An AI-based Approach to Intelligent Waste Collection Optimization

Mattéo Audigier Jakob Leitner Michał Binda
Claudia De Carlo

Barcelona School of Informatics (FIB-UPC)
Intelligent Systems Project – Milestone 1: Project Definition

October 2025

Contents

1 Introduction

1.1 Definition of the Problem

Efficient waste collection is a key challenge for modern cities aiming to become more sustainable. Poorly timed pickups or inefficient routes increase CO_2 emissions, operational costs, and overall environmental impact; moreover, delayed collections lead to overflowing bins, litter and odors, degrading street cleanliness and public hygiene. At the same time, the growing deployment of Internet of Things (IoT) sensors in waste bins enables real-time monitoring of fill levels, opening the door to smarter and more efficient waste management strategies.

The goal of this project is to design an AI-based system capable of predicting the filling levels of waste bins and explicitly leaving route optimization/pathfinding as a downstream extension for future contributors. We start from real-world sensor data collected in Melbourne's Argyle Square Smart Bins dataset, while keeping our approach flexible enough to generalize later to other cities.

1.2 Problem Statement

Given a network of bins equipped with sensors, we aim to predict when each bin is likely to reach its maximum capacity. The final objective is to create a data-driven intelligent system that opens the doors for reduced unnecessary trips, lower emissions and improvements in urban service efficiency.

1.3 Challenges and Complexities

- Sensor data are often incomplete, noisy or limited to short time periods.
- External factors e.g., weather, holidays and area characteristics (such as proximity to commercial zones) influence waste generation. Therefore, not all variance is explainable from the available dataset, with some patterns driven by unobserved factors.
- Ensuring generalization across different cities and data sources is potentially non-trivial.

2 Literature Review

2.1 Previous Studies

A recent and relevant paper, *Forecasting the Status of Municipal Waste in Smart Bins* [1], introduced a machine-learning-based approach for predicting bin fill levels using IoT sensor data. The authors obtained interesting results, but their approach did not include external contextual variables such as holidays or location meta data and it was not directly compared to other optimization-based systems.

In our project, we plan to:

- Reproduce their experimental results as a baseline.

- Extend the approach by evaluating alternative regressors and integrating additional data sources.
- Evaluate simpler regression and time-series baselines (SARIMAX/Prophet, tree-based regressors).
- Compare our approach to theirs to better understand its effectiveness and limitations.

2.2 Relevant AI Techniques

- **Gradient Boosting (XGBoost):** Gradient Boosting builds an ensemble of decision trees that iteratively correct errors, which makes it effective on tabular data with nonlinear relationships. In our setting, it can naturally exploit lagged targets and rolling-window statistics and provides model explanations (e.g., SHAP) [2].
- **SARIMAX:** SARIMAX is a classical time-series model that combines autoregressive and moving-average components with differencing and allows exogenous regressors such as weather or holidays. It is useful here as an interpretable baseline that captures weekly/seasonal patterns and yields confidence intervals out of the box [3]
- **Prophet / NeuralProphet:** Prophet models time series as an additive combination of trend, multiple seasonalities and holiday effects, while NeuralProphet augments this with autoregressive terms and a light neural component [4].

3 Project Overview

3.1 Goals

Our goals are to develop a predictive model for bin fill levels using the Melbourne dataset, to evaluate its predictive and computational performance, and to design the system so that it generalizes to other cities and datasets. Furthermore, we will compare simple and advanced forecasters, report interpretability and expose outputs in a format so that route optimization/pathfinding can be added downstream.

3.2 Motivation

Waste collection is simultaneously an environmental, operational, and social challenge. Static pickup schedules can lead to avoidable kilometres, higher fuel use and CO_2 emissions, and, when collections are delayed, overflowing bins and subsequently reduced public hygiene. With the increasing availability of IoT sensors and open data, there is a timely opportunity to forecast demand more accurately and inform better decisions.

This project focuses exclusively on forecasting bin fill levels and producing reliable outputs. Route planning/pathfinding is explicitly out of scope. Our forecast's output will be designed to be easily consumable by downstream systems.

We expect benefits across the following dimensions:

- **Sustainability:** Fewer unnecessary trips and lower CO_2/energy use (when forecasts are integrated by downstream routing systems).

- **Public hygiene & equity:** Reduced overflow risk and more consistent service across neighbourhoods.
- **Operational efficiency:** Better planning of crews and fleets through more reliable forecasts.
- **Scalability:** The system should generalize well across different cities.

3.3 Requirements

We define the following requirements:

3.3.1 Functional

- Predict the future fill level of bins.
- Optimize waste collection routes dynamically.
- Support continual learning by retraining (on schedule) as new data arrives.
- Provide decision support.
- Provide downstream systems with formatted, documented forecast outputs for easy integration.

3.3.2 Non-Functional

- Scalable to different city sizes and data sources.
- Robust to missing or noisy data.

4 Initial Strategy

The project will be developed in three stages:

1. **Data Analysis:** Explore, clean and understand the Melbourne dataset. Identify trends and anomalies.
2. **Prediction:** Build baseline models and later experiment with more modern architectures.
3. **Evaluation:** Quantify performance with e.g RMSE/R^2 and compare with the system proposed in [1].

5 Preliminary Analysis

5.1 Dataset Description: Argyle Square Smart Bins

The dataset used in this work originates from the public waste bins managed by the Wyndham City Council in Melbourne, Australia. It is publicly available through the Australian Government open data portal. The dataset [5] spans the period from July 2018 to May 2021 and is provided in JSON format, where records are organized under descriptive JSON tags.

It includes sensor readings from 32 smart bins, with data collected daily from various public locations within the Wyndham City area. Each entry contains several attributes describing the bin's characteristics and status, including its geographical coordinates, fill-level and timestamp. Additional fields define identifiers, descriptive labels, and operational status indicators (e.g., empty or full).

Before model training, the dataset undergoes pre-processing to handle missing values and ensure data consistency, forming a reliable foundation for the subsequent tasks. When appropriate, we will enrich the dataset with external features retrieved via third-party APIs (e.g., weather variables and location metadata).
Although this dataset is geographically limited to Argyle Square, it offers a solid foundation for testing predictive methods and building an initial prototype. Our system shall be extended and tested with other sources once additional data becomes available.

5.2 Future Data Expansion and Model Improvement

To increase the diversity of our data, we have contacted the Barcelona City Council, explaining the academic and non-commercial purpose of our work. We are currently waiting for their response and hope to access local waste management data that could complement the Melbourne dataset. This would allow us to validate the generalization capabilities of our approach on a different city context.

In the meantime, rather than immediately adding external factors such as weather or location metadata, we will first focus on getting the most out of the data we already have. Our next steps include experimenting with baseline models, Gradient Boosting and hybrid or ensemble methods to improve prediction accuracy and robustness. Once a solid baseline is established, we plan to integrate external variables to further refine and test our system.

5.3 Sustainability and Responsible AI

- **Environmental:** Reduce fuel consumption and CO_2 emissions by optimizing truck routes.
- **Social:** Improve waste management fairness between city areas.
- **Ethical:** Ensure transparent decision-making and explainable AI processes.

6 Project Management

6.1 Methodology

We follow an agile, iterative plan:

1. Weeks 1–2: Literature review & baseline replication
2. Weeks 3–4: Data exploration & cleaning
3. Weeks 5–6: Model development
4. Week 7: Model evaluation

5. Weeks 8–10: Final testing, documentation & report

6.2 Simplified Gantt Chart

Task	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
Literature Review & Baseline	X	X								
Data Exploration & Cleaning			X	X						
Model Development					X	X				
Model Evaluation							X			
Final Testing & Report								X	X	X

7 Initial Risks Identification

- **Limited Data:** The Melbourne dataset is small and potentially limits which technology we will be able to use.
- **Data Quality:** Sensor readings may be missing or inconsistent.
- **Generalization:** Ensuring model transferability across cities.

8 Conclusion

This first milestone defines a clear direction for our project: to build an AI-based system that makes urban waste collection smarter, more sustainable and adaptable to different environments. Our immediate goal is to replicate previous studies using the Melbourne dataset and improve prediction accuracy with approaches mentioned above. As we gather more data, including potential contributions from the Barcelona City Council, we will extend our system.

References

[1] João Ferreira, Nuno Lopes, Bernardo Ribeiro, Hugo Silva, and João Martins. Forecasting the status of municipal waste in smart bins. *Sensors*, 22(24):9779277, 2022.

[2] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[3] Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3):1–22, 2008.

[4] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[5] City of Melbourne Open Data. Argyle square smart bins. data.gov.au dataset listing; City of Melbourne Open Data portal, 2025. Accessed: 2025-10-05.