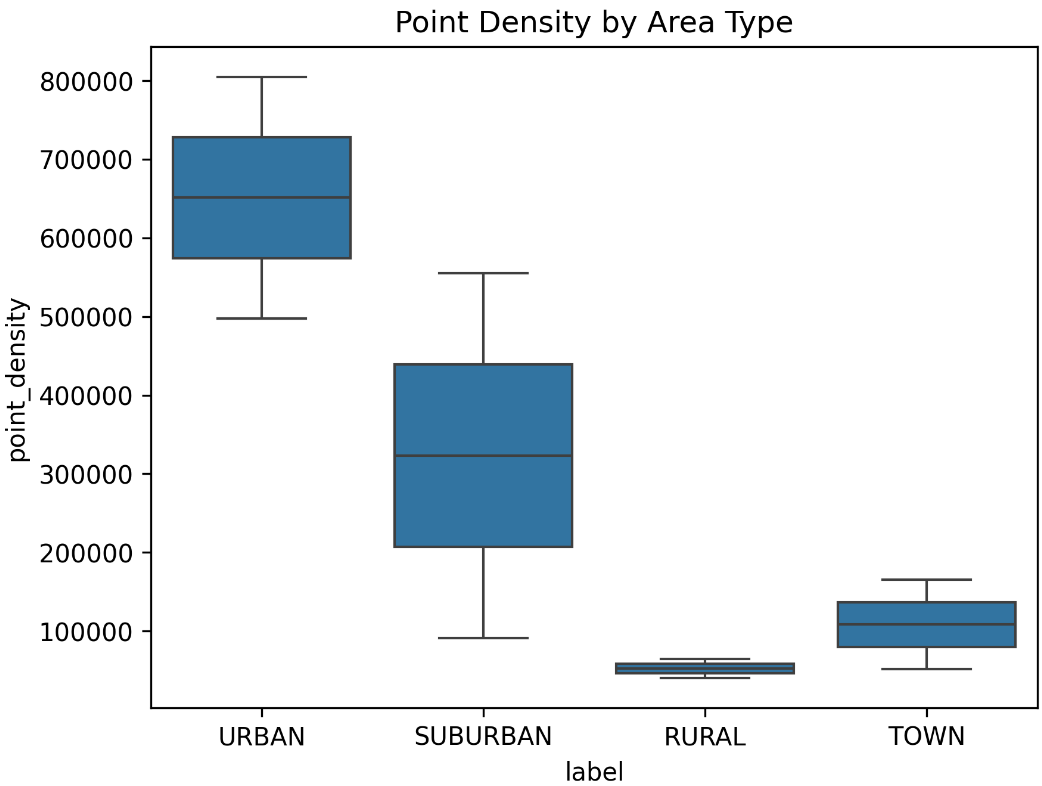

Figure 3.6: Boxplot of point density grouped by area label

The point density boxplots for each label shows a clear difference between the labels. Urban trackings appear to have the highest density and rural trackings the lowest, with suburban and town trackings falling in the middle with a wider variability.

This validates that the point density is a strong feature for classifying trackings to the four labels.

3.4 Solution Approach

This chapter outlines the methods used to analyse and claaify the GPS tracking data collected by waste collection vehicles. The approach is structured in multiple stages, beginning with exploratory data analysis, proceeding with feature extraction, data cleaning, outlier detection, clustering and finally training a classifier. The goal is to identify tracking patterns that allow for structural classification of trackings into four categories: Urban, Suburban, Town and Rural.

3.4.1 Exploratory Data Analysis

The foundation of this thesis is the dataset provided by *infeo GmbH*. It contains anonymized GPS track data collected by waste collection vehicles during real world operation. Due to its large size of more than 100.000 GPS tracks and its anonymization, a clear picture of the data needed to be established as the first step. With visual examination of different random trackings a sample of eight manually labeled trackings was gathered. The selected trackings were labeled to four categories, two trackings per category.

- Urban (Dense building structures, complex road networks, and minimal spacing between stops)
- Suburban (Moderately spaced residential areas with organized street layouts)
- Town (Smaller clustered settlements with simpler road structures)
- Rural (Sparse housing, long roads, large distances between collection points)

With the manually selected trackings varying in geographical structure, a exploratory data analysis helped transfer the visual distinction to clear differences found in the data of each tracking category.

3.4.2 Feature Extraction from Sample Dataset

To enable the use of machine learning algorithms, each tracking originally represented by a consecutive list of GPS waypoints with latitude and longitude must be transformed into feature vectors. Feature extraction methods were applied to derive representative metrics for each tracking.

This approach is well established and used in similar studies such as the classification of transportation modes by Etemad et al. [9] and clustering of GPS data using distance-based features by Koh et al. [10]

3.4.2.1 Features Available in Trackings

1. **Length:** Total distance of the route, in kilometers.
2. **Duration:** Total time spent for the trackings, captured.

3.4.2.2 Features Extracted from Waypoints

1. **Number of Points:** The total number of GPS waypoints recorded in the tracking.
2. **Bounding Box Area:** The area of the minimum bounding box that encloses all latitude and longitude waypoints in one tracking.
3. **Point density:** The number of waypoints divided by the bounding box area. This feature indicates how closes packed the points are.
4. **Average segment distance:** The mean distance between consecutive waypoints, computed using geodisc distance.
5. **Number of Stops:** Waypoints with speed of zero. Originally considered as a potential feature, it was found out to be redundant as most waypoints have a speed of zero.

3.4.3 Filtering and Preprocessing the Full Dataset

There are a lot of faulty datapoints that are not relevant for the clustering, that are filtered out. Before applying machine learning methods, extensive filtering was required to remove invalid and irrelevant data from the dataset. These included:

1. Trackings with a too low duration to be realistic collection trackings.
2. Trackings that recorded movement only in and around a parking lot.
3. Trackings left running overnight with no significant movement.
4. Trackings with unrealistic GPS jumps and anomalies (possibly due to device errors or compression errors introduced by *infeo GmbH*).

Removing these faulty datapoints was curcial for ensuring the quality of the clustering and reducing the computation times.

3.4.4 Clustering and Pattern Discovery

To discover patterns in the unlabeled dataset, clustering algorithhms were applied. Initially, K-Means++ was used to cluster the data into four clusters. However the result showed that one cluster mostly consisted of outlier routes, with unrealistic GPS jumps.

To solve this issue, the process was improved by detecting and removing outliers using DBSCAN (Density-Based Spatial Clustering for Applications with Noise). DBSCAN effectively removed most anomalies in the dataset. After filtering the outliers with DBSCAN, K-Means++ clustering was applied again, resulting clusters that represented the four defined categories better.

3.4.5 Training a Classification Model

To automatically classify new trackings into one of the four categories (Urban, Suburban, Town, Rural), a supervised machine learning model was trained using the labeled dataset from the clustering step.

The labeled dataset was split into training and test sets using an 80/20 split. Evaluation of the classifier on the test sset showed a high overall accuracy.

The trained classifier was saved for the later use in a sandbox API to classify new trackings.

3.4.6 Comparison with OSM Geofences

3.4.7 API Integration

3.5 Structure of the Work

The remainder of this thesis is structured in the following chapters:

- **Chapter 4: Implementation:** Provides a detailed description of how the previously introduced methos and algorithms were implemented. This includes the data preprocessing, feature extraction, clustering pipeline and classifier training.
- **Chapter 5: Evaluation and Discussion:** Presents the results of the clustering and classification. Performance metrics, visualizations, and analysis of the results are discussed alongside future reasearch and application.
- **Chapter 6: Conclusion:** Summarization of the contributions of the thesis, highlighting key takeaways and shortcomings.

This structure follows the development of the project, from understanding the problem and analysing the data, to buildinng and evaluating a solution to address the objective of classifying GPS trackings into structural categories.

4 Implementation

4.1 Implementation of the Big Picture

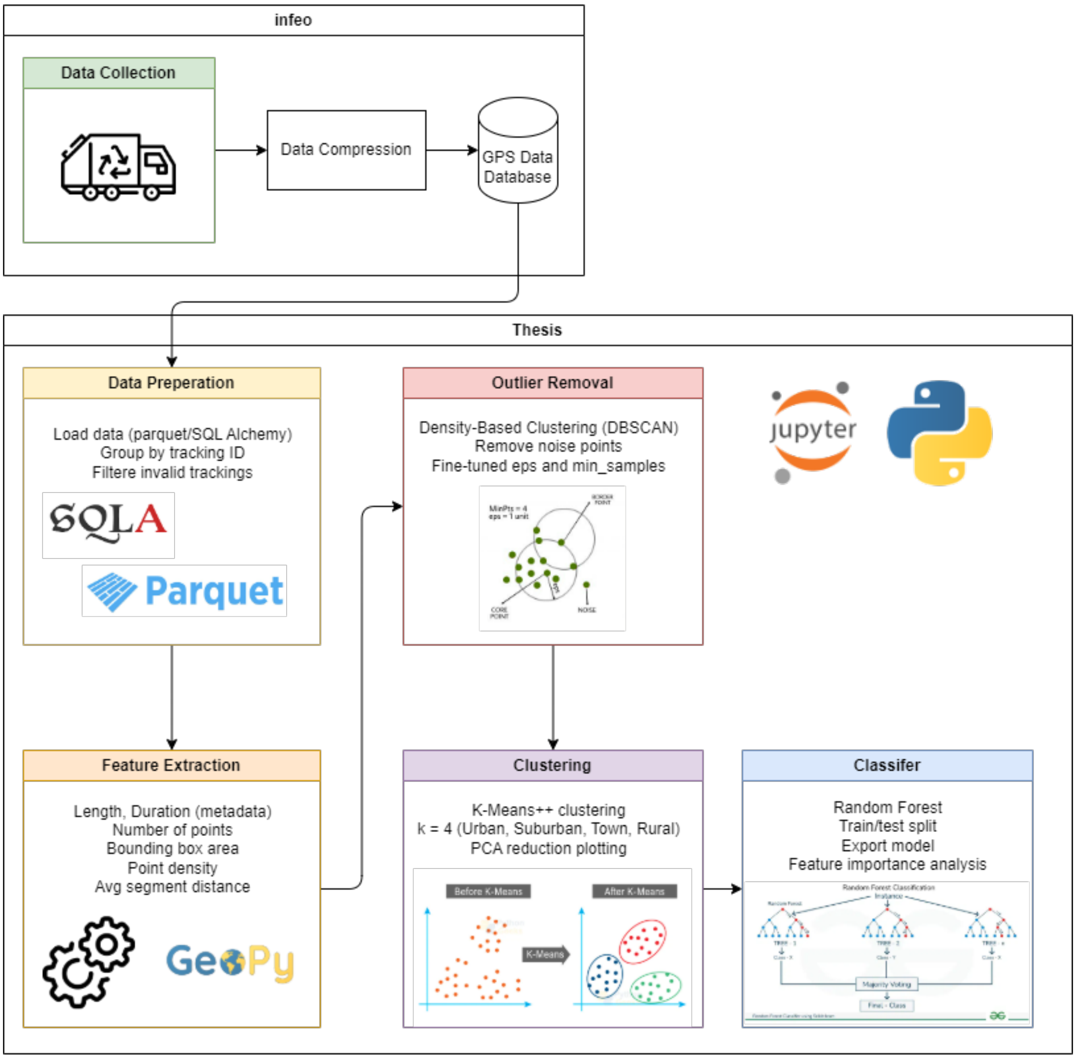

Figure 4.1: Big picture implementation pipeline

4.2 Data Handling and Preprocessing

Preprocessing is a crucial step in the machine learning pipeline.

Each row in the dataset is a single GPS waypoint and consist of the following columns:

- **id_tracking:** Unique identifier for each tracking.
- **sequence:** The order of the waypoint in the tracking.
- **latitude/longitude:** GPS coordinates.
- **speed:** Vehicle speed at the waypoint.
- **tracking_duration:** Total time of the tracking.
- **tracking_length:** Total dinstance of the tracking in km.

4.2.1 Loading and Structuring the Dataset

4.2.2 Filtering Faulty Trackings

The raw data included many invalid or irrelevant trackings that could impact the analysis and clustering. These trackings were removed based on these criteria:

- **Duration Filter:** Trackings shorter than 30 minutes or longer than 10 hours were excluded.
- **Length Filter:** Trackings with a total length lower than 50 km or above 850km were excluded.
- **Minimal Waypoint Count:** Trackings with less than 10 recorded waypoints were excluded.
- **Minimal Bounding Box:** Trackings with a bounding box smaller than about 50 meters in latitude or longitude were excluded. This was a helpful step to reduce the amount of trackings that were exclusively recorded in a parking lot.

The filters were directly implemented as a SQL query executed in an jupyter notebook. This reduced the further computatinal times for clustering and training the classifier model. The SQL query implementing these filters looked as follows:

```sql
SELECT
    wp.id_tracking, wp.id, wp.time, wp.type, wp.sequence,
        wp.comment,
    wp.speed, wp.heading, wp.duration, wp.block_type, wp.
       log,
    wp.latitude, wp.longitude, wp.altitude, wp.meta_tag,
       wp.meta_value,
    t.length AS tracking_length,
    t.duration AS tracking_duration
FROM waypoint wp
JOIN tracking t ON wp.id_tracking = t.id
WHERE
    t.duration BETWEEN 18000000000 AND 360000000000
    -- 0.5 to 10 hours
    AND t.length BETWEEN 50 AND 850  -- 50 to 850 km
    AND EXISTS (
        SELECT 1 FROM waypoint w
        WHERE w.id_tracking = t.id
        HAVING COUNT(*) > 10  -- Min. 10 waypoints
    )
    AND (
        (SELECT MAX(latitude) FROM waypoint WHERE
           id_tracking = t.id) -
        (SELECT MIN(latitude) FROM waypoint WHERE
           id_tracking = t.id)
    ) > 0.0005  -- Min. ~50m latitude span
    AND (
        (SELECT MAX(longitude) FROM waypoint WHERE
           id_tracking = t.id) -
        (SELECT MIN(longitude) FROM waypoint WHERE
           id_tracking = t.id)
    ) > 0.0005;  -- Min. ~50m longitude span
```

Listing 4.1: SQL query used for filtering the GPS tracking dataset

Each Filter step reduces the dataset by this many trackings:

Filter Step	Remaining Trackings	Removed	Reduction (%)
All trackings	101353	–	–
Duration filter	66477	34876	34.410%
Length filter	45668	20809	31.303%
Minimal Waypoints filter	45599	69	0.151%
Min Lat/Long span filter	45598	1	0.002%
Total Remaining	**45598**	**55755**	**55.011%**

Table 4.1: Tracking reduction per filtering step

4.3 Feature Extraction

4.3.1 Available Metadata Features

Tracking length Tracking duration

4.3.2 Computed Features from Waypoints

4.3.2.0.1 Number of points The total number of valid GPS points n_p for a given route:

$$n_p = |\{(lat_i, lon_i)\}|$$

4.3.2.0.2 Bounding box area The bounding box area A_{bbox} is calculated as:

$$A_{bbox} = (\max(lat_i) - \min(lat_i)) \cdot (\max(lon_i) - \min(lon_i))$$

This gives an approximation of the route's spatial extent. A similar measure is used in **yang2009spatial** for comparing track spread.

4.3.2.0.3 Point density To measure how dense the GPS sampling is in space, we define point density d_p as:

$$d_p = \frac{n_p}{A_{bbox} + \varepsilon}$$

where ε is a small constant to avoid division by zero.

4.3.2.0.4 Average segment distance The average distance between consecutive waypoints d_{avg}:

$$d_{avg} = \frac{1}{n_p - 1} \sum_{i=1}^{n_p-1} dist(p_i, p_{i+1})$$

where *dist* refers to the geodesic distance between two GPS points (in meters), calculated using the WGS84 ellipsoid model.

4.3.2.0.5 Number of stops

A stop is defined as a waypoint where the recorded speed equals zero. The number of such points n_s is computed as:

$$n_s = |\{i \mid speed_i = 0\}|$$

4.3.3 Feature Storage and Output

The extracted features were stored in a structured csv file, with one row per tracking_id and all corresponding features. This resulting dataset forms the basis for all future analysis, clustering and classification applications in the next steps of the pipeline.

Feature	**Description**
`num_points`	Total GPS waypoints in the track
`bbox_area`	Area covered by the track's bounding box
`point_density`	Points per unit area
`avg_segment_distance`	Mean distance between consecutive points
`num_stops`	Count of zero-speed positions
`duration`	Duration of tracking (from metadata)
`length`	Route length in meters (from metadata)

Table 4.2: Overview of extracted features per tracking.

4.4 Clustering Implementation

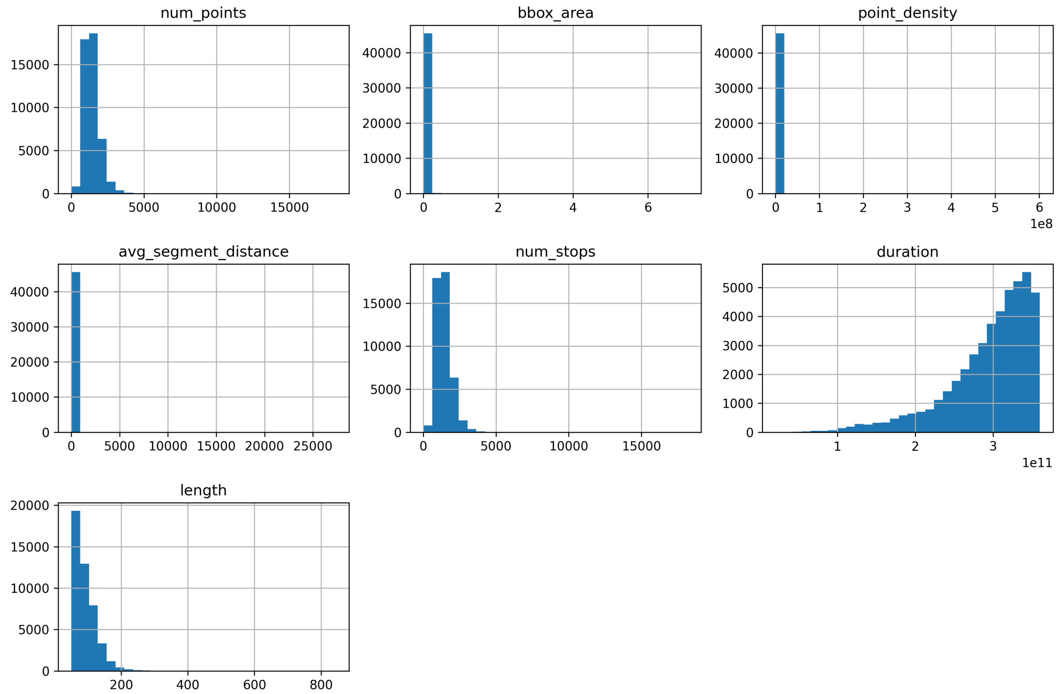

Figure 4.2: Histogram

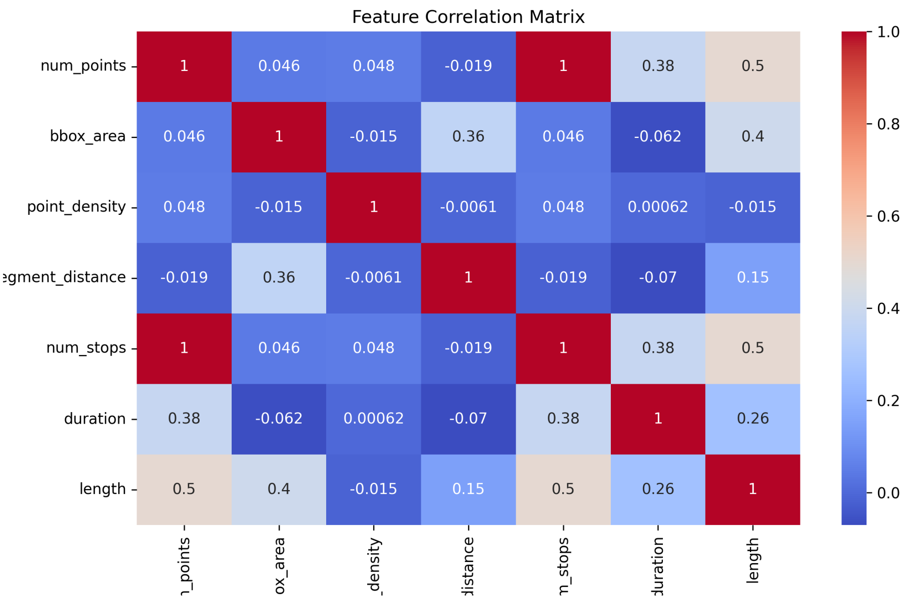

Figure 4.3: Correlation matrix

4.4.1 DBSCAN for Outlier Removal

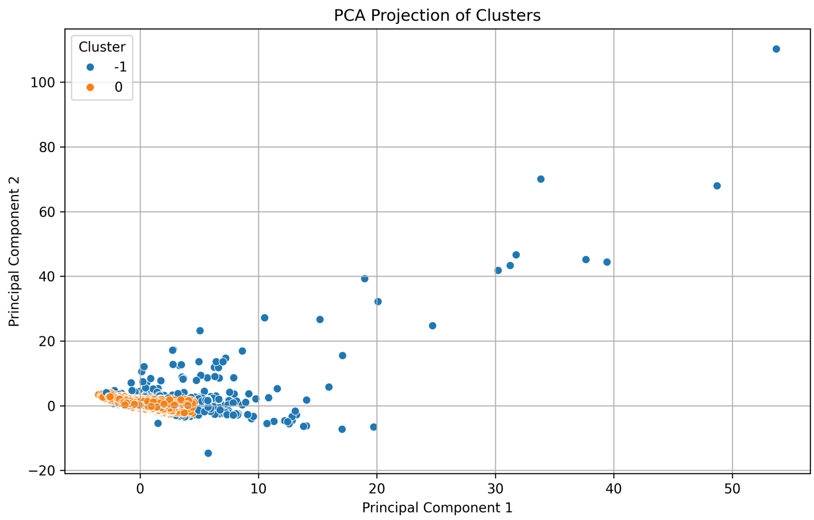

Figure 4.4: DBSCAN

4.4.2 K-Means++ for Category Discovery

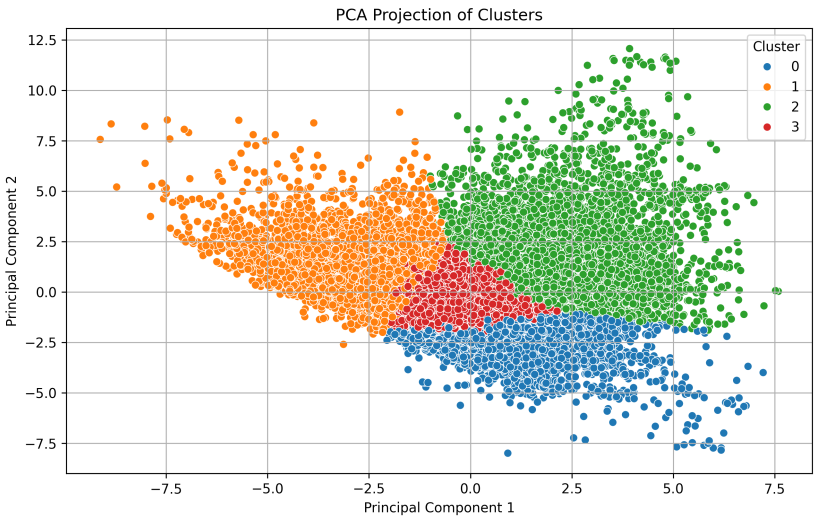

Figure 4.5: K-means++ clustering

4.5 Classification Model

4.5.1 Model Selection and Training

4.5.2 Feature Importance Analysis

4.5.3 Saving and Exporting the Classifier

4.6 Integration with existing systems

Formatvorlage für den Fließtext.

5 Evaluation and Discussion

5.1 Definition of the data sets used for the evaluation

Formatvorlage für den Fließtext.

5.2 Evaluation of the results

Formatvorlage für den Fließtext.

5.3 Reflection on the results

Formatvorlage für den Fließtext.

6 Conclusion

6.1 Future Directions

Formatvorlage für den Fließtext.

6.2 Limitations

Formatvorlage für den Fließtext.

Literaturverzeichnis:

Bibliography

[1] "The world's most valuable resource is no longer oil, but data," *The Economist*, ISSN: 0013-0613. [Online]. Available: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (visited on 04/06/2025).

[2] T. Petroc, *Data growth worldwide 2010-2028*, en. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/ (visited on 04/11/2025).

[3] *Europe Waste Management Market Size and Share Analysis 2030*, en. [Online]. Available: https://www.nextmsc.com/report/europe-waste-management-market (visited on 04/24/2025).

[4] *Solid waste management*, en. [Online]. Available: https://www.eib.org/en/projects/topics/energy-natural-resources/solid-waste/index (visited on 04/24/2025).

[5] *Gemeinsam mit unseren Kunden revolutionieren wir die digitale Kreislaufwirtschaft*, de-AT. [Online]. Available: https://infeo.at/index.php/revolution/ (visited on 04/24/2025).

[6] K. P. Sinaga and M.-S. Yang, "(PDF) Unsupervised K-Means Clustering Algorithm," en, *ResearchGate*, Dec. 2024. DOI: 10.1109/ACCESS.2020.2988796. [Online]. Available: https://www.researchgate.net/publication/340813602_Unsupervised_K-Means_Clustering_Algorithm (visited on 04/25/2025).

[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," en,

[8] J. Liang, J. Cui, J. Wang, and W. Wei, "Graph-based semi-supervised learning via improving the quality of the graph dynamically," en, *ResearchGate*, Dec. 2024. DOI: 10.1007/s10994-021-05975-y. [Online]. Available: https://www.researchgate.net/publication/351571830_Graph-based_semi-supervised_learning_via_improving_the_quality_of_the_graph_dynamically (visited on 04/25/2025).

[9] M. Etemad, A. S. Junior, and S. Matwin, "Predicting Transportation Modes of GPS Trajectories using Feature Engineering and Noise Removal," in vol. 10832, arXiv:1802.10164 [cs], 2018, pp. 259–264. DOI: 10.1007/978-3-319-89656-4_24. [Online]. Available: http://arxiv.org/abs/1802.10164 (visited on 06/05/2025).

[10] Z. Koh, Y. Zhou, B. P. L. Lau, R. Liu, K. H. Chong, and C. Yuen, *Clustering and Analysis of GPS Trajectory Data using Distance-based Features*, arXiv:2212.00206 [cs], Dec. 2022. DOI: 10.48550/arXiv.2212.00206. [Online]. Available: http://arxiv.org/abs/2212.00206 (visited on 06/05/2025).

[11] *Route zum Nachfahren bereitstellen – awm.* [Online]. Available: https://wiki.awm.infeo.at/index.php/Route_zum_Nachfahren_bereitstellen#Aufzeichnung_komprimieren (visited on 06/05/2025).

Affidavit

I hereby declare in lieu of oath that I have written this Bachelor thesis independently and without the use of aids other than those specified. The passages taken directly or indirectly from other sources directly or indirectly from other sources are marked as such. The thesis has not been neither in the same nor in a similar form to any other examination authority nor has it been published.

Dornbirn, on 15. May 2025 Matthias Hefel