Transport for NSW

Urban Freight Forecasting Model

Technical Guide

September 2023
Version 2.0

transport.nsw.gov.au

NSW GOVERNMENT

Table of Contents

September 2023

What's new in this version.

The version includes:

- Hotels as a building land use
- More survey data incorporated in back-end regression analysis that improves the results particularly for larger buildings

Summary

The purpose of this document is to provide technical guidance for the Urban Freight Forecasting Model (UFFM) so that users may understand the data and methodology behind its outputs.

The model aims to support decision-making processes through better estimates of the levels and types of freight & servicing activity generated by buildings, as well as the performance of their dock provisions.

At its simplest, the model applies regression analysis to a consolidated dataset of freight survey activity. The coefficients produced by this are combined with the building information entered by the user and fed into a simulation. The simulation provides the model with a statistical distribution of outcomes where averages are taken from and presented in various dashboards for the user.

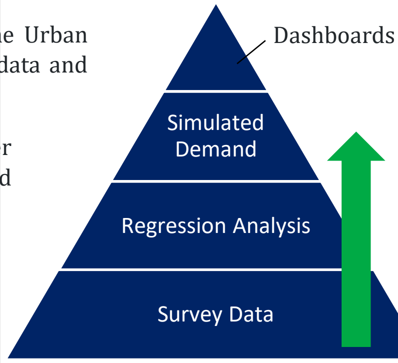

Figure 1: Model Structure

The project was undertaken in close collaboration with researchers from University of Melbourne who led the development of the model and its methodology. A research paper about the model and methodology was submitted to the Australasian Transport Research Forum (Aljohani et al. Predicting freight demand for planning loading docks).

Aljohani et al
Predicting freight der

Model Process Map

The below diagram shows how the model works at a very high level.

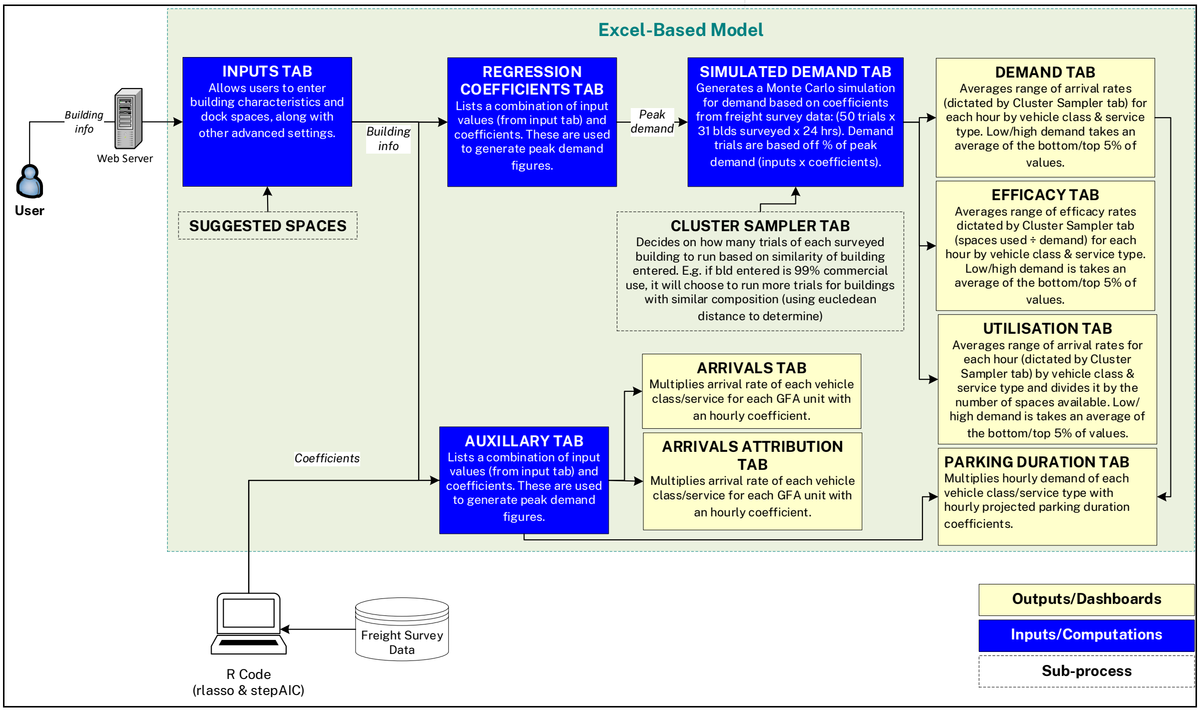

Figure 2: High level summary of model

Inputs

The 'Inputs' tab acts as the interface between user and model and can be split into two main sections with the first section focusing on the building information, and the second focusing on dock performance.

Building Information

These inputs are key predictors of freight and ultimately determine the demand and arrival rates generated by the building. They combine with the regression coefficients to generate levels of peak demand which feed into the 'Simulated Demand' tab as well as coefficients that forecast the total volume of freight across the day.

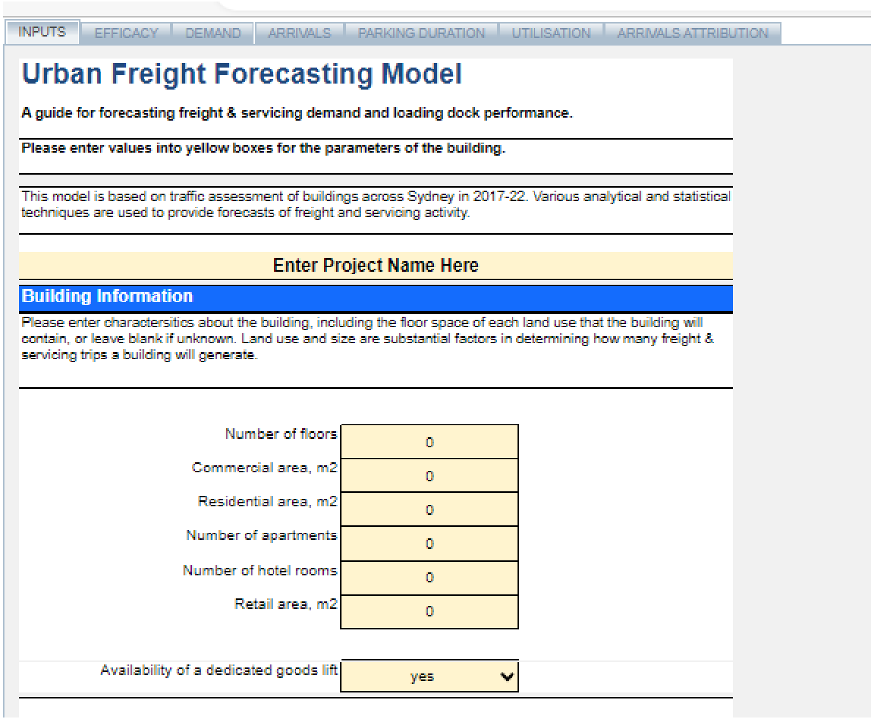

Figure 3: Building Information Inputs

- Number of floors and availability of dedicated goods lift affect service times
- Commercial, residential, hotel rooms and retail floor space are combined with regression coefficients and are also used by the model to compare the building with those surveyed for relevancy.

Dock Information

This section allows users to test different combinations of loading dock spaces while also suggesting an optimal starting point.

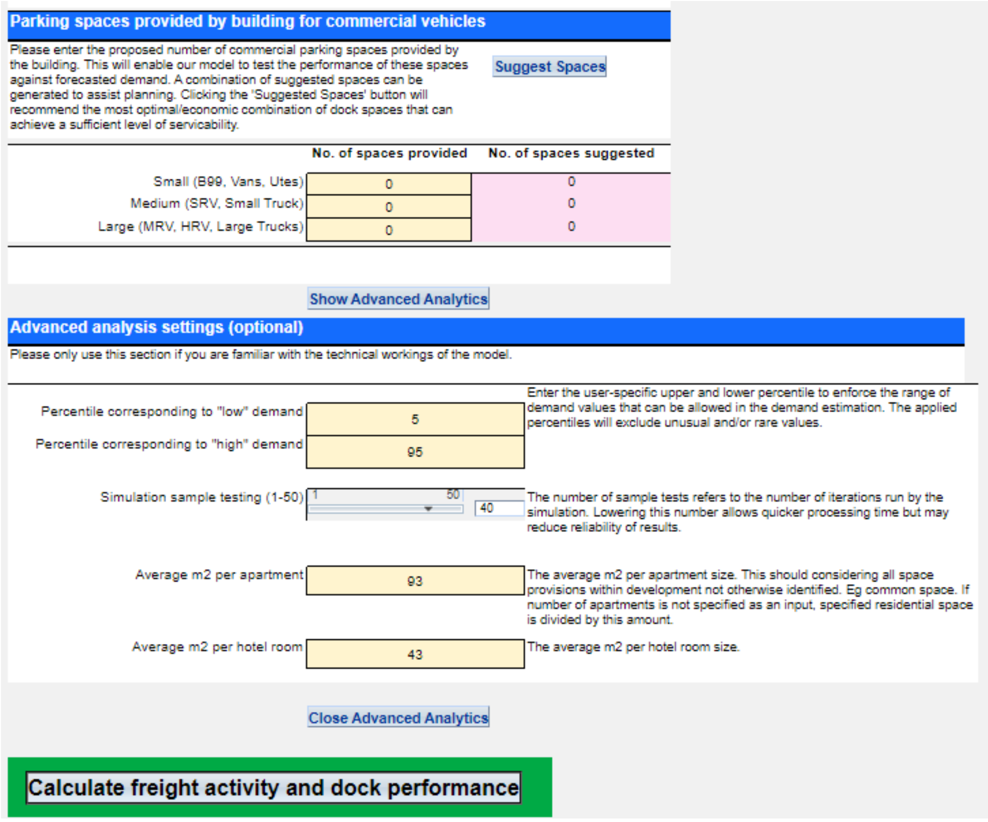

Figure 4: Dock inputs

- The number of dock spaces entered is used by the ‘Efficacy’ tab to assess the percentage of vehicles that were able to be accepted by the dock and determine how many of each vehicle class will be rejected back onto the network.
- The advanced settings allow the user to manually override what the model deems “low” and “high” scenarios in the ‘demand table’ under the ‘Demand’ tab. By default, the model considers the top and bottom 5% of iterations/outcomes as high/low scenarios. A lower number will reduce processing time but may affect the reliability of results.
- The average metres per apartment combines with an entry of residential area to determine the number of apartments.
- The average room per hotel room is multiplied with the number of rooms entered to calculate the size of a hotel.

Suggested Spaces Sub-Process

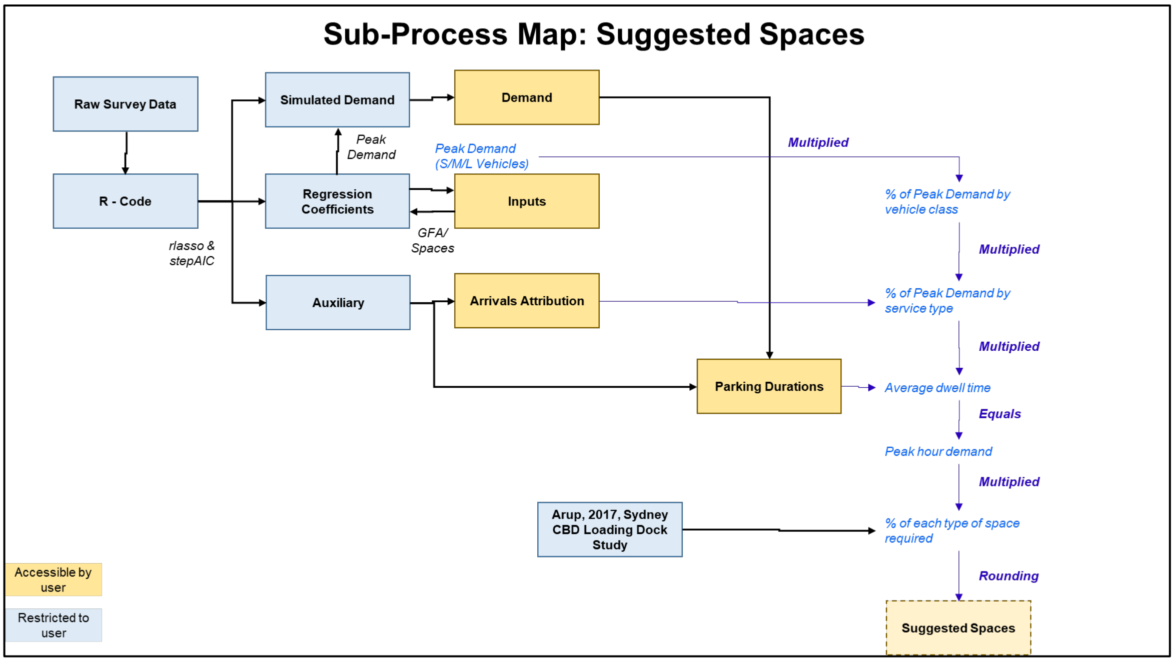

Figure 5: Suggested spaces sub-process

Data Sources

A comprehensive data collection effort was conducted to survey 44 buildings of varying sizes, locations, land use sizes and types: the first in November 2017, and the second in June 2022.

Surveys of each building were conducted on three consecutive weekdays (Tuesday to Thursday) to capture:

- Freight & servicing activity
- Parking in the dock
- Parking and walking in from nearby kerbside spaces

A total of approximately 9000 movements were captured by a combination of video surveillance and manual staff where obstructions or multi-directional arrivals were initially observed.

Four different land uses are included in the model being commercial, residential, retail and hotel. While the primary land use of many buildings is commercial, many of these contained retail spaces of up to 3,000m^2 which have been captured by the model as shown in Table 1.

Figure 6 shows an overview of freight and servicing activity on a typical weekday for the four different land uses. The graphs show the average number of freight movements by the hour on a typical weekday for each land use, highlighting their peak deliveries (i.e. the highest average number of movements generated in the hour).

Building id	Primary Use	Total Area	Location	Survey date
1	Commercial	39,539	Sydney CBD	10/2020
2	Residential	14,200	North Sydney	11/2020
3	Residential	21,550	North Sydney	11/2020
4	Residential	15,350	Sydney CBD	11/2020
5	Residential	41,500	Sydney CBD	11/2020
6	Residential	27,165	Green Square	11/2020
7	Commercial	25,000	Parramatta	12/2020
8	Commercial	22,000	Parramatta	12/2020
9	Commercial	7,651	Parramatta	12/2020
10	Commercial	10,940	Parramatta	12/2020
11	Commercial	28,500	North Sydney	12/2020
12	Commercial	18,000	Parramatta	12/2020
13	Commercial	22,262	North Sydney	12/2020
14	Commercial	40,695	Sydney CBD	12/2020
15	Commercial	42,965	Sydney CBD	12/2020
16	Commercial	17,758	Sydney CBD	12/2020
17	Commercial	62,070	Sydney CBD	12/2020
18	Commercial	46,160	Sydney CBD	6/2021
19	Commercial	6,020	Parramatta	6/2021
20	Commercial	21,996	North Sydney	6/2021
21	Residential	21,500	Parramatta	6/2021
22	Residential	17,900	Macquarie Park	6/2021
23	Commercial	45,455	North Sydney	6/2021
24	Commercial	39,865	Sydney CBD	6/2021
25	Commercial	39,853	Sydney CBD	6/2021
26	Residential	20,000	Sydney CBD	6/2021
27	Commercial	24,739	Sydney CBD	6/2021
28	Residential	24,885	Parramatta	6/2021
29	Commercial	26,480	Parramatta	6/2021
30	Hotel	15,064	North Sydney	6/2021
31	Commercial	12,325	North Sydney	6/2021
32	Hotel	20,900	Sydney CBD	11/2018
33	Hotel	11,835	Parramatta	11/2018
34	Hotel	17,190	Sydney CBD	11/2018
35	Commercial	37,473	North Sydney	8/2022
36	Commercial	52,608	North Sydney	6/2022
37	Commercial	19,128	North Sydney	8/2022
38	Commercial	4,930	North Sydney	8/2022
39	Commercial	13,738	North Sydney	8/2022
40	Retail	9,245	North Sydney	8/2022
41	Hotel	32,600	Sydney CBD	11/2017
42	Residential	25,000	Sydney CBD	11/2017
43	Residential	27,000	Sydney CBD	11/2017
44	Residential	27,500	Sydney CBD	11/2017

Table 1 Buildings survey

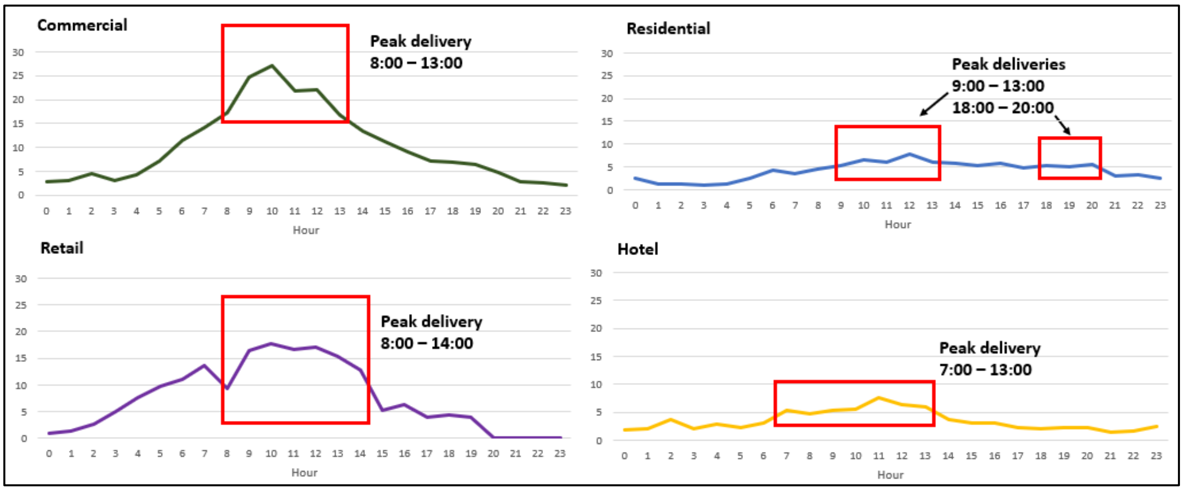

Figure 6: Average number of freight movements by hour on a typical weekday

Survey Template

The following template was used to convert video footage of freight and servicing events into rows of records for analysis in Excel. Each event/row represents an arrival and departure of a freight & servicing activity to a building.

Seq	Site	Event Date & Week day	Park Location	Cam Location	Vehicle ID	Arrival Time	Queue Length	Entry Time	Dwell Time	Vehicle Type	Visit Type	Activity Type	Exit Time	Remarks
1	Site 1	15/06/2021 Tuesday	Outside Bld X	Outside of Loading dock	1	5:51:12		5:51:12	6:46:37	Ute	Trade & Service	Service	12:37:49	White colour ute with no logo

Table 2: Freight assessment template

Vehicle, visit and activity type fields contained in the following data validation lists were used to ensure consistent categorisation during the input process. A visual guide was provided to staff to assist with the categorisation of vehicles and activities.

Vehicle type	Visit type	Activity type
Ute	Trade & Service	Service
SUV	Removalist	Parking
Car	Private parking	Waste
SRV "Small Rigid Vehicle"	Waste collection	Delivery/Pickup
Pedestrian	Courier	Passenger Services
Van	Other Freight	Unidentifiable
Motorbike	Laundry Service	Construction
Bicycle	Food	
MRV "Medium Rigid Vehicle"	Bulky Delivery	
Taxi	Passenger Services	
Station Wagon	Beverages	
	Unknown	
	Construction	

Table 3: Freight activity & vehicle classifications

The screenshot below shows a sample of video footage recorded by the freight surveys. Many sites have multiple points of access for freight & servicing like dock and lobby entrances – it is important to consider the complexity and added costs these require to monitor when selecting sites to survey.

In dense urban areas where on-site dock parking and nearby kerbside parking is limited, additional cameras are often required to capture the arrival characteristics of events occurring outside of the nearby proximity. In the screenshot below, freight & servicing activity could have come from outside of camera view (blue lines), so more cameras were needed to capture along the street. When choosing sites like this, it is often more cost effective to measure a neighbour building which shares street parking.

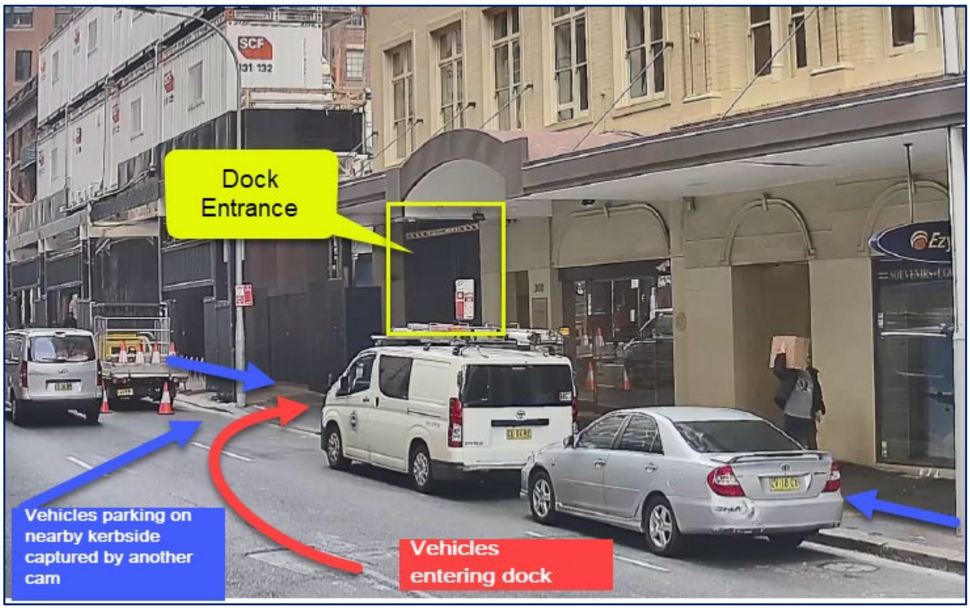

Figure 7: Freight assessment video screenshot

Future Data Expansion

By having a consistent input template for capturing freight & servicing events, the same process can be reliably duplicated and more easily fed into the regression analysis. The Urban Freight team is open to accepting submissions of data collected externally to incorporate into the model provided they are:

Robust

Enough sites and days should be captured to give a meaningful picture of freight demand for each building. The methodology/camera set-up would need to include *all freight & servicing activities* that each building generates. This includes drivers entering the dock and those parking/walking from nearby streets and loading zones.

Statistically distributed and relevant

The number and types of sites should be diverse enough to provide a significant distribution of data points for the regression model to work effectively. This is especially important for the land use, size and locations of buildings where high correlations were found for freight demand.

Consistent

For future datasets to be accepted by the regression analysis, the data needs to be in a format consistent with previous entries and definitions. (e.g., using the same weekdays for survey timelines and having the same definitions of what constitutes a bulky delivery etc.)

Have high quality assurance

Previous data collection methods involved manual video processing to convert footage of events to Excel datasets. Considerable time was spent reviewing and cleansing this data to ensure that it was reliable and accurate. While emerging technology like AI and machine learning are proving to be promising, a level of assurance needs to be applied to ensure the data reported accurately reflects the events being recorded.

Cost-effective

Adding new data to the regression model is most cost effective when done in batches of several buildings. These would take some time to validate, incorporate, and test the model and would need a minimum time interval of every 6 months to be actioned regularly.

Methodology

Mathematical Approach

The approach is summarised below:

1. **Data entry –** For each of the 44 buildings, vehicle records were captured over three working days. Data was used to generate descriptive statistics.
2. **Demand inference –** Demand was calculated for each time point *t* (minute) of each of 3 days for each building for small, medium and large vehicles.
3. **Compute peak demand for each vehicle class for each building –** Peak demand from vehicles of size *j* in building *i* ($Demand\_peak_{ij}$) is the maximum demand observed over the three days corrected for outliers' presence. The 99th percentile was used instead of the maximum so that extremely unusual situations occurring less than 1% of the time are neglected to exclude sporadic demand.
4. **Predictive modelling of peak demand based on a building's characteristics**

   - **Conceptual Modelling**

   $Demand\_peak_{ij}$ was modelled by creating 3 models: one for each vehicle class, based on 44 observations (1 per building). This model was intended to predict peak demand for each vehicle class for a given building, using residential area, number of apartments, commercial, hotel rooms and retail area, and availability of a dedicated elevator as inputs. The potential pool of predictors was based on the conceptual model, which considered moderating effects of the number of floors, presence of a dedicated elevator, and primary use type on the relationship between residential, commercial, hotel rooms and retail space peak demand.

   - **Modelling method justification**

   The estimation procedure was selected according to several criteria that allow the inclusion of nonlinearities (such as log-linear relationships instead of linear ones) and interactions to account for the possibility of a differential effect of one variable on the outcome depending on the value of another variable. The model should be able to work with many predictors when the sample size is small. Regularised generalised linear regression models appeared to meet the above criteria. Regularisation essentially means eliminating irrelevant predictors. Two variable selection techniques were used, including LASSO and the stepwise variable selection minimisation technique.

   - **Model Results**

   The demand for small vehicles is determined by *retail_area*, *commercial_area*, *hotel_area, log(residential_apts+1)* and the interaction between (i.e. product of) *commercial_area* and *elevator_dedicated*. The model implies that having dedicated elevator results in a lower accumulation of small vehicles in the commercial's building dock if there is a dedicated elevator.

   The demand for medium vehicles is determined by *log (retail_area + 1)*, *commercial_area*, *log (residential_apts + 1)*, *log(floors), hotel_area and commercial_area: elevator_dedicated*.

   The demand for large vehicles is determined by *log (retail_area + 1), commercial_area*, *hotel_area* and *log (floors)*.

   - **Model Validation**

   Despite the natural presence of some unexplained heterogeneity in peak demand across buildings, the resulting set of 3 models was shown to fit data reasonably well. Predicted demand values and actual demand across the 44 buildings are strongly correlated as Pearson's correlation coefficient of 0.78 represents a high correlation. The mean absolute error (MAE) is 2 vehicles, which is a good result considering that the total demand varies across 44 buildings from 5 to 21 with a mean of 8.9 and standard deviation of 5.6. The R^2 of the predictive model is 0.61, which indicates a reliably accurate model prediction for the demand.

   - **Inferring common shapes of intra-day relative demand dynamics**

Although peak demand "maximum parking demand throughout the whole day" varies depending on the building's size and other factors, intra-day trends of relative demand (percentage of peak demand) were found to be consistent across buildings of the same primary use category. Simply, the demand patterns based on relative demand is more accurate than values of demand, i.e. more accurate to predict that 45% of peak demand in a building happens at 9:35AM than predicting 3 small spaces are specifically occupied at 9:35 AM.

- **Unsupervised clustering of buildings by intra-day dynamics**

Buildings were clustered using all 18 variables produced using the 6 time periods and 3 vehicle classes present in the model to reveal common shapes of intra-day dynamics of relative demand e.g., average relative parking demand for small vehicles from 6 am to 9 am).

Two distinct clusters were discovered using a model-based clustering technique. Cluster 1 is distinguished by a distinct peak between 9 am and 12 pm. Cluster 2 lacks a consistent peak and has a more evenly distributed parking demand during the day. It was discovered that clusters are linked to the primary use sort. Cluster membership prediction is rational and straightforward since all primarily commercial buildings belong to cluster 1 and all primarily residential buildings belong to cluster 2.

Regression Outputs

The mathematical approach produces three sets of regression coefficients used by the model:

1. Predictors of Freight Demand – feeds into 'Simulated Demand' output tab
2. Arrival Rate by Land Use – feeds into the 'Arrivals Attribution' output tab
3. Parking Duration – feeds into the 'Parking Duration' output tab

Predictors of Freight Demand Coefficients

The UFFM was developed using a predictive modelling approach that incorporates regression analysis with a clustering technique. The selection of the mathematic approach to building the model using predictive modelling was based on several factors, namely:

- promoting a versatile and scalable method for model inputs,
- allowing different datasets and new variables to be used (e.g., new land-use types or vehicle types),
- being ideal for typical land-use styles and building sizes.

The approach (regularised general linear model (GLM) with clustering) allows for convenient and practical updating of results when new parking surveys become available. The methods are universal, general-purpose and are neither limited by the number of buildings, nor by the number, nor type of variables involved in peak demand prediction.

The initial output of the regression analysis reveals a large set of potential model predictors, e.g. 15-20 different variables that may be selected as candidate predictors. However, for this model, the number of predictors has been refined down to only 5-8 to select the most critical/statistically significant predictors for the final model. The R code included below includes a regression algorithm "glmnet & step functions, which are a RLASSO function for GLM-based variable selection in R" that allows such a selection process to identify the key predictors.

Example: The regression analysis of existing 44 buildings revealed the following predictors - *log(retail_area + 1)*, *commercial_area*, *log(floors)*, *log(residential_apts + 1)*, *retail_area, hotel_area and commercial_area multiplied by elevator_dedicated which are shown.* We have 4 parameters in the model for large vehicles, 5 – for medium vehicles and 5 – for small vehicles columns in the table should always be "class" (vehicle size), "term" (the predictor), "estimate" (the coefficient of the predictor) as shown in the screenshot below. Similarly, the regression of new datasets (new buildings) may produce similar or different predictors for the vehicle classes and land-use units.

class	term	estimate
Large	(Intercept)	[redacted]
Large	log(retail_area + 1)	[redacted]
Large	commercial_area	[redacted]
Large	log(floors)	[redacted]
Large	hotel_area	[redacted]
Medium	(Intercept)	[redacted]
Medium	log(retail_area + 1)	[redacted]
Medium	commercial_area	[redacted]
Medium	log(residential_apts + 1)	[redacted]
Medium	log(floors)	[redacted]
Medium	commercial_area:elevator_dedicated	[redacted]
Medium	hotel_area	[redacted]
Small	(Intercept)	[redacted]
Small	retail_area	[redacted]
Small	commercial_area	[redacted]
Small	log(residential_apts + 1)	[redacted]
Small	commercial_area:elevator_dedicated	[redacted]
Small	hotel_area	[redacted]
Bike	(Intercept)	[redacted]
Bike	retail_area	[redacted]
Bike	commercial_area	[redacted]
Bike	log(residential_apts + 1)	[redacted]
Bike	commercial_area:elevator_dedicated	[redacted]
Bike	hotel_area	[redacted]

Figure 8: Regression coefficients

Arrival Rate by Land Use Coefficients

The model utilises Excel's inbuilt LINEST() function which returns the coefficient of a straight line with known x and y values. In the case x values = land use and y values = daily raw counts of vehicle type across buildings.

Parking Duration Coefficients

This regression links log-transformed parking duration (in minutes) with *activity_type*, *vehicle_size* and the interactions of hours with the % of commercial and residential areas. Thus, it allows parking duration to vary across activity types, vehicle classes and assumes that the marginal impact of each hour differs depending on the % of residential/commercial area in the building.

Similarly, new types of areas can be included, e.g., log(*hotel_perc+1*). Log-transformation of *resid_perc* and *commercial_perc* was found useful to prevent too fast increase of log(p*arking_duration_num*) in response to increase of *resid_perc* and *commercial_perc*. Overall, log-transformations are often useful if they allow making the distributions of variables more symmetric (less skewed and thus less prone to the impact of rarely observed values).

Only events where parking duration was less than 7 hours were included as inputs into the regression analysis.

Monte Carlo Simulation (Simulated Demand)

The 'Simulated Demand' worksheet is a Monte Carlo simulation that generates the demand of the building.

It contains a maximum of 50 demand profile combinations (% of peak demand) for each hour of each building recorded in the survey dataset for a total of 34,800 combinations.

Percentage of peak demand was used to calculate each iteration of the simulation and used to derive absolute demand. This is because the clustering analysis of the data revealed two main intra-day patterns of residential and commercial buildings. While absolute demand *(number of utilised parking spaces at time point t)* varies from building to building quite substantially, patterns of relative demand *(proportion of utilised parking spaces at time point t to the maximum observed demand for a bldg.)*, can be more accurate and predictable. Simply, the demand patterns based on relative demand are more accurate than values of demand (i.e., more accurate to predict that xx% of peak demand happens at 9:35AM than predicting x small spaces at 9:35).

For this reason, the simulated demand figures express demand values for the 44 buildings as % of peak demand. These were converted from absolute demand values by using the R code 'Simulated Demand Script' mentioned in the previous section.

There are three types of demand calculated in this worksheet (Figure 9) which are used for different outputs. These values, along with efficacy are calculated for every iteration of the simulation.

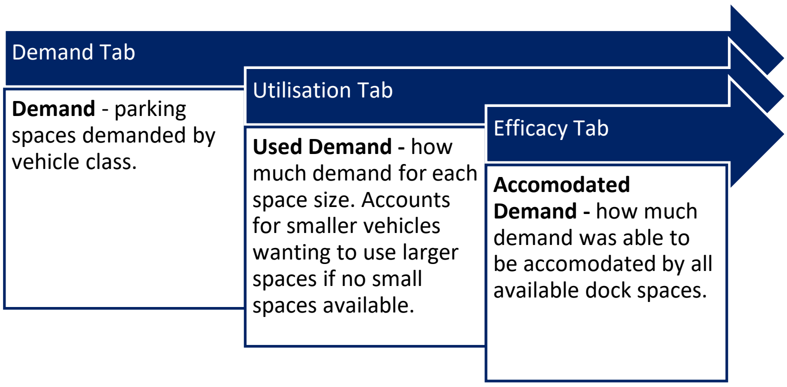

Figure 9: Explanation of demand values calculated by the Simulated Demand tab and their uses throughout the model

Cluster Sampler Sub-Process

This worksheet acts as a sub-process the model uses to filter the most relevant iterations produced the simulated demand computation based on similarity between the building characteristics entered by the user and those of the buildings surveyed.

This is done by plotting each surveyed building by land use percentage and using Euclidean distance to find the most similar.

$$\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}.$$

The inverse of this result was then multiplied with the maximum sample size of the simulation to determine how many samples of each building surveyed to capture in the 'Simulated Demand' tab.