GREEN CITIES
GREEN LOGISTICS FOR GREENER CITIES

7th International Conference Green Cities 2026 – Green Logistics for Greener Cities, 15-17 April 2026, Szczecin, Poland

Predictive Models for Loading Zone Availability: A Case Study

Hamrah Kor[a,b,1], Lele Zhang[a,b], Mark Fackrell[a,b], Russell G. Thompson[c]

[a]*School of Mathematics and Statistics, The University of Melbourne, Parkville 3010, Australia.*
[b]*ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Australia.*
[c]*Department of Infrastructure Engineering, The University of Melbourne, Parkville 3010, Australia.*

Abstract

Efficient management of kerbside loading spaces in central business districts is critical for optimising urban freight operations and reducing traffic congestion and emissions. Providing information on loading bay availability and expected waiting times for vacant bays can inform couriers' delivery planning and reduce cruising time for parking. In this paper, we develop a prediction framework with a suite of models for predicting availability and waiting time, integrating non-homogeneous Poisson processes for modelling vehicle arrivals, Coxian phase-type distributions for modelling waiting time, and machine learning classifiers for predicting availability. These predictive models are designed to operate under both real-time and sensor-latency conditions, providing capabilities that can be integrated into IT-based freight distribution platforms. We validate the framework using in-ground sensor data from Melbourne's central business district, a dense commercial and tourist area, across both individual loading zones and loading zone clusters. Real-time evaluations demonstrate the framework's ability to accurately predict both bay availability probabilities and expected waiting durations, supporting data-driven decision-making tools for urban freight operations. These predictive models provide logistics operators and urban planners with tools to make better routing decisions and allocate kerb space more effectively, improving the efficiency of urban freight distribution and loading zone utilisation.
Keywords: Loading zone; last mile delivery, Coxian phase-type distribution, non-homogeneous Poisson process, machine learning, predictive model.

1 Introduction

Urban freight delivery systems face increasing challenges due to intensified competition for limited kerb space. Loading zones (LZs), designated areas for commercial vehicle loading and unloading, are essential infrastructure supporting last-mile delivery operations; however, their utilisation is highly variable and difficult to predict (Vuchic, 1999). This uncertainty causes unnecessary search time, delivery delays, illegal parking, and local congestion, resulting in significant economic and environmental costs (Dablanc, 2007). Research on loading zone management has evolved from survey-based analyses to real-time predictive modelling. Earlier studies applied statistical distributions such as negative binomial and Weibull to identify temporal patterns of truck parking demand (Kalahasthi et al., 2022) and conducted establishment-level surveys to examine relationships among land use, freight demand, and dwell time factors, including building attributes and vehicle types (Holguín-Veras et al., 2016; Kim et al., 2021). Recent studies have demonstrated that real-time information on kerb availability reduces cruising distance and time for courier services (Chiara et al., 2022). Machine learning (ML) models have become effective tools, with ensemble methods such as CatBoost, XGBoost, and Random Forest applied to analyse occupancy factors (Castrellon et al., 2023; Provoost & Kamilaris, 2020), and deep learning techniques, including long short-term memory (LSTM) and gated recurrent unit (GRU), employed to capture temporal dependencies in bay availability (Lyu et al., 2024; Martinez, Linares, and Casanovas, 2020). Nevertheless, most studies have examined arrival patterns or dwell times independently (Jain et al.,2024; Ogulenko et al., 2022), and none simultaneously produce bay-level availability probabilities alongside conditional expected waiting times derived from the same sensor data.

This study addresses the identified gap by developing a loading zone prediction (LZP) framework consisting of two models. The first model uses a machine learning classifier to predict bay availability at a given time. The

[1] Corresponding author. e-mail address: h.kor@student.unimelb.edu.au

second model estimates the expected waiting time when a bay is occupied, employing a Coxian phase-type distribution via the mean residual life (MRL) function. The framework incorporates a non-homogeneous Poisson process (NHPP) to capture time-varying arrival patterns and benchmarks the Coxian approach against multiple regression models from machine learning and deep learning. Because waiting times are not directly recorded by sensors, they are inferred from a tagged-driver simulation using observed arrival and departure records. Validation uses sensor data from Melbourne's central business district across individual and clustered bays, with evaluation conducted in both offline and real-time scenarios, including conditions with sensor data latency.

2 Loading Zone Usage Data in Melbourne

This study utilises publicly available on-street parking data from in-ground sensors in Melbourne's central business district (CBD) for the full calendar year from January to December 2019 (City of Melbourne, 2019), which record vehicle arrival and departure timestamps every two minutes. We define a "loading bay" as a single designated space and a "cluster" as a group of spatially clustered bays. Our analysis is restricted to the commercial operating hours of 07:30–18:30, excluding dwell times under 2 minutes or over 4 hours as outliers, while explicitly accounting for the sensor's two-minute registration latency in our methodology. As summarized in Table 1, the dwell times exhibit a strongly right-skewed distribution; the standard deviation exceeds the mean because a minority of exceptionally long durations significantly inflates the average despite the vast majority of stays being short. Finally, to examine trends in vehicle turnover, waiting times, and occupancy rates, we aggregated the data into 30- and 60-minute time bins.

Table 1. Summary statistics of the dataset.

Statistic	Value	Statistic	Value
No. of arrivals	1,246,898	Min duration	2.00 min
No. of bays	257	25th percentile	6.00 min
Mean	20.26 min	Median	13.00 min
Std. Dev.	27.88 min	75th percentile	28.00 min
Skewness	3.327	95th percentile	70.00 min
Kurtosis	14.456	Max duration	240.00 min

3 Methodology

This section explains the methods used to predict loading bay availability and driver waiting times. Availability refers to whether a single bay (or at least one bay in a cluster) is vacant upon arrival. Waiting time is the duration a driver must wait until a bay becomes available. A key data limitation is that the sensor records only vehicle arrival and departure timestamps, not the drivers' queueing times. To address this and generate ground-truth waiting times, we design a tagged driver simulation. We model vehicle inter-arrival patterns using an NHPP to capture time-varying arrival rates, while dwell times are modelled using a Coxian phase-type distribution fitted to observed sensor data. For each simulated arrival, the model checks the bay occupancy status; if all relevant bays are occupied, the waiting time is calculated as the time difference between the tagged driver's arrival and the next earliest departure event.

The LZP framework consists of models to predict bay availability and waiting times. We predict bay availability using statistical methods and ML classification models, specifically support vector classifier (SVC) and LightGBM (Ke et al., 2017). When a bay is predicted to be occupied, we estimate the expected waiting time using a statistical model based on Coxian phase-type distributions, benchmarked against ML and deep learning regression models. We evaluate the framework across both individual loading zones and cluster configurations under two scenarios:

- **S1 with historical data:** The driver plans a future visit with no access to live sensor data. Availability and expected waiting time are predicted solely from historical patterns.
- **S2 with real-time sensor latency:** While en route, the driver reads live sensor data at time t_1(the reading time) and plans to arrive at $t_2 > t_1$. The sensor carries a latency of $\delta \geq 0$ minutes, meaning the data read at t_1 reflects the true bay state at $t_1 - \delta$. This gives rise to two cases: case A (bay occupied), the sensor confirms occupation began at $t_0 < t_1 - \delta$. In this case, the confirmed elapsed occupation is $\tau = (t_1 - \delta) - t_0$; in case B (bay free), the bay was unoccupied at $t_1 - \delta$. The uncertainty window, the interval from the last state to the planned arrival, has duration $\Delta = t_2 - (t_1 - \delta)$. The special case $\delta = 0$ corresponds to zero sensor delay (that is, instantaneous data), giving $\tau = t_1 - t_0$ and $\Delta = t_2 - t_1$.

3.1 *Statistical models*

We model vehicle arrivals as an NHPP with a piecewise-constant intensity function to capture time-varying demand. Let t denote continuous time within the operating window (07:30–18:30) and let K denote the number of equal-length time bins partitioning this window. Within bin $k = 1,2,\dots,K$, the arrival rate $\lambda_k > 0$ (vehicles per minute) is assumed constant and estimated as the mean observed arrivals in that bin across all weekdays, that is,

$$\lambda(t) = \lambda_k, \quad t \in [t_{k-1}, t_k), k = 1,2,\dots,K. \tag{1}$$

Let $\Lambda(t) = \int_0^t \lambda(s)\, ds$ denote the cumulative mean of arrivals up to time t. The total arrival count satisfies $N(t) \sim \text{Poisson}(\Lambda(t))$, and for any sub-interval $[a, b]$, the number of arrivals $N(b) - N(a)$, follows a Poisson distribution with mean $\Lambda(b) - \Lambda(a)$.

The Coxian distribution (Cox, 1955; Neuts, 1975; Fackrell, 2009) models the absorption time of a continuous-time Markov chain, with one absorbing state, progressing linearly through $m > 0$ transient states. From state i, the chain transitions to the state $i + 1$ at the forward rate $\lambda_i > 0$ or is absorbed (exits) at a rate $\mu_i > 0$, reducing the size of the parameter space to $(2m - 1)$ compared to a general phase-type distribution. The distribution is fully specified by the initial probability vector $\boldsymbol{\alpha} = (1,0, \dots ,0)$ — meaning the chain always starts in state 1 — and the sub-generator matrix $\boldsymbol{Q}$, with entries $q_{i,i+1} = \lambda_i$ and $q_{i,m+1} = \mu_i$. The resulting probability density function is

$$f_S(t) = \boldsymbol{\alpha}\, e^{\boldsymbol{Q}t}\, \boldsymbol{q}, \qquad \boldsymbol{q} = -\boldsymbol{Q}\boldsymbol{1}, \tag{2}$$

where $\boldsymbol{q}$ is the exit rate vector and $\boldsymbol{1}$ is a column vector of ones. The number of phases m is selected by fitting models with increasing m and choosing the smallest value for which AIC and BIC cease to improve. We estimate parameters (λ_i, μ_i) using the expectation-maximisation (EM) algorithm (Dempster, Laird, Rubin, 1977), iterated until the log-likelihood improvement falls below 10^{-3}. The first two moments, used in subsequent calculations, are computed from the Coxian parameters as

$$\mathbb{E}[S] = -\boldsymbol{\alpha}\, \boldsymbol{Q}^{-1}\boldsymbol{1}, \qquad \mathbb{E}[S^2] = 2\, \boldsymbol{\alpha}\, \boldsymbol{Q}^{-2}\boldsymbol{1}. \tag{3}$$

We employ Coxian phase-type distributions for two distinct modelling purposes. Where $\alpha = (1,0, \dots ,0)$ is the initial probability vector and Q denotes the sub-generator matrix encoding transition and absorption rates:

1. **Dwell-time modelling:** The dwell-time random variable S_D — the duration any parked vehicle physically occupies the bay — is modelled as $S_D \sim \text{Coxian}(\boldsymbol{\alpha}, \boldsymbol{Q}_D)$. Where $\boldsymbol{Q}_D$ is the sub-generator matrix of the Coxian distribution fitted to observed parking durations from sensor data.
2. **Waiting-time modelling:** The waiting-time random variable S_W — the time a driver must wait when finding the bay occupied on arrival — is modelled as $S_W \sim \text{Coxian}(\boldsymbol{\alpha}, \boldsymbol{Q}_W)$, where $\boldsymbol{Q}_W$ is the sub-generator matrix fitted to synthetic waiting durations from the tagged driver simulation.

S1. Historical prediction

We now use time-dependent $M(t)/G/c$ queueing models to predict bay availability and waiting time at a future time t using historical data. Let c denotes the number of parallel loading bays and $\bar{D} = \mathbb{E}[S_D]$ the mean dwell time. For a cluster ($c > 1$), we assume homogeneous bays, meaning S_D is fitted to the pooled dwell times of all bays in the cluster. The time-varying occupancy rate, that is, the offered load per server, is

$$\rho(t) = min\left(1, \frac{\lambda(t) \cdot \bar{D}}{c}\right). \tag{4}$$

A single bay (C1) is modelled as an $M(t)/G/1$ queue, and then the probability that it is occupied is estimated by

$$P(occupied\ at\ t) \approx \rho(t). \tag{5}$$

Bay cluster (C2) is modelled as an $M(t)/G/c$ queue; the Erlang-C formula gives the probability that all c bays are simultaneously occupied, that is,

$$P(all\ occupied\ at\ t) \approx \frac{\frac{(c\,\rho)^c}{c!} \cdot \frac{1}{1-\rho}}{\sum_{k=0}^{c-1} \frac{(c\,\rho)^k}{k!} + \frac{(c\,\rho)^c}{c!} \cdot \frac{1}{1-\rho}}, \qquad \rho = \rho(t). \tag{6}$$

The inspection paradox (Ross, 2019) states that a driver arriving at a random point during an ongoing occupation is more likely to arrive during a long stay than a short one; thus, the expected residual stay exceeds the mean dwell time. Let, W_t denote the random waiting time experienced by a driver arriving at time *t*. The expected waiting time $\mathbb{E}[W_t]$ combines occupancy probability with the residual dwell time, that is,

$$\mathbb{E}[W_t] = \mathbb{P}(occupied\ at\ t) \times \frac{\mathbb{E}[S_D^2]}{2\, \mathbb{E}[S_D]}. \tag{7}$$

S2. Real-time prediction

The real-time scenario partitions into two cases depending on whether the sensor confirms the bay is occupied or free at time $t_1 - \delta$. We introduce the following notation to model the bay-state evolution over the uncertainty window $[t_1 - \delta, t_2]$. Let $F_{S_D}(\cdot)$ denote the cumulative distribution function (CDF) of the dwell-time random variable S_D, $\bar{F}_{S_D}(\cdot) = 1 - F_{S_D}(\cdot)$ its survival function, and $f_{S_D}(\cdot)$ its probability density function (PDF); each vehicle entering the bay draws an independent dwell time from S_D. Let $s \in [t_1 - \delta,\, t_2]$ denote the departure time of the vehicle currently occupying the bay at $t_1 - \delta$ (integration variable). Let $r \in [s,\, t_2]$ denote the arrival time of the next vehicle to enter after that departure (integration variable). Let u denote a generic dummy variable of integration within inner integrals. Let W denote the random waiting time experienced by the driver who is expecting to arrive at time t_2.

Case A: Sensor confirms bay is occupied at reading time

Assuming the bay has been occupied since $t_0 < t_1 - \delta$, the confirmed elapsed occupation is $\tau = (t_1 - \delta) - t_0$. The following mutually exclusive probabilities partition the event space:

Scenario 1 — Bay is continuously occupied until arrival.

The vehicle currently occupying the bay at $t_1 - \delta$ remains present throughout $[t_1 - \delta,\, t_2]$, the probability is

$$P_1 = \frac{\bar{F}_{S_D}(\tau + \Delta)}{\bar{F}_{S_D}(\tau)}. \tag{8}$$

Scenario 2 — Current vehicle departs; no subsequent arrival before t_2.

The current occupying vehicle departs at some time $s \in [t_1 - \delta,\, t_2]$ and no new vehicle arrives in $[s,\, t_2]$, so the bay is free at t_2. The factor $\exp\left(-\int_s^{t_2} \lambda\,(u)\,du\right)$ is the probability of zero NHPP arrivals in $[s,\, t_2]$, therefore,

$$\mathrm{P}_2 = \int_{t_1-\delta}^{t_2} \frac{f_{S_D}(s - t_0)}{\bar{F}_{S_D}(\tau)} \cdot \exp\left(-\int_s^{t_2} \lambda\,(u)\,du\right) ds. \tag{9}$$

Scenario 3 — Current vehicle departs; exactly one subsequent arrival occupies bay at t_2.

The current occupying vehicle departs at $s \in [t_1 - \delta,\, t_2]$, the next vehicle arrives at $r \in [s,\, t_2]$ with instantaneous rate $\lambda(r)$, with no other arrival in $[s,\, r]$; and this next vehicle has not yet departed at t_2 (with probability $\bar{F}_{S_D}(t_2 - r)$), yielding

$$\mathrm{P}_3 = \int_{t_1-\delta}^{t_2} \frac{f_{S_D}(s - t_0)}{\bar{F}_{S_D}(\tau)} \int_s^{t_2} \lambda\,(r) \exp\left(-\int_s^{r} \lambda\,(u)\,du\right) \bar{F}_{S_D}(t_2 - r)\,dr\,ds. \tag{10}$$

Scenario 4 — Current vehicle departs; next arriving vehicle also departs before t_2, bay free.

As Scenario 3, but the next arriving vehicle also departs before t_2 (probability $F_{S_D}(t_2 - r)$) and no further vehicle arrives in $[r,\, t_2]$, that is,

$$P_4 = \int_{t_1-\delta}^{t_2} \frac{f_{S_D}(s - t_0)}{\bar{F}_{S_D}(\tau)} \times \int_s^{t_2} \lambda\,(r) \exp\left(-\int_s^{r} \lambda\,(u)\,du\right) F_{S_D}(t_2 - r) \exp\left(-\int_r^{t_2} \lambda\,(u)\,du\right) dr\,ds. \tag{11}$$

Scenario 5 — Multiple turnovers; bay occupied at t_2.

The complement of Scenarios 1–4: the currently occupying vehicle departs, and two or more subsequent vehicles arrive before t_2, with the bay occupied at t_2, that is

$$P_5 = 1 - (P_1 + P_2 + P_3 + P_4). \tag{12}$$

The probability that the bay is occupied at the driver's arrival time t_2, and its complement that it is free, are

$$P_{occupied} = P_1 + P_3 + P_5, \tag{13}$$

$$P_{free} = P_2 + P_4. \tag{14}$$

The expected waiting time $\mathbb{E}[W]$ is a scenario-weighted sum of residual dwell times. Let $MRL_{S_D}(a)$ denote the mean residual life of S_D at elapsed occupancy time $a \geq 0$, that is, the expected remaining dwell time given the vehicle has already been present for a minutes,

$$MRL_{S_D}(a) = \frac{\int_a^\infty \bar{F}_{S_D}(x)\,dx}{\bar{F}_{S_D}(a)}. \tag{15}$$

Under Scenario 1, the driver waits for the residual dwell time of the current occupying vehicle, which has been present for $\tau + \Delta$ minutes by time t_2. Under Scenario 3, the driver waits for the residual dwell time of the next arriving vehicle, weighted by its possible arrival time r. Under Scenario 5, the expected waiting time is approximated by $\mathbb{E}[S_W]$, the mean of the fitted waiting-time distribution. The expected waiting time is then

$$\mathbb{E}[W|occupied] = \frac{P_1}{P_{occupied}} \cdot MRL_{S_D}(\tau + \Delta) + \frac{P_3}{P_{occupied}} \cdot \mathbb{E}_r\left[MRL_{S_D}(t_2 - r) \mid \mathcal{E}_3\right] + \frac{P_5}{P_{occupied}} \cdot \mathbb{E}[S_W], \tag{16}$$

where $\mathcal{E}_3$ denotes the event corresponding to Scenario 3, the P_3-weighted conditional expectation of the residual dwell time of the next arriving vehicle over its arrival time r is

$$\mathbb{E}_r\left[MRL_{S_D}(t_2 - r) \mid \mathcal{E}_3\right] = \frac{1}{P_3}\int_{t_1-\delta}^{t_2} \frac{f_{S_D}(s - t_0)}{\bar{F}_{S_D}(\tau)} \times \int_s^{t_2} \lambda(r)\, exp\left(-\int_s^r \lambda(u)\, du\right) \bar{F}_{S_D}(t_2 - r)\, MRL_{S_D}(t_2 - r)\, dr\, ds. \tag{17}$$

Case B: Sensor confirms bay is free at reading time

If the sensor data at t_1 confirms the bay was free at $t_1 - \delta$, there is no current occupying vehicle to track. The uncertainty window begins with an empty bay, and occupancy at t_2 depends entirely on new arrivals. We define four exhaustive and mutually exclusive scenarios for the free-state case.

Scenario F1 –The probability that no vehicles arrive during the uncertainty window $[t_1 - \delta, t_2]$,

$$Q_1 = exp\left(-\int_s^r \lambda(u)\, du\right). \tag{18}$$

Scenario F2 –The probability that exactly one vehicle arrives at $r \in [t_1 - \delta, t_2]$ and its dwell time exceeds $t_2 - r$,

$$Q_2 = \int_{t_1-\delta}^{t_2} \lambda(r)\ exp\left(-\int_{t_1-\delta}^{r} \lambda(u)\ du\right) \bar{F}_{S_D}(t_2 - r)\, dr. \tag{19}$$

Scenario F3 – The probability that exactly one vehicle arrives at $r \in [t_1 - \delta, t_2]$, departs before t_2, and no further vehicle arrives,

$$Q_3 = \int_{t_1-\delta}^{t_2} \lambda(r)\, exp\left(-\int_{t_1-\delta}^{r} \lambda(u)\ du\right) F_{S_D}(t_2 - r)\ exp\left(-\int_r^{t_2} \lambda(u)\ du\right) dr. \tag{20}$$

Scenario F4 – Multiple turnovers, bay occupied at t_2. The complement of scenarios F1-F3,

$$Q_4 = 1 - (Q_1 + Q_2 + Q_3). \tag{21}$$

Following the notation established previously, the probability of the bay being occupied at the driver's arrival time t_2, given it was free at $t_1 - \delta$, is

$$P_{\text{occupied|free}} = Q_2 + Q_4. \tag{22}$$

The expected waiting time, conditioned on the bay being free at reading time, combines the residual life of the arrival in $F2$ with the mean synthetic waiting time $\mathbb{E}[S_W]$ for F4 is

$$\mathbb{E}[\,W \mid \text{free}\,] = \frac{Q_2}{P_{\text{occupied|free}}} \cdot \mathbb{E}_r\left[\text{MRL}_{S_D}(t_2 - r) \mid \mathcal{E}_{F2}\right] + \frac{Q_4}{P_{\text{occupied|free}}} \cdot \mathbb{E}[S_W]. \tag{23}$$

3.2 Machine learning models

We applied ML and deep learning models to estimate bay availability and waiting times. For immediate bay availability, LightGBM and SVC produced availability probabilities. For the waiting time, we used regression models including Random Forest, Extra Trees, XGBoost, LightGBM, two deep neural networks (PyTorch and TensorFlow), and a Stacking Ensemble that combines base-learner predictions (Géron, 2022; Awan et al., 2020). All models utilise a shared feature set including sine-cosine encodings of hour-of-day h and day-of-week d $(\sin(2\pi h/24), \cos(2\pi h/24))$ and $(\sin(2\pi d/7), \cos(2\pi d/7))$ to capture cyclic time structure, dwell-time statistics (mean, median, standard deviation per bay), binary peak-hour indicators, and historical averages of occupancy and waiting times. Regression models were trained by minimising the mean squared error (MSE).

3.3 Evaluation metrics

We assessed the goodness-of-fit of the NHPP model using the KS test applied to the inter-arrival time distributions within each time bin. The fitted Coxian dwell-time and waiting-time distributions were evaluated using four complementary statistics: the Akaike information criterion (AIC) and Bayesian information criterion (BIC), which balance goodness of fit against model complexity and penalise over-parameterisation; the KS p-value, testing whether the fitted CDF deviates significantly from the empirical distribution; and the Anderson–Darling (AD) statistic, which places greater weight on discrepancies in the tails of the distribution.

3.4 ML model validation.

We validated the models separately for the two prediction tasks. Specifically, we evaluated the classification models (bay availability) using the area under the ROC curve (AUC). Aggregating true versus false positive rates across all probability thresholds, AUC measures the model's ability to discriminate between available and occupied bays. We evaluated waiting time models using R^2 and mean absolute error (MAE). R^2 measures how well the model explains the observed variance in the data, ranging from 0 (no better than a mean baseline) to 1 (perfect fit). MAE captures the average size of prediction errors in the original time units, giving an interpretable sense of average deviations.

4 Numerical Analysis

This section presents numerical results for both a single loading zone and a cluster of loading zones under historical and real-time scenarios. We selected six representative loading zones in the Melbourne CBD, with bay IDs 735, 1175, 1496, 2622, 6075, and 6263, for the single-zone analysis. Also, we considered five adjacent loading bays with IDs 6222, 6223, 6229, 6230, and 6231 in the Docklands area as a cluster of loading zones.

4.1 Single loading zone analysis

The NHPP model adequately captured inter-arrival dynamics across all bays, with KS p-values exceeding 0.05 (see Table 2). Mean waiting times, generated via the tagged driver simulation, ranged from 15.62 to 27.90 minutes, with availability from 5.10% to 25.05%. A five-phase Coxian model provided the best fit for waiting-time distributions across all bays, balancing goodness of fit and complexity, as shown in Figure 1 and Table 3.

Table 2: Waiting time statistics and NHPP validation results.

Bay ID	735	1175	1496	2622	6075	6263
Total arrivals	3908	1536	6459	7568	8091	8091
Average waiting time	27.90	25.66	23.06	15.62	23.04	22.21
Median waiting time	16.41	12.99	9.13	5.81	12.56	12.47
Maximum waiting time	232.82	173.47	235.78	232.84	231.89	227.27
Availability percentage	10.26	16.80	20.16	25.05	5.10	6.82
KS P-Value	0.1695	0.2138	0.4354	0.1694	0.0805	0.094

Table 3. Statistics of phase-type models for fitting waiting times across different loading bays.

Bay ID	AIC	BIC	KS P-Value	AD Statistic	N Observations	R^2
735	31103.92	31159.38	0.43	0.23	3507	0.9968
1175	11295.10	11341.48	0.39	0.19	1278	0.9902
1496	44432.59	44491.52	0.44	0.36	5157	0.9977
2622	44960.99	45020.78	0.46	0.24	5672	0.9975
6075	63948.52	64011.03	0.43	0.29	7678	0.9982
6263	62528.79	62591.15	0.43	0.16	7539	0.9990

After fitting the NHPP and Coxian phase-type distributions to historical data, we evaluated the model in individual loading zones. Table 4 shows that, in the absence of live sensor data (Scenario S1), the statistical model outperforms LightGBM and SVM in bay availability prediction (AUC: 0.825–0.986) than LightGBM (0.519–0.982) and SVM (0.468–0.964), particularly for heavily-demanded bays such as bay IDs 6075 and 6263. For waiting-time estimation, all models performed similarly in terms of MAE. This MAE indicates that, on average, the models predicted waiting times deviate from the actual waiting times by roughly 20 to 25 minutes.

Table 4. Evaluation results for bay availability and waiting time predictions.

Bay IDs	LZ availability prediction (AUC) Statistical Model	LightGBM classifier	SVM	Waiting time prediction (MAE) Statistical Model	Random Forest	Extra Trees	XGBoost	PyTorch NN	TensorFlow NN	Stacking Ensemble
735	0.8248	0.7451	0.7234	24.16	23.74	24.33	23.58	24.62	22.26	22.20
1175	0.9864	0.9824	0.9638	20.43	23.55	21.80	23.67	25.87	22.81	22.31
1496	0.8444	0.6291	0.5612	25.55	28.29	28.29	27.30	26.68	26.60	26.50
2622	0.8278	0.5927	0.5250	22.08	20.42	20.36	19.93	20.38	19.42	20.44
6075	0.9580	0.5588	0.4912	23.11	21.07	21.24	20.70	20.13	20.38	20.69
6263	0.8978	0.5190	0.4684	21.42	19.81	19.83	19.41	18.98	18.85	19.66

In the second scenario S2 with live data, we examined the models for three different values of δ (sensor latency): 0, 2, and 5 minutes. Table 5 shows that access to real-time sensor data substantially improves the prediction performance across both tasks relative to the historical baseline for bay ID 735. At zero sensor latency ($\delta = 0$), the statistical model achieves an AUC of 0.944 for availability prediction, a great improvement over the historical AUC of 0.825. However, as sensor latency increases from $\delta = 0$ to $\delta = 5$ minutes, the statistical model's availability AUC degrades smoothly (from 0.9435 to 0.9075), whereas the ML models, especially LightGBM, maintain stable performance (AUC $\approx$ 0.913–0.936). This divergence reflects the statistical model's direct dependence on the confirmed elapsed occupancy time τ, which becomes unreliable at higher latencies. Similarly, for waiting time prediction, access to real-time data improves accuracy relative to the historical baseline. For example, at zero latency, the MAE for bay 735 drops from over 24 minutes (historical) to between 18.2 and 20.7 minutes (real-time). However, both models show moderate degradation as latency increases, with MAE rising by approximately 1.3 minutes from $\delta = 0$ to $\delta = 5$, suggesting that even moderate sensor delays meaningfully erode precision.

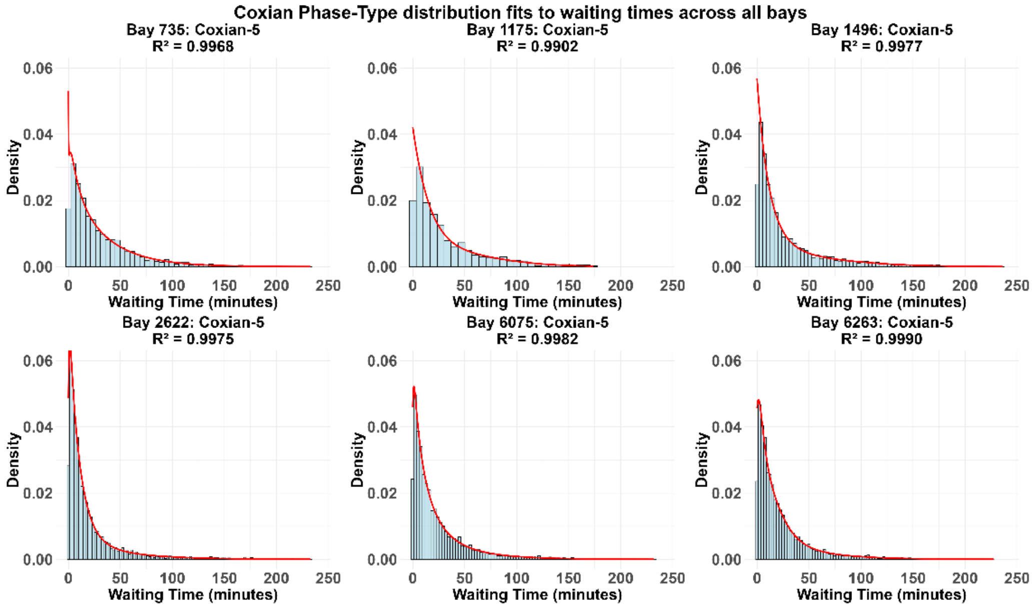

Fig. 1. Waiting time empirical histograms and fitted distributions for six selected loading bays.

Table 5. Accuracy of bay availability and waiting time estimation in a real-time scenario with different latencies for bay ID 735

	LZ availability prediction (AUC)			Waiting time prediction (MAE)						
δ (min)	Statistical Model	LightGBM classifier	SVM	Statistical Model	Random Forest	Extra Trees	XGBoost	PyTorch NN	TensorFlow NN	Stacking Ensemble
0	0.9435	0.9125	0.8857	23.42	18.94	18.28	19.06	22.87	19.63	18.24
2	0.9207	0.9363	0.8935	23.41	19.43	18.90	19.69	23.35	20.06	18.87
5	0.9075	0.9203	0.8963	23.41	19.76	19.36	20.14	23.62	20.30	19.36

4.2 Loading cluster analysis

To ensure a fair comparison of space resources, we analysed the efficiency of pooling bays into a unified cluster versus operating them as scattered, independent zones facing proportional demand. According to the resource pooling principle in multi-server ($M/G/c$) queueing theory, a unified cluster of c bays achieve a strictly lower probability of all bays being simultaneously occupied than c separate, independent single-bay queues under the same total load. Table 6 confirms this effect in practice, by grouping adjacent bays into a pooled loading cluster, the mean probability of finding an available bay significantly improved from 15.0% (the average for an independent single bay) to 32.2% (the unified cluster). To evaluate waiting times when the cluster is fully occupied, Coxian phase-type modelling confirmed that a five-phase structure provided the best fit for the non-zero waiting times (

Table 7). Finally, comparing the cluster's waiting-time errors (Table 8) with those of individual bays under both historical and real-time conditions (Table 4 and Table 5) reveals a considerable improvement in predictive accuracy. Single-bay MAEs remain high (18–28 minutes) across all scenarios, whereas cluster MAEs drop significantly to 3.4–4.9 minutes. Although sensor delays degrade the estimation power of both models, the statistical model predicts waiting times more accurately, whereas the ML model performs better at predicting availability.

Table 6. Summary of performance metrics for a loading cluster.

A=Prob. Availability (5-bay cluster)	B=Prob. Availability (avg single bay)	Absolute increase Δ=A-B)	Mean gain ratio (A / B)
0.322	0.150	+0.172	2.23

Table 7. Coxian Phase-Type Distribution Fit Results for Non-Zero Waiting Times.

Phases	AIC	BIC	KS p-value	AD Statistic
3	100,156.2	100195.2	0.5758	0.6085
5	100,135.3	100205.6	0.9792	0.1909
10	100,154.4	100302.6	0.9232	0.1754

Table 8. Accuracy of bay availability and waiting time prediction in a real-time scenario with different latencies on a cluster of LZs.

	LZ availability prediction (AUC)			Waiting time prediction (MAE)						
Scenario	statistical model	LightGBM classifier	SVM	Statistical Model	Random Forest	Extra Trees	XGBoost	PyTorch NN	TensorFlow NN	Stacking Ensemble
Historical	0.6386	0.6254	0.6347	3.905	4.9102	4.9487	4.6765	4.5811	4.5439	4.6278
δ=0	0.8570	0.9024	0.8889	3.480	4.3428	4.3040	4.4294	4.5257	4.3621	4.2646
δ=2	0.8440	0.8974	0.8832	3.511	4.3623	4.3414	4.4759	4.5208	4.4725	4.2845
δ=5	0.8218	0.8920	0.8774	3.555	4.3704	4.3513	4.4743	4.5248	4.3843	4.3070

5 Conclusion

This research presents a prediction framework that combines NHPP, Coxian phase-type distributions, and ML methods to predict bay availability and waiting times under different conditions. The statistical approach integrates NHPP with Coxian phase-type distributions, driven by continuous-time Markov chains and matrix analytic methods. Specifically, for real-time predictions under sensor latency, this statistical framework utilises a state-dependent probability structure of a five-scenario model if the bay is initially occupied, and a four-scenario model if the bay is initially free. In addition, ML models were applied using a range of regression and classification algorithms. Results indicate that while both approaches are effective, the ML methods achieved higher accuracy in estimating waiting times and proved more robust to sensor latency when predicting real-time bay availability.

Future work could focus on incorporating these predictive insights into real-time routing planning to optimise last-mile delivery and improve urban logistics efficiency. Also, further research is needed to address limitations in sensor accuracy and potential geographic mismatches between suggested and destination-proximate bays.

References

Awan, F. M., Saleem, Y., Minerva, R., & Crespi, N. (2020). A Comparative Analysis of Machine/Deep Learning Models for Parking Space Availability Prediction. *Sensors*, 20(1), 322.

Castrellon, J. P., Sánchez-Díaz, I. D., & Kalahasthi, L. K. (2023). Enabling factors and durations data analytics for dynamic freight parking limits. *Transportation Research Record*, 2677, 219–234.

Chiara, G. D., Krutein, K. F., Ranjbari, A., & Goodchild, A. V. (2022). Providing curb availability information to delivery drivers reduces cruising for parking. *Scientific Reports*, 12(1), 19355.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51, 313–319.

Dablanc, L. (2007). Goods transport in large European cities: Difficult to organize, difficult to modernize. *Transportation Research Part A: Policy and Practice*, 41, 280–285.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological),* 39(1), 1–22.

Fackrell, M. (2009). Modelling healthcare systems with phase-type distributions. *Health Care Management Science*, 12, 11–26.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media. ISBN: 978-1-098-12597-4.

Holguín-Veras, J., Lawson, C. T., Wang, C., Jaller, M., González-Calderón, C. A., Campbell, S., Kalahasthi, L., Wojtowicz, J., & Ramírez-Ríos, D. (Eds.). (2016). Using commodity flow survey microdata and other establishment data to estimate the generation of freight, freight trips, and service trips: Guidebook. The National Academies Press.

Jain, M., Amatya, V.C., Bleeker, A., Vasisht, S., Feo, J.T., & Wolf, K.E. (2024). Predicting curb side parking availability for commercial vehicle loading zones. *International Journal of Intelligent Transportation Systems Research, 22*, 614 - 628.

Kalahasthi, L. K., Sánchez-Díaz, I. D., Castrellon, J. P., Gil, J., Browne, M., Hayes, S., & Sentís Ros, C. (2022). Joint modeling of arrivals and parking durations for freight loading zones: Potential applications to improving urban logistics. *Transportation Research Part A: Policy and Practice*, 166, 307–329.

Kim, H., Goodchild, A. V., & Boyle, L. N. (2021). Empirical analysis of commercial vehicle dwell times around freight-attracting urban buildings in downtown Seattle. *Transportation Research Part A: Policy and Practice*, 147, 320–338.

Lyu, M., Ji, Y., Kuai, C., & Zhang, S. (2024). Short-term prediction of on-street parking occupancy using multivariate variable based on deep learning. *Journal of Traffic and Transportation Engineering (English Edition)*, 11, 28–40.

Martínez, J. A., Linares, M. P., & Casanovas, J. (2020). Characterizing parking systems from sensor data through a data-driven approach. *Transportation Letters*, 13(3), 183–192.

City of Melbourne. (2019). On-street car parking sensor data – 2019, https://data.melbourne.vic.gov.au/explore/dataset/on-street-parking-bay-sensors/information/.

Neuts, M. F. (1975). Computational uses of the method of phases in the theory of queues. *Computers & Mathematics with Applications*, 1, 151–166.

Ogulenko, A., Benenson, I., & Fulman, N. (2022). The nature of the on-street parking search. *Transportation Research Part B: Methodological.* 166, 48-68.

Provoost, J. C., Kamilaris, A., Wismans, L. J. J., van der Drift, S., & van Keulen, M. (2020). Predicting parking occupancy via machine learning in the web of things. *Internet of Things*, 12, 100301.

Ross, S. M. (2019). Introduction to Probability Models (12th ed.). Academic Press.

Vuchic, V. R. (1999). Transportation for Livable Cities. Routledge.

Zhang, L., & Thompson, R. G. (2019). Understanding the benefits and limitations of occupancy information systems for couriers. *Transportation Research Part C: Emerging Technologies*, 105, 520–535.