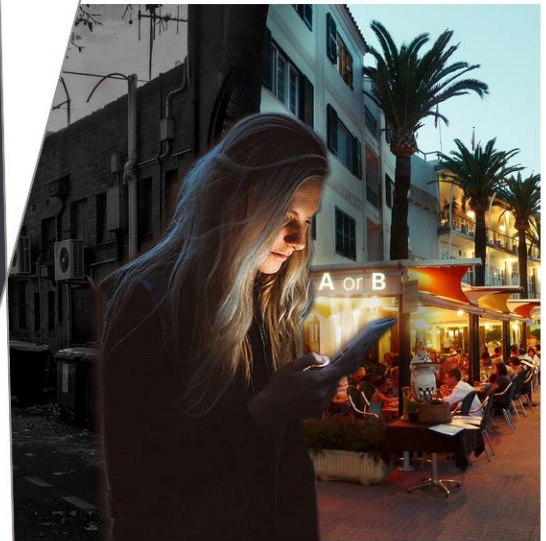


80021010-TR-TEN-0003 Technical Note

Passive Surveillance Index

80021010



Prepared for
Transport for New South Wales

23 September 2021

Contact Information

Cardno (NSW/ACT) Pty Ltd

ABN 95 001 145 035

Level 9 - The Forum
203 Pacific Highway
St Leonards NSW 2065
Australia

www.cardno.com

Phone +61 2 9496 7700

Fax +61 2 9439 5170

Document Information

Prepared for	Transport for New South Wales
Project Name	Passive Surveillance Index
File Reference	80021010-TR-TEN-0003 Technical Note.docx
Job Reference	80021010
Date	23 September 2021
Version Number	A

For more information, please contact syd.transport@cardno.com.au

Document History

Version	Effective Date	Description of Revision
A	11/11/2020	Draft
A	23/09/2021	Final

© Cardno. Copyright in the whole and every part of this document belongs to Cardno and may not be used, sold, transferred, copied or reproduced in whole or in part in any manner or form or in or on any media to any person other than by agreement with Cardno.

This document is produced by Cardno solely for the benefit and use by the client in accordance with the terms of the engagement. Cardno does not and shall not assume any responsibility or liability whatsoever to any third party arising out of any use or reliance by any third party on the content of this document.

Our report is based on information made available by the client. The validity and comprehensiveness of supplied information has not been independently verified and, for the purposes of this report, it is assumed that the information provided to Cardno is both complete and accurate. Whilst, to the best of our knowledge, the information contained in this report is accurate at the date of issue, changes may occur to the site conditions, the site context or the applicable planning framework. This report should not be used after any such changes without consulting the provider of the report or a suitably qualified person.

Table of Contents

1	Introduction	1
	1.1 Background	1
	1.2 Purpose	1
	1.3 Project team	2
	1.4 Methodology	2
	1.5 Known limitations	3
	1.6 Study area selection	4
2	Data collection and processing	5
	2.1 Survey	5
	2.2 Footpath data	10
	2.3 Business data	11
	2.4 Land use data	14
	2.5 Road hierarchy data	14
	2.6 Lighting data	14
3	Production	15
	3.1 Chosen approach	15
	3.2 Method	15
	3.3 Results	19
4	Validation	21
	4.1 Validation results	21
	4.2 Validation comparing predicted score and mean survey score	22
	4.3 Validation comparing predicted scores between survey locations and footpaths	23
	4.4 Validation of predicted score with time variation	24
5	Visualisation	25
6	Considerations for expansion	27
	6.1 Customer experience	27
	6.2 Business data	27
	6.3 Survey	27
	6.4 Lighting data	27
	6.5 Other geospatial data	28
	6.6 Other models	28

Tables

Table 2-1	Benchmark assessment variables	5
Table 2-2	Benchmark assessment	6
Table 2-3	Footpath classification	10
Table 2-4	Simplified business type	13
Table 3-1	Example – spatially joined data sets for a typical road segment	15

Table 4-1	Predicted score and mean survey score values for each location	22
-----------	--	----

Figures

Figure 1-1	Strategic approach	2
Figure 1-2	Process flow chart	3
Figure 1-3	Study area	4
Figure 2-1	Survey photograph locations	8
Figure 2-2	Score distribution for each survey location	10
Figure 2-3	Sample polygon segments	11
Figure 2-4	Business locations - shopping centres	13
Figure 3-1	Example - Forest-based Classification and Regression tool input	17
Figure 3-2	Example – Esri’s Model Builder in ArcGIS Pro	18
Figure 3-3	Output message from ArcGIS Pro Forest-based Classification and Regression (Spatial Statistics) tool	19
Figure 3-4	Boxplot of Survey Results and Predicted Score using the Forest-based Classification and Regression tool	20
Figure 4-1	Output from ArcGIS Pro displaying the distribution of R-squared for validating the model	21
Figure 4-2	Validation results from ArcGIS Pro Forest-based Classification and Regression (Spatial Statistics) tool	21
Figure 4-3	Predicted score versus mean survey score	22
Figure 4-4	Predicted score for survey data points and footpath polygons on Sunday 12am. Basemap supplied by OpenStreetMap	23
Figure 4-5	Mean predicted evening friendliness score for all footpaths on Friday	24
Figure 5-1	Comparison of Predicted Evening Friendliness score	25
Figure 5-2	Legend showing predicted evening friendliness layer for Friday selected	26
Figure 5-3	Filter to select day and time to view evening friendliness score	26

1 Introduction

The Passive Surveillance Index is provided in the following portal:

<https://cardnoanz.maps.arcgis.com/apps/webappviewer/index.html?id=82110546dbce47f58a4bbb47e94a6779>

This website works best using the Google Chrome browser.

1.1 Background



Transport for New South Wales (TfNSW) is looking for innovative and data-driven solutions to improve safety after dark from women, especially in a public transport environment.

To respond to TfNSW's challenge, UNSW and Cardno formed a collaborative project team which identified a critical challenge to understanding safety as not only be the threat of crime, but also the psychological perception of danger that may reduce participation of women in the community at night.

The project team have collaborated to develop a proof of concept for a **Passive Surveillance Index (PSI)** that estimates the level of safety based on passive surveillance items such as open businesses and available lighting.

The PSI dataset quantifies safety for pedestrians by assigning a safety score to walking routes via a predictive ranking system. These can then be visualised on a map, using a simple to understand cartography, or integrated into more sophisticated mobile routing applications.

Although criminal behaviour cannot be completely stopped, women can be provided tools to make more informed choices on how they travel.

This project is the first of its kind as our team would be the first to score walking routes using a unique set of geospatial data sources validated with experiential input. The chosen geospatial data approach also allows this process to be repeatable across other study areas with adjusted sensitivity, without relying on user input through incident reporting.

1.2 Purpose

The purpose of this Technical Note is to provide TfNSW with the methodology and processes undertaken by Cardno and UNSW developing the PSI, as part of the Safety After Dark Innovation Challenge.

This Technical Note documents the overall methodology and identifies critical decisions points, assumptions made, current limitations, and how these limitations may be mitigated in future iterations of the PSI.

This document is intended to be referred to when the Passive Surveillance Index (PSI) is scaled-out and expanded upon as per the high-level expansion approach detailed in **Figure 1-1**.

This Technical Note is current as of February 2021 and includes all progress to date and will be finalised upon completion of the PSI.

Figure 1-1 Strategic approach



1.3 Project team

Cardno’s Transport Advisory Team in Sydney is collaborating with Dr Meead Saberi from the UNSW School of Civil and Environmental Engineering (the project team).

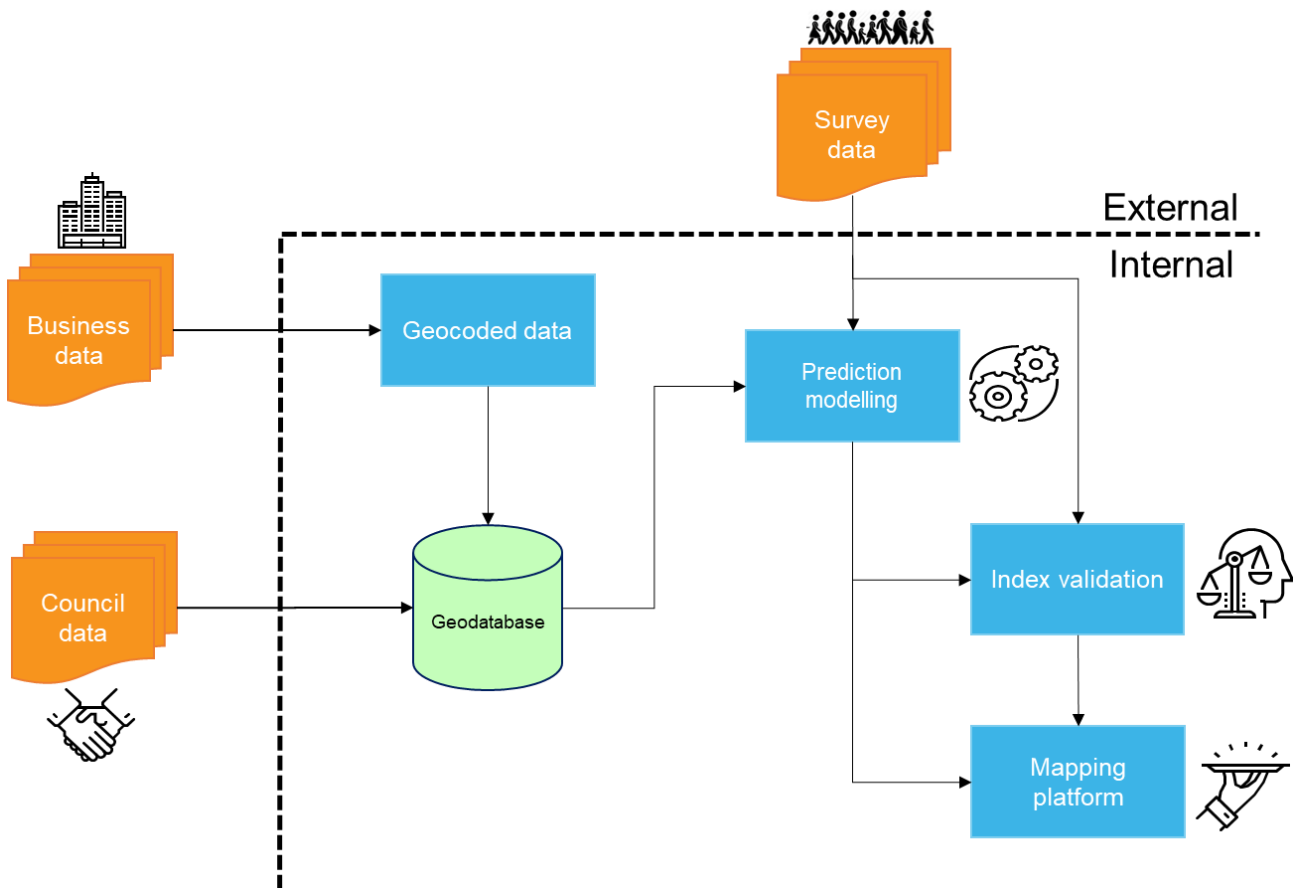
Parramatta Council’s Night Time Economy Officer is a key stakeholder in this project.

1.4 Methodology

This Technical Note outlines the processes used to develop a scalable and repeatable PSI methodology as a proof of concept. The proposed methodology and use of data has been outlined in **Figure 1-2** and describes a four-phase process of:

- > Data collection & processing (see **Section 2**);
- > Production (see **Section 3**);
- > Validation (see **Section 4**); and
- > Visualisation (see **Section 5**).

Figure 1-2 Process flow chart



1.5 Known limitations

A summary of the key limitations of the methodology are outlined in this section and discussed in context throughout this Technical Note. It is intended to improve upon this process for future stages of the PSI.

- > Business data was collated from multiple sources. The resultant dataset was incomplete and required general assumptions on opening and closing times of local businesses and services. It is recommended that a consolidated data set is used for future development, such as the Google Places API. The use of Google Places API is currently being investigated which is further explained in **Section 2.3**.
- > The accuracy of business opening and closing times from the selected dataset may be affected by changing conditions due to COVID-19. The business data is current as of 2020 and it is unknown whether the business data reflects changes due to COVID-19.
- > Validation of the PSI is based on experiential data gathered from an online survey comprised of a sample size of 47 participants. A larger sample size is recommended for increased accuracy. This is further explained in **Section 2.1**.
- > Machine learning was used to make predictions of safety by correlating survey outcomes with geo-spatial data, such as lighting, business (opening hours) and land use. Survey results were carried out in specific locations that had characteristic elements represented in the data, i.e. photos shown to respondents which depicted night-time locations within the city, had corollary data points within a GIS that allowed for a machine-learning model to make predictions elsewhere within the study area, where there had been no survey. It should be noted however, that the predictions made by the model can only be as good as the input data and as such, improved data is necessary to make more robust and accurate predictions. The use of machine learning is further explained in **Section 3.2**.
- > Photographs taken at night for the survey were produced to match the photographer’s own sense of how the location appeared. While every effort was taken to ensure accuracy and consistency, photographs are primarily intended to provide a prompt for the respondents to recall their own personal experiences within a particular location.

1.6 Study area selection

Key criteria considered for study area selection were:

- > Large metropolitan centre with lots of activity;
- > Diverse demographics; and
- > Active or emerging night life economy.

The two locations that were considered were the Sydney CBD and the Parramatta CBD.

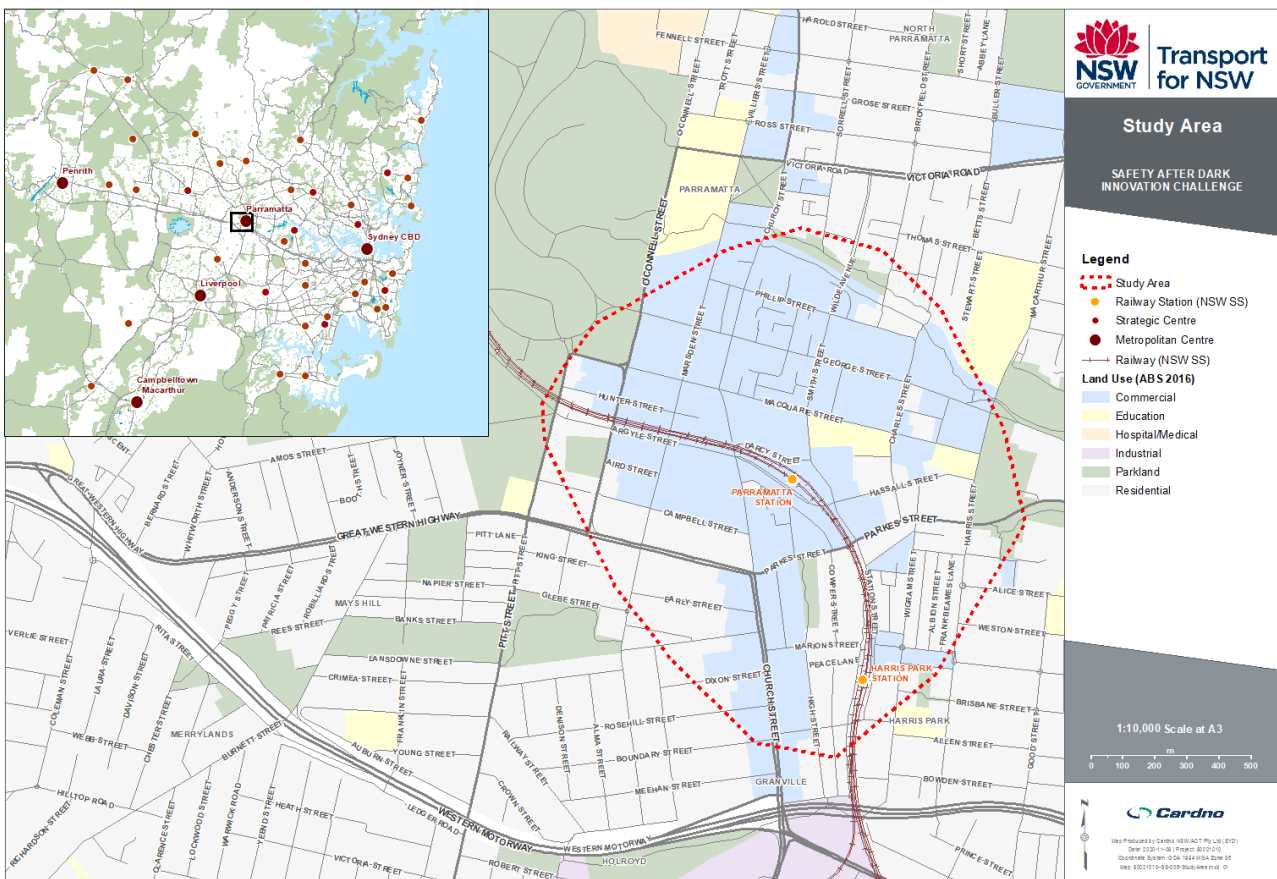
The Parramatta study area was selected due to its strategic significance in Sydney's transport network, as a key interchange point for public transport services. Parramatta is undergoing significant development to emerge as a metropolitan centre in Sydney's metropolis of three cities, including expansion of night life and economy in the centre. Currently, there is an apparent disconnect between night life locations and the public transport interchange.

Development of the PSI relies on assessment of key criteria datasets. It was important that reliable data was available for the chosen study area. Parramatta Council confirmed the availability and suitability of the data.

The study area extent was determined as the 800-metre walking catchment from the two entrances to Parramatta Station. 800 metres equates to an approximate ten-minute walk. It extends to cover the commercial core of the Parramatta CBD. The extent was calculated using the catchment tool in ArcMap software.

The study area is shown in **Figure 1-3**.

Figure 1-3 Study area



2 Data collection and processing

This section outlines the various data sets used in the development of the PSI, assumptions made and limitations.

2.1 Survey

The purpose of the survey was to gather information from women regarding their perception of safety within a specified geographic location where they have experience walking at night. This survey data was then used to both train and validate a PSI machine learning model.

The survey was created using Qualtrics with UNSW’s institutional license. Qualtrics is an easy to use, web-based platform for creating and distributing online surveys.

- Validation is further detailed in **Section 4** and the development of the machine learning model is explained in more detail in **Section 3.1**.

2.1.1 Design

The survey was designed by UNSW to measure perceptions of safety from a sample of women who regularly walk within the study area.

Survey participants were shown a series of photos taken within the study area, at night, and asked to score on a scale of 1 to 10, how safe they “felt” within each area based on both their experience and what was visible in the photograph.

The photos were taken at a variety of distinctive locations within the study area, with highly variable land uses, lighting and infrastructure provision. Each location was photographed with camera settings that matched the photographer’s own sense of how the location appeared at night.









The variables for the benchmark assessment are shown in **Table 2-1**.



Table 2-1 Benchmark assessment variables

Variable	Score		
	0	1	2
Land use	Either residential, commercial, park, CBD, train station		
Businesses (retail, restaurant, café etc)	Closed or none	Open	-
Lighting level	Dark	Medium	Bright
People crowd	No people around	Some people around	Crowded
Traffic presence	No traffic	Some traffic	Congested
Footpath width	Narrow	Medium	Wide

The photos used in the survey and the project team’s benchmark assessment are provided in **Table 2-22**.

Table 2-2 Benchmark assessment

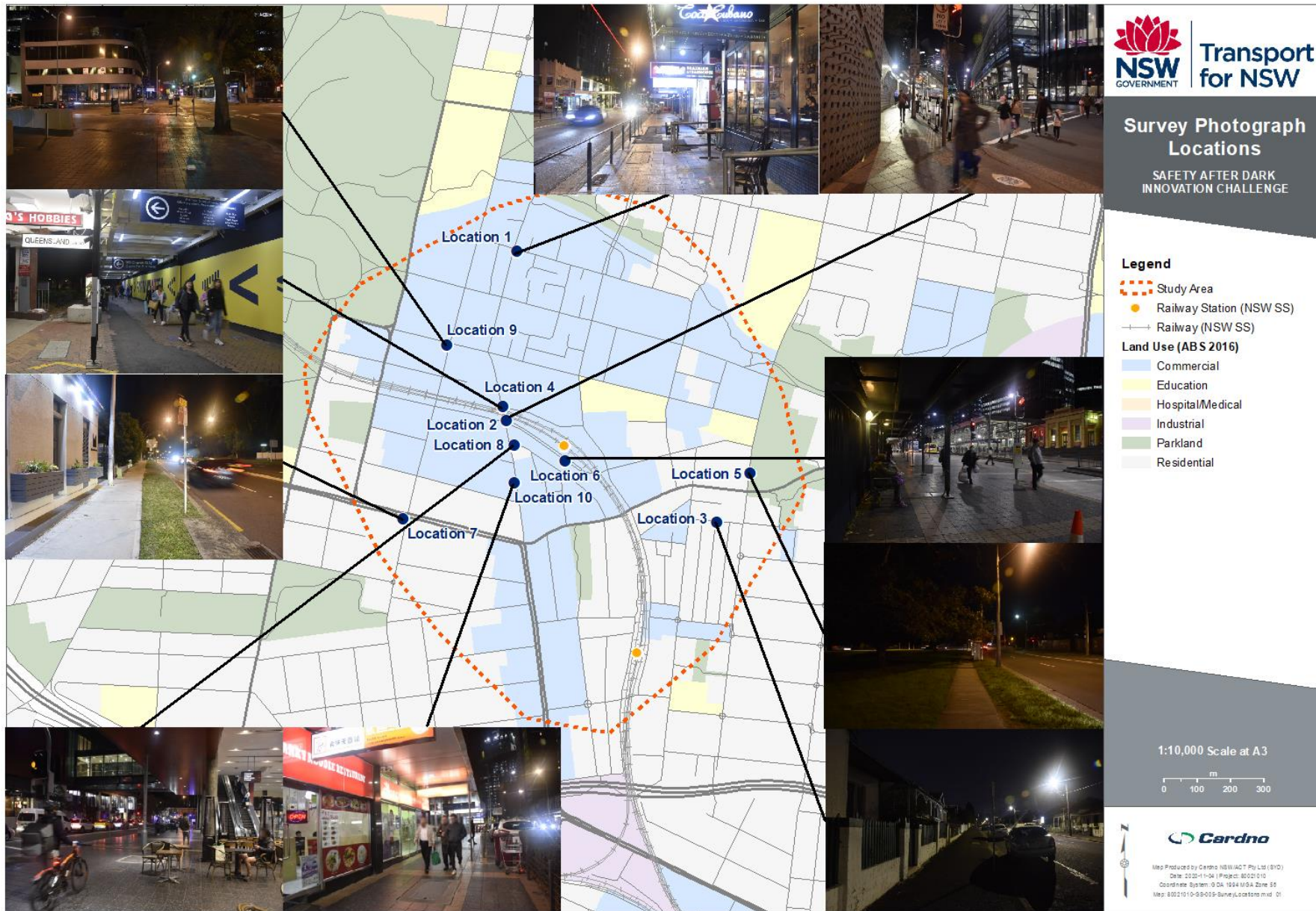
#	Photo	#	Photo
1		2	
3		4	
5		6	
7		8	

#	Photo	#	Photo
9		10	

The survey was created using Qualtrics software. It should be noted that a number of suitable alternative programs exist.

The location of each photograph is shown on **Figure 2-1**.

Figure 2-1 Survey photograph locations



2.1.2 Privacy

A privacy statement was included on the first page in the survey. This statement identified that TfNSW is committed to protecting any private information for survey participants and that no personal information would be disclosed to any third parties unless consent was first obtained.

A complementary privacy statement from UNSW was also included in the survey along with information on who was running the survey and what information was being collected.

The survey was conducted as part of UNSW's general research program to study pedestrian walking behaviour with an approved ethics application (HC190093).

2.1.3 Criteria for participation

The criteria used to determine eligibility for participation in the survey was:

- > Provision of an email address;
- > Gender was female; and
- > In a usual week, at least one walking trip is made within the City of Parramatta.

2.1.4 Advertisement activities

The survey was advertised by UNSW via social media platforms, Facebook and Instagram, for a period of one week. These were paid-for advertisements which incentivised participants to complete the survey for \$100.

The advertisement included a link to a short survey that asked interested participants to identify their eligibility criteria as per **Section 2.1.1**.

After indicating their eligibility and interest, potential participants were asked to provide an email address and await further instructions.

2.1.5 Distribution

The survey was distributed via email to selected participants and was active for a period of 4 weeks. The survey was reviewed and approved by TfNSW prior to commencement.

2.1.6 Survey reliability

The sample size of the survey was relatively small (47 participants). A larger sample size would produce a more robust model and generate more accurate results.

For this proof of concept, however, the sample size was deemed appropriate, as it clearly demonstrates that the methodology is sound but should be expanded upon when this study progresses to larger study areas.

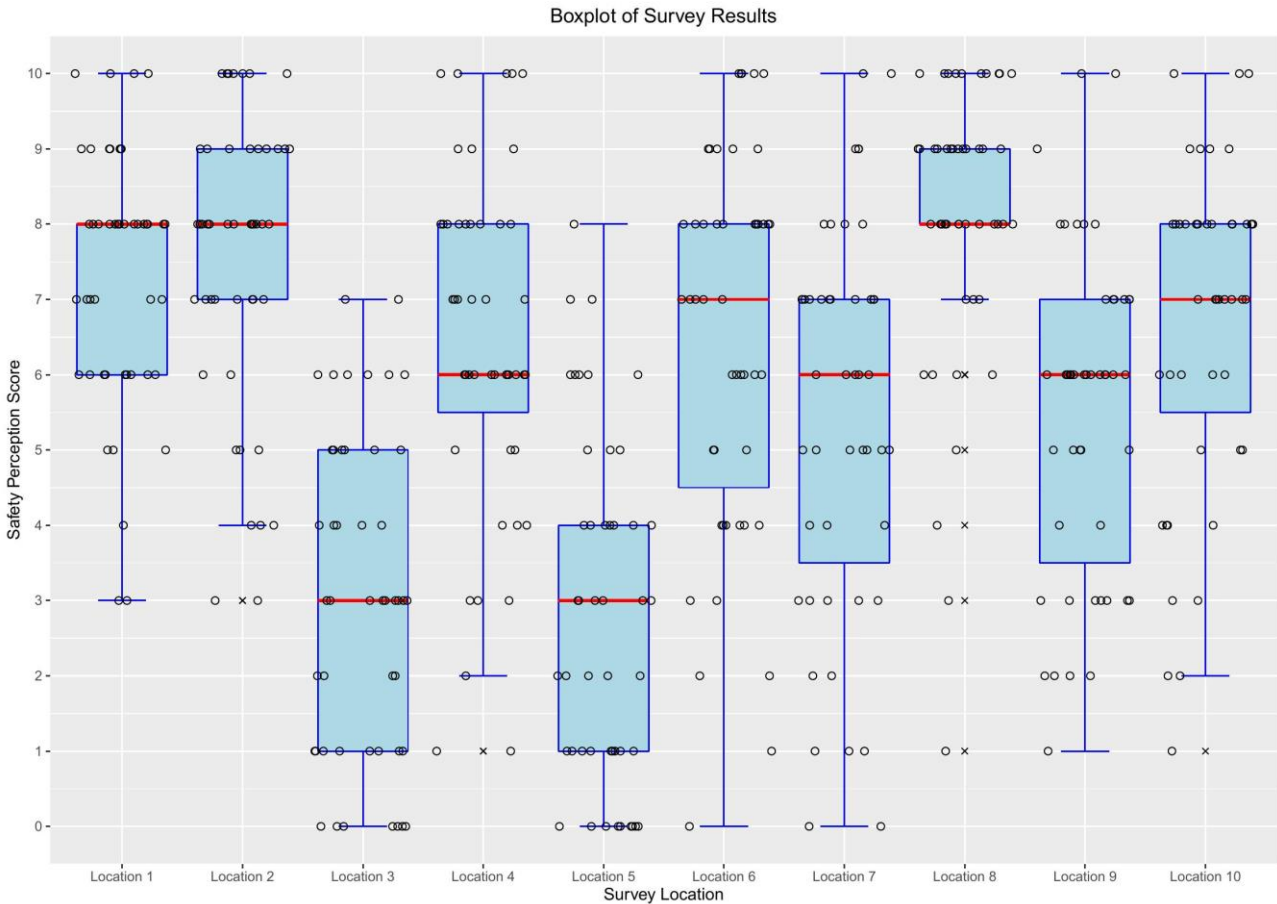
2.1.7 Results

The score distribution for each location is shown in the box plot in **Figure 2-2**. The red line represents the median score, and the blue box represents the 1st and 3rd percentile for each survey location. Furthermore, the blue lines extending from the box represent the minimum and maximum scores.

Also shown in **Figure 2-2** are the individual responses from each of the 47 participants displayed as circles. The variation of the circles in the left-right direction is random and is purely to aid in understanding the number of responses per score. Some responses were found to be statistical outliers and are shown as a cross, most notably in location 8.

The photo locations that participants deemed to provide the highest scores for safety are locations 1, 2, and 8. This may be due to the bright lighting or proximity to Parramatta Station. Locations 3 and 5 had the lowest scores possibly due to photos taken in those locations having less lighting and fewer people around.

Figure 2-2 Score distribution for each survey location



2.2 Footpath data

Footpath data for both sides of each road was provided by Parramatta Council as a polygon shapefile. It was decided that the PSI will represent a safety rating for both sides of each road segment (between intersections). Therefore, the footpath polygon layer was chosen as the base layer for the index and all data was spatially joined to this layer.

The model required the width of each footpath segment but the provided data lacked this attribute so the project team derived it by simply dividing any given footpath segment's length, by its area.

Each footpath segment was then assigned a width category from 0 (narrow) to 2 (wide) seen in **Table 2-3** below. It is based on Austroads guidelines.

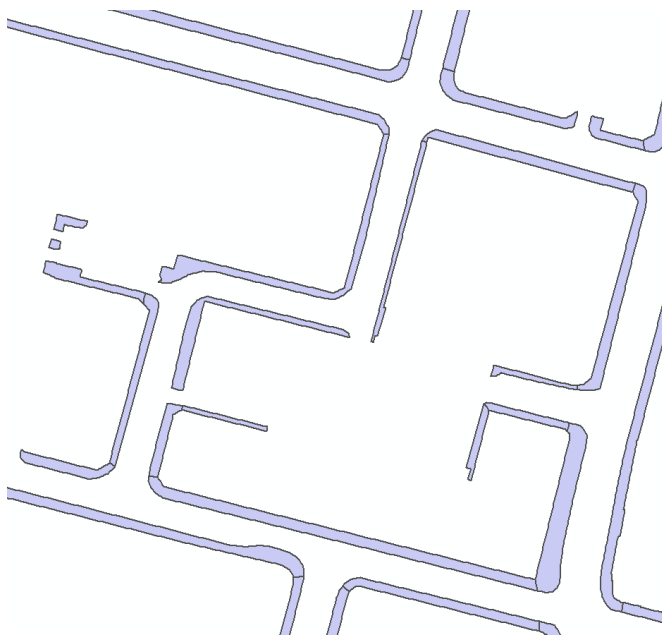
Table 2-3 Footpath classification

Footpath Width (m)	Footpath Category
< 1.2	0
≥ 1.2 and ≤ 1.8	1
> 1.8	2

Streets with missing footpath segments still require an index rating, so a polygon was drawn to represent the verge and the value for the width was entered as zero.

A sample of footpath polygon segments are shown in **Figure 2-3**.

Figure 2-3 Sample polygon segments



2.3 Business data

Business location and opening and closing time data was a key criterion in the development of the PSI. The importance of the data assumed that open businesses on street frontages provide passive surveillance and lighting to pedestrians and therefore increased the perception of safety. This section outlines the process of obtaining and analysing the business data.

2.3.1 Source of data

The use of Google Places API was investigated as a primary source of business data. The use of the API in the production of new content is not legally permitted by Google. A partnership with Google is currently being investigated in order to secure usage of the Google Places API for future work.

The adopted method for collection of business data involved collating data from multiple sources and combining them into a master data set in the form of a point feature class.

The sources are:

- > HERE;
- > Foursquare; and
- > Open Street Maps.

A business data superset was created by combining the business data-sources into a single Esri Feature class. Within the study area, HERE had 936 businesses and Foursquare had 1,024 businesses. By spot-checking the data-sources, it was decided to omit OpenStreetMap as HERE and Foursquare contained more businesses and had more accurate positional and relevant attribution information.

The superset contained the following common attributes:

- > Business name;
- > Business location (X,Y);
- > Business type (high level classification i.e. restaurant, fast food, shop, bank etc); and
- > Opening and closing times (all days).

2.3.2 Business data considerations

Cardno's model benefited greatly by sanitising the business data, however, this was a time-consuming exercise which could be mitigated using a single data source, such as the Google Places API.

Some of the primary issues encountered during data sanitisation included:

- > Inconsistent format of data;
- > Business street address not provided - location information is provided only through geo-coordinates and these are often inaccurate i.e. placed in the middle of a road carriageway. In instances like these, Cardno made various assumptions to plot them. This is discussed further in **Section 2.3.3**.
- > Some businesses have missing attribution, such as opening and closing times. This is further discussed in **Section 2.3.4**.
- > Shopping centres contain many businesses however they are not all street-facing and therefore provide no level of passive surveillance to pedestrians on the street. Business data for shopping centres required manual cleansing (further outlined in **Section 2.3.3.1**).

2.3.3 Data processing

A key criterion for the PSI is the number of businesses open for each road segment. Each business needed to be spatially joined/ attributed to its associated side of road. As previously discussed, some business point locations are inaccurate. The adopted method of assigning each business point to its associated side of the road (footpath polygon) incorporates a scalable and repeatable process, as follows:

1. Business points associated with shopping centres addressed as per **Section 2.3.3.1**.
2. Create a five-metre buffer polygon for each footpath polygon to create a larger area capturing business locations. This was done using the Buffer tool on ArcMap;
3. Spatially count the number of businesses within each polygon buffer; and
4. For the points that do not lie within the polygon buffer (due to inaccuracies of point placement or other issues) use the Near¹ association tool for ArcMap to spatially join to the footpath buffer polygon layer.

2.3.3.1 Consideration of shopping centres

Shopping centres in Parramatta were identified using NSW Spatial Services points of interest geo-data and associated with cadastre areas provided by Parramatta Council.

The businesses located within the shopping centre cadastre area were discounted. An example of this is illustrated in **Figure 2-4**.

¹ More information on this tool can be viewed here: <https://desktop.arcgis.com/en/arcmap/10.3/tools/analysis-toolbox/near.htm>

Figure 2-4 Business locations - shopping centres



2.3.4 Opening and closing times

Opening and closing time data for businesses were collected for the periods of Monday to Thursday, Friday, Saturday and Sunday. Each business was assigned a general type, and standard opening and closing times were assigned to those businesses where this data was unavailable. General business types and assumptions made on opening and closing business times are shown in **Table 2-4**.

Table 2-4 Simplified business type

General type	Assumed opening and closing time (where data is unavailable)			
	Monday to Thursday	Friday	Saturday	Sunday
Business	9:00am – 5:00pm	9:00am – 5:00pm	9:00am – 5:00pm	10:00am – 4:00pm
Restaurant	11:00am – 9:00pm	11:00am – 10:00pm	11:00am – 10:00pm	11:00am – 9:00pm
Fast food	10:00am – 10:00pm	10:00am – 11:00pm	10:00am – 11:00pm	10:00am – 10:00pm
Education	9:00am – 5:00pm	9:00am – 5:00pm	Closed	Closed
Grocery store	7:00am – 8:00pm	7:00am – 8:00pm	7:00am – 8:00pm	7:00am – 8:00pm
ATM	24 hours	24 hours	24 hours	24 hours
Business opened at night	11:00am – 11:00pm	11:00am – 11:59pm	11:00am – 11:59pm	11:00am – 10:00pm

2.4 Land use data

The Australian Bureau of Statistics provides land use information for areas across Australia based on the 2016 Census. In Parramatta, they are typically classified as 'Commercial', 'Residential', 'Parkland' or 'Education'.

Different classifications could indicate what areas are more likely to be safe after dark. For example, 'Parkland' areas may be less safe due to less lighting and people whilst 'Commercial' areas could be safer due to more businesses opened and more people about.

2.5 Road hierarchy data

Data from NSW Spatial Services for the type of road was obtained and used as a criterion in the index. In Parramatta, roads are typically classified as 'Local Road', 'Distributor Road', 'Dedicated Busway', 'Arterial Road' or 'Urban Service Lane'.

This data was used under the assumption that the road type would influence perceived levels of safety. For example, it was assumed that an 'Urban Service Lane' has less activity, poorer lighting and amenities than an 'Arterial Road' or 'Dedicated Busway' and as such would be categorised as less safe. This assumption, was largely validated by the survey results.

2.6 Lighting data

Lighting data was initially proposed to be sourced from Parramatta Council, however, this data is owned by Endeavour Energy and its use was not permitted without a license.

Street lighting data was therefore captured using recent available aerial imagery from Nearmap and in areas of dense tree cover, Google street view was used to confirm the location of light poles.

The number of light poles were spatially counted for each footpath segment, and was used as a criteria in analysis. It is assumed that all light poles provide a similar level of brightness/ illumination, and are generally the same height.

3 Production

This section outlines the chosen approach and methodology for calculating the PSI following from data collection and processing.

3.1 Chosen approach

The project team have developed a scalable methodology that heavily utilised Esri’s ArcGIS platform. One component in particular, ‘Forest-based Classification and Regression’², was used to perform the analytics and produce the final pilot PSI model.

This Forest-based Classification and Regression tool was utilised because of its ability to create powerful machine learning models which can make accurate predictions based on learning from the correlation of training data (survey results) with explanatory datasets (geospatial datasets).

In this instance, the method was employed to analyse how survey respondents have scored locations (training), and predict how they may have scored remaining unsurveyed locations within the study area by correlating their answers with a variety of geospatial datasets, such as lighting, business opening hours, land use etc (explanatory).

3.2 Method

The method for the Forest-based Classification and Regression process is as follows:

1. All relevant data sets were spatially joined to the footpath polygon layer (base layer) to enrich the footpath’s attribution data and assign each side of each road a value for all of the criteria.

An example of the spatially joined data for a typical footpath segment is shown in **Table 3-1**.

The ‘Number of Businesses Opened’ variable in **Table 3-1** changes based on the day of the week and time of day. More information on how this was calculated is described in **Section 3.2.1**.

Note that in **Table 3-1**, the road hierarchy and land use variables are both categorical data (cannot order the types of road/ land use in a numerical manner) whilst the other variables are continuous (can be ordered on a numerical scale)

Table 3-1 Example – spatially joined data sets for a typical road segment

Footpath segment unique ID	Road Hierarchy	Land Use	Light Pole Count	Footpath Width Category	Number of Businesses Opened
1	Local Road	Residential	5	2	37
2	Distributor Road	Commercial	7	3	15
3	Local Road	Parkland	2	1	3

2. The results of the survey were attached to the locations within the survey photos taken in Parramatta (a score between one and ten). Information described in **Table 3-1** was also attached to each of the ten locations.

However, for the ‘Number of Businesses Opened’ variable, the number of businesses within 30 metres of the survey location was used, regardless of opening hours. Since the respondents to the survey were not notified of the time a photo was taken, it was decided that would be a suitable proxy for this variable

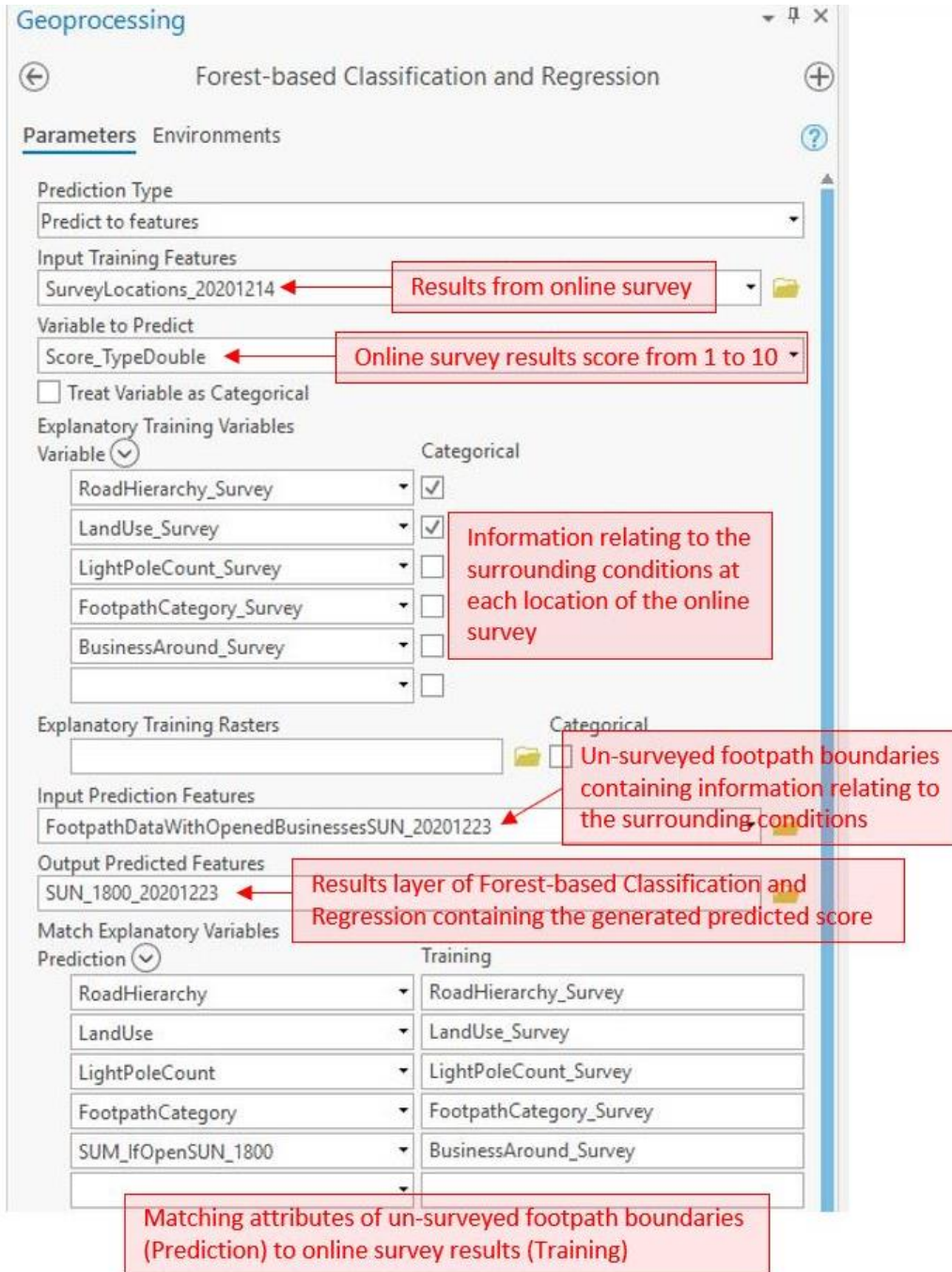
² More information on this tool can be found at: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/forestbasedclassificationregression.htm> and <https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/using-forest-based-classification-and-regression-to-model-and-estimate-house-values/>

3. The Forest-based Classification and Regression tool requires a minimum of 20 inputs on which to run the process of machine learning. Each survey point (10 locations) was duplicated for the number of participants/ scores for each location (47 times) to achieve the required number of input data points.
4. The tool compares the survey scores against the five variables outlined in **Table 3-1**, known as model training. It then determines which of these variables is most closely aligned with the survey scores. It also provides other statistical tests such as R-squared and p-values.

Furthermore, the tool uses 90% of the data for model training. It uses the final 10% of data to validate the outputs from the model. This is explained more in **Section 4**.
5. After training the model and statistical tests are satisfied, the tool is used to predict the score (PSI) at unsurveyed footpath locations. That is, the data collected from respondents for the ten survey locations is used to predict the score at unsurveyed footpath locations. The two datasets are linked by the five variables outlined in **Table 3-1**.
6. This process is repeated for the twelve hours from 6pm to 6am, and for the four days Thursday, Friday, Saturday, Sunday. It was assumed that the opening hours for businesses remains constant from Monday to Thursday.

An example of the Forest-based Classification and Regression tool input is shown in **Figure 3-1**.

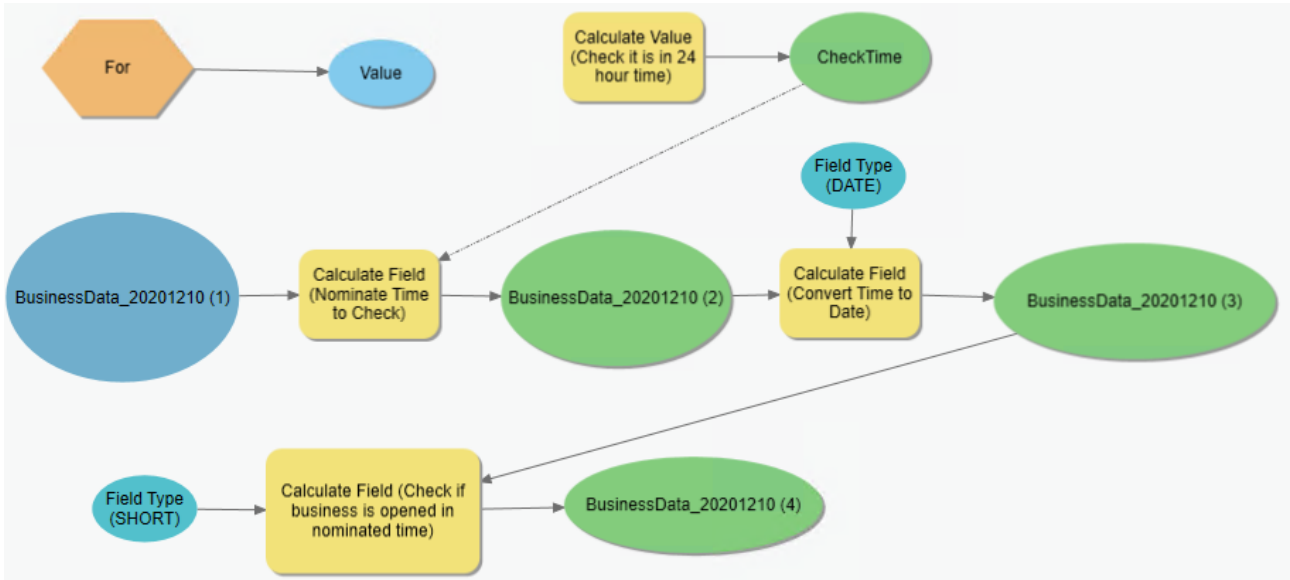
Figure 3-1 Example - Forest-based Classification and Regression tool input



3.2.1 Process to determine if a business is opened

The flowchart in **Figure 3-2** relates to determining whether a business is opened in a nominated time. If it is opened, it is given the value of 1; if closed, it is given the value of 0. The program iterates through all nominated time periods (6pm to 6am) to check whether a business is opened. The number of businesses opened for each time interval is attached to the closest footpath segment.

Figure 3-2 Example – Esri’s Model Builder in ArcGIS Pro



3.3 Results

Results from the model’s training are displayed in **Figure 3-3**. It shows that the most important of the five variables described in **Table 3-1** are land use (48%). That is, the forest-based classification and regression tool finds that land use is most important for the prediction of scores. The importance of the number of businesses opened variable is 14% indicating that it has limited importance when predicting scores.

R-squared for the training data was found to be 0.339 indicating that 33.9% of the observed variation in the survey data can be explained by the survey inputs described in **Table 3-1**.

The R-squared value of 0.339 is low, and would be improved by using a larger and more diverse sample size of input data. This translates to more survey points located across the study area in diverse land use types with varying levels of lighting, different street types and businesses.

Figure 3-3 Output message from ArcGIS Pro Forest-based Classification and Regression (Spatial Statistics) tool

```

▼ Environments
Random Number Generator 1 ACM599

▼ Messages
Running script Forest...
Random Seed: 1
----- Model Characteristics -----
Number of Trees 100
Leaf Size 5
Tree Depth Range 1-6
Mean Tree Depth 2
% of Training Available per Tree 100
Number of Randomly Sampled Variables 1
% of Training Data Excluded for Validation 10

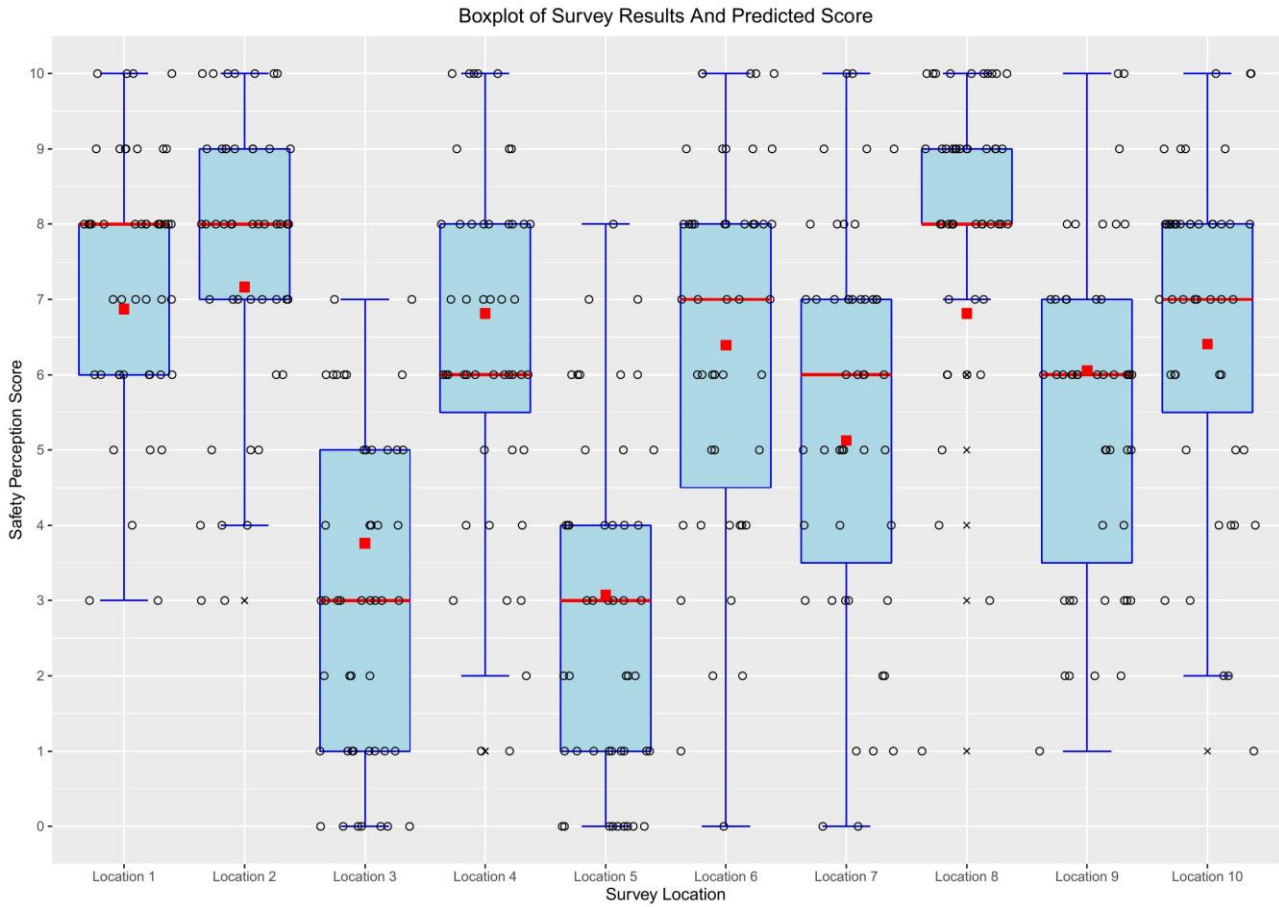
----- Model Out of Bag Errors -----
Number of Trees 50 100
MSE 5.238 5.202
% of variation explained 30.348 30.819

----- Top Variable Importance -----
Variable Importance %
LandUse_Survey 443.07 48
FootpathCategory_Survey 152.74 16
LightPoleCount_Survey 137.59 15
BusinessAround_Survey 133.86 14
RoadHierarchy_Survey 65.32 7

----- Training Data: Regression Diagnostics -----
R-Squared 0.339
p-value 0.000
Standard Error 0.019
*Predictions for the data used to train the model compared to the observed categories for those features
    
```

The tool provides a predicted score at each of the ten survey locations. This is seen below in **Figure 3-4**. Note that is the same image as in **Figure 2-2** but now includes a red square which indicates the predicted score based on the tool. Most predicted scores are within the 1st and 3rd quartile except for Location 8.

Figure 3-4 Boxplot of Survey Results and Predicted Score using the Forest-based Classification and Regression tool



4 Validation

ArcGIS Pro is a professional desktop GIS application from Esri, which allows users to explore, visualise and analyse data. This software was chosen for developing the PSI due to its high processing capability and in-built modelling tools including the forest-based regression model.

4.1 Validation results

The forest-based regression and classification tool uses 90% of the survey data for training purposes. This trained output is then used to predict the withheld 10% of the survey data³. The predicted scores are compared to the withheld survey data to determine how well the tool performed with the given variables.

The validation process was run 30 times and the R-squared for each iteration was calculated. The distribution of R-squared can be seen below in **Figure 4-1**.

Figure 4-1 Output from ArcGIS Pro displaying the distribution of R-squared for validating the model

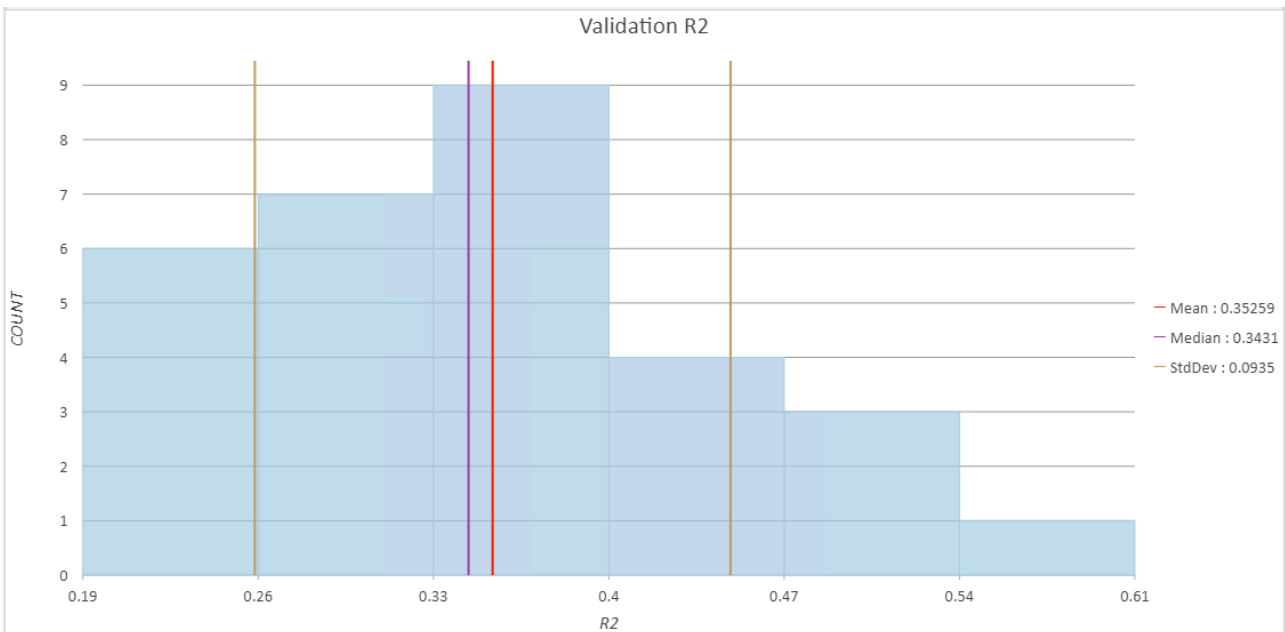


Figure 4-2 displays the results from the validation of the model. It records a median R-squared of 0.343. This indicates that the trained model (using 90% of the survey data) is explaining 34.3% of the variation in the remaining survey data.

Analysing R-squared is a form of validating the model with an R-squared of 100% indicating that the input variables perfectly fit the observed data. A higher number of input variables would improve this R-squared value and better validate the model.

Figure 4-2 Validation results from ArcGIS Pro Forest-based Classification and Regression (Spatial Statistics) tool

```

---- Validation Data: Regression Diagnostics ----
R-Squared          0.343
p-value           0.000
Standard Error     0.061
*Predictions for the test data (excluded from model training) compared to the observed values for those test features

Median R2 0.343 was approximately reached at seed 294103
    
```

³ More information can be found at: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-forest-works.htm> and <https://www.esri.com/content/dam/esrisites/en-us/events/conferences/2020/federal-gis/the-forest-for-the-trees-making-predictions-using-forest-based-classification-and-regression.pdf>

4.2 Validation comparing predicted score and mean survey score

A comparison between the mean score from the survey and the predicted score generated from the forest-based classification and regression tool is seen in **Table 4-1** below.

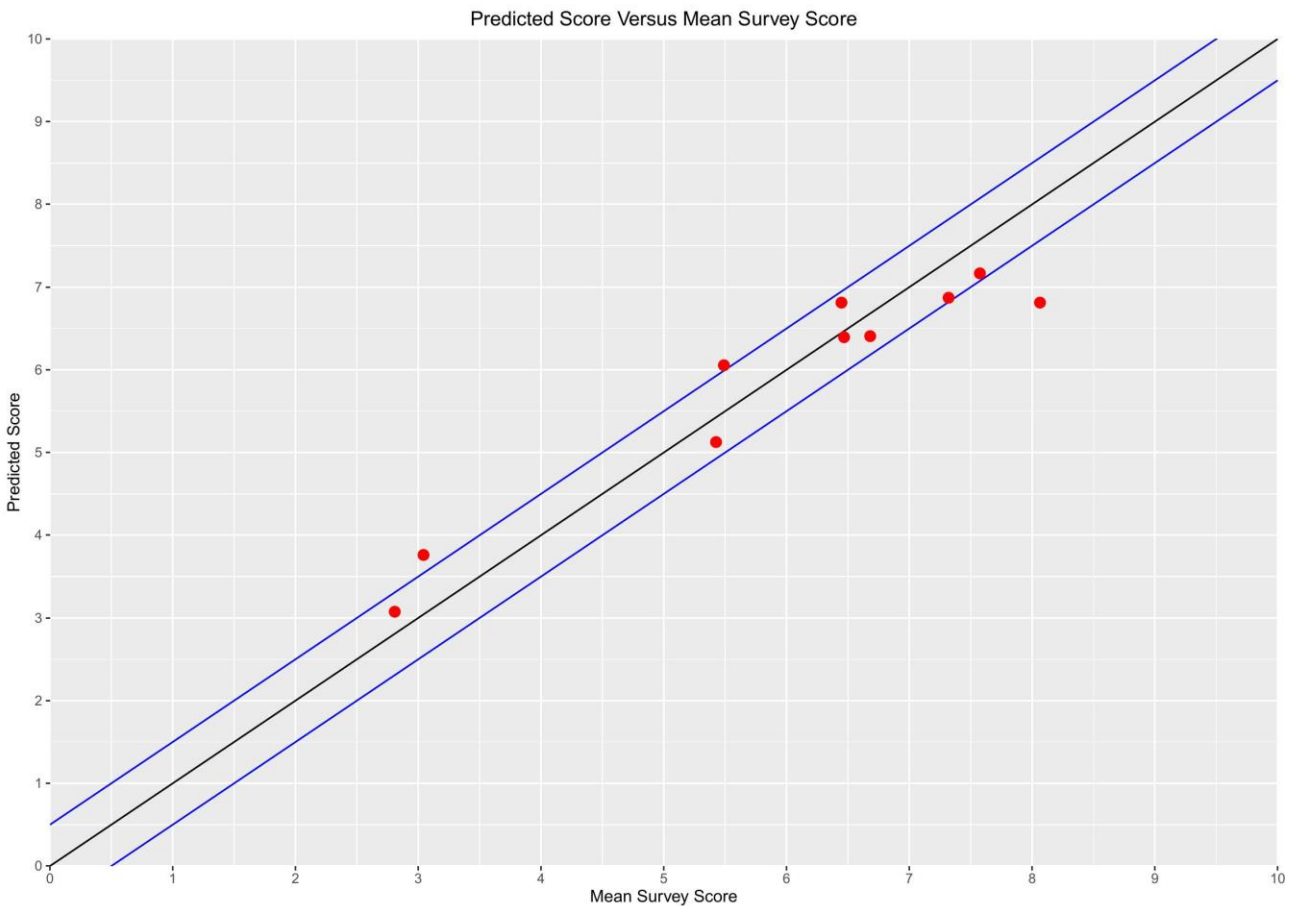
Table 4-1 Predicted score and mean survey score values for each location

Location	Predicted Score	Mean Survey Score	Absolute Difference
Location 1	6.87	7.32	0.45
Location 2	7.17	7.57	0.41
Location 3	3.76	3.04	0.72
Location 4	6.81	6.45	0.37
Location 5	3.07	2.81	0.27
Location 6	6.39	6.47	0.07
Location 7	5.13	5.43	0.30
Location 8	6.81	8.06	1.25
Location 9	6.05	5.49	0.57
Location 10	6.41	6.68	0.27

Figure 4-3 outlines the data of **Table 4-1** on a graph. It includes a black 45-degree line. If predicted scores were exactly the same as the mean score from the survey for each location, the red point would fall on this line. There are also two blue 45-degree lines which are ± 0.5 units of the black line.

A possible reason that predicted scores are not perfectly located on the black line could be due to the fact that the five variables used are only explaining 33.9% of the variation in the model. Nevertheless, the plot shows that in general, predicted scores are within ± 0.5 units of the mean survey score.

Figure 4-3 Predicted score versus mean survey score



4.3 Validation comparing predicted scores between survey locations and footpaths

The tool then predicts evening friendliness scores for unsurveyed footpaths in Parramatta based on the variable information attributed to each footpath and the machine learning using the survey data.

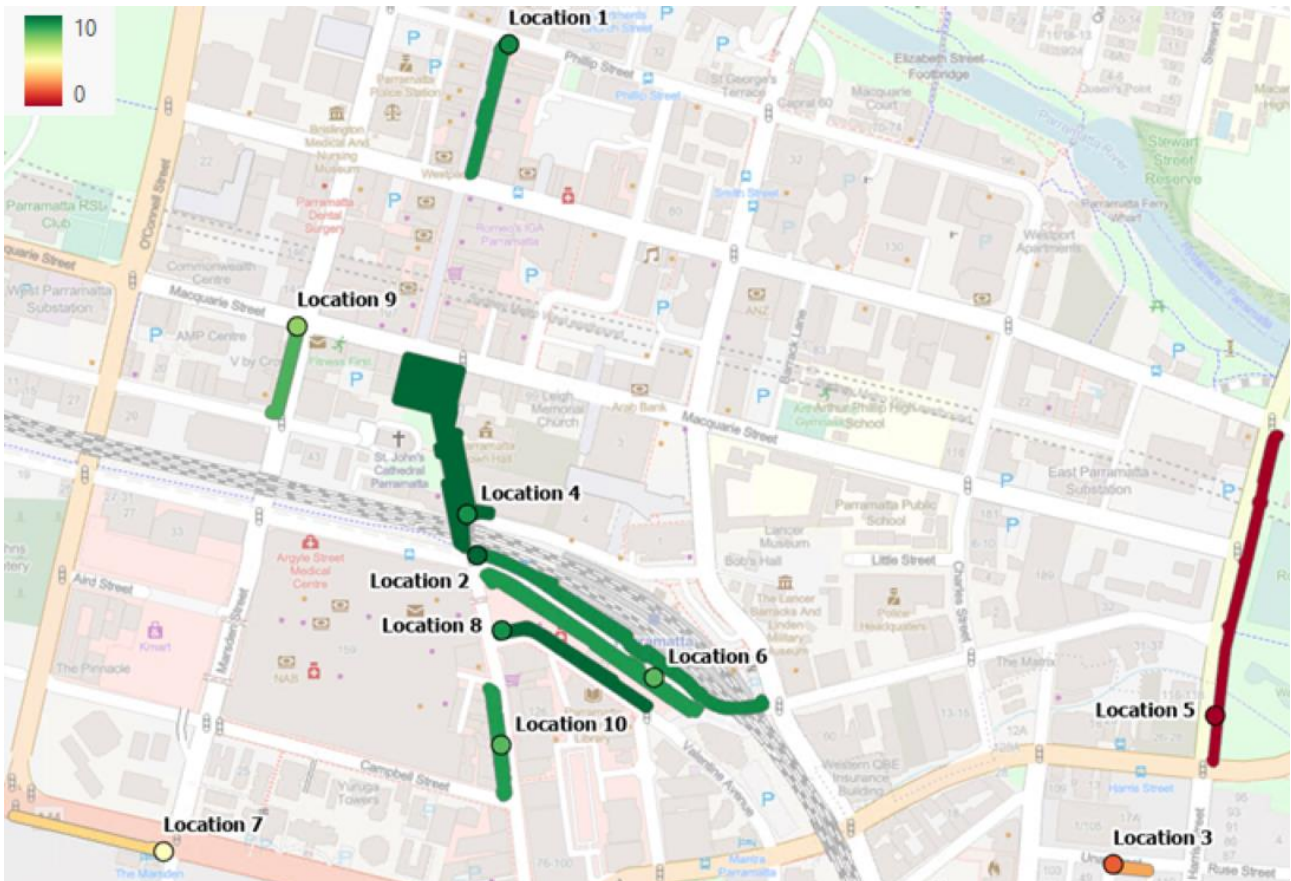
Figure 4-4 below shows a comparison between the predicted scores for each survey location and the predicted score for the closest footpath to each survey location. Note that all other footpaths have been removed for easier viewing. The figure shows the scores for Sunday 12am.

The colour distribution which represents predicted score (red is low and green is high) for the survey location point and footpath polygon are on the same scale. The colour tends to be fairly similar between the survey location point and the footpath polygon since they share four of the five variables (road hierarchy, land use, light pole count, footpath width category).

The last variable (number of businesses opened) contains different attribute information between the data sets as for the survey data, the total number of businesses within 30 metres was used whilst for the footpath data, the number of businesses opened at Sunday 12am was used. This could explain the slight difference in colour between some of the survey data points and footpath.

Nevertheless, predicted scores from the survey locations are being transferred with high accuracy to unsurveyed footpath locations.

Figure 4-4 Predicted score for survey data points and footpath polygons on Sunday 12am. Basemap supplied by OpenStreetMap



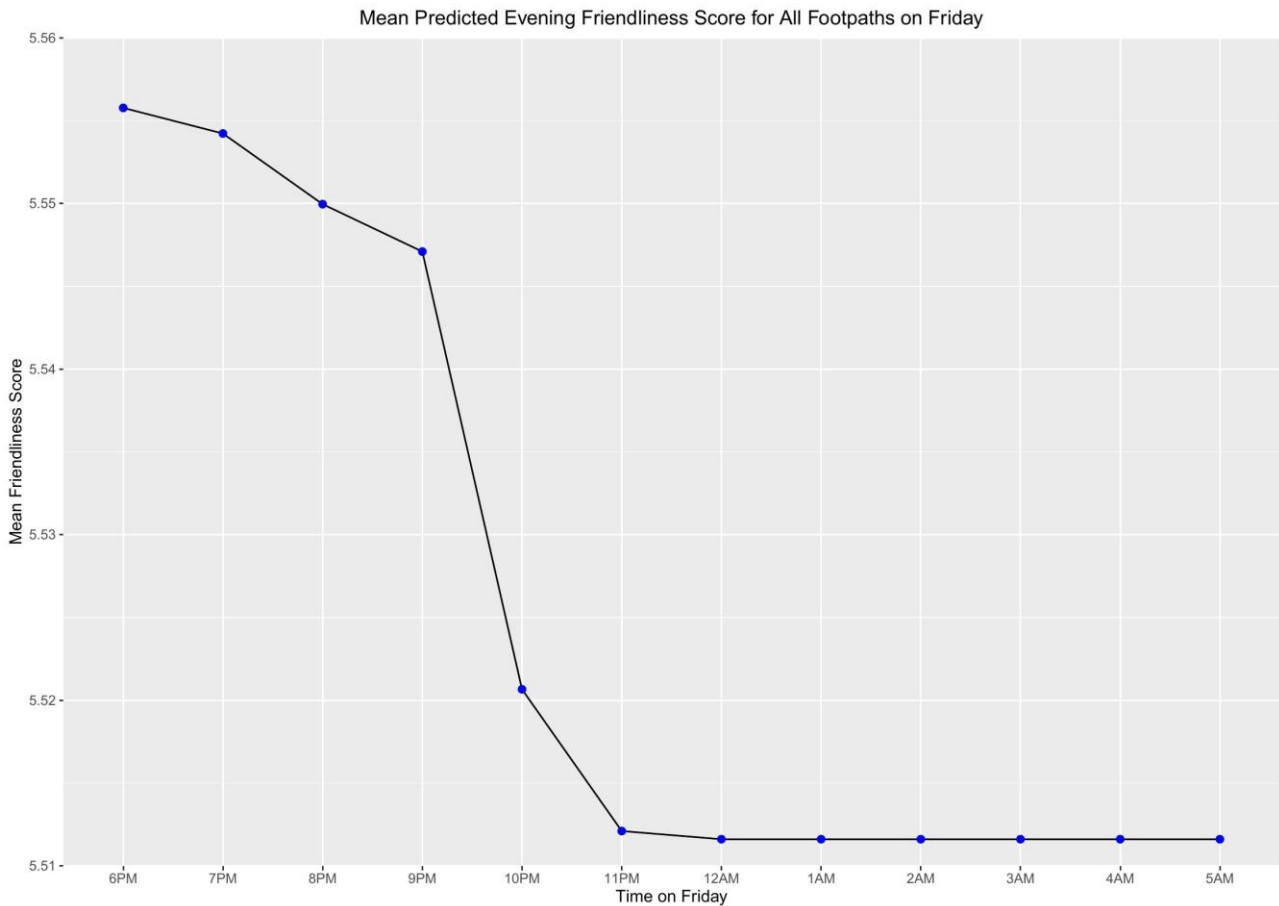
4.4 Validation of predicted score with time variation

Whilst the importance of the number of businesses opened variable was found to be only 14% as seen in **Section 3.3**, it still affected the predicted score throughout different times of the day.

The mean score for all footpaths during each time period was calculated. **Figure 4-5** below shows the mean score for Friday and how it changes throughout the evening. The score slowly decreases (i.e. becomes less evening friendly) from 6pm to 9pm. It then rapidly drops at 10pm and 11pm. This would be due to fewer businesses remaining opened at this time. It then remains constantly low from 12am to 5am.

The change in score in **Figure 4-5** below is similar to what would occur in reality, where evening friendliness decrease towards the early hours of the morning.

Figure 4-5 Mean predicted evening friendliness score for all footpaths on Friday



5 Visualisation

Predicted evening friendliness scores (PSI) were generated for all footpaths in the Parramatta area using forest-based classification and regression. These scores change based on the time from 6pm to 6am and the day of the week that is selected. The dynamic variable that controls the change in the index is the number of businesses open at the time. It is assumed that Monday to Thursday have the same score.

The results are publicly available and can be viewed on an online web portal. Different layers representing the day of the week can be turned on and off. The time of day can be selected using a filter. The portal is available at the following website:

<https://cardnoanz.maps.arcgis.com/apps/webappviewer/index.html?id=82110546dbce47f58a4bbb47e94a6779>

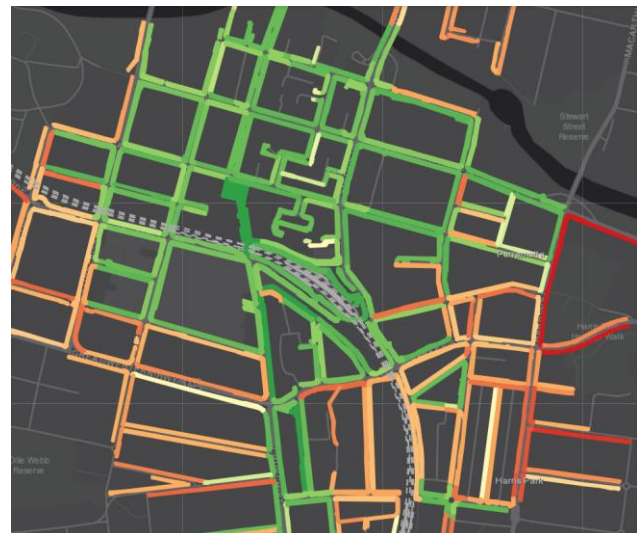
An example of the portal is shown below in **Figure 5-1**. It shows a comparison between Friday 6pm and Friday 4am. A green colour indicates that a footpath is more evening friendly whilst a red colour indicates that a footpath is less evening friendly. The most notable difference is shown in the south-eastern corners of the study area in **Figure 5-1**, as the index changes from moderately to less friendly as businesses close over the night.

Figure 5-1 Comparison of Predicted Evening Friendliness score

Friday 6pm

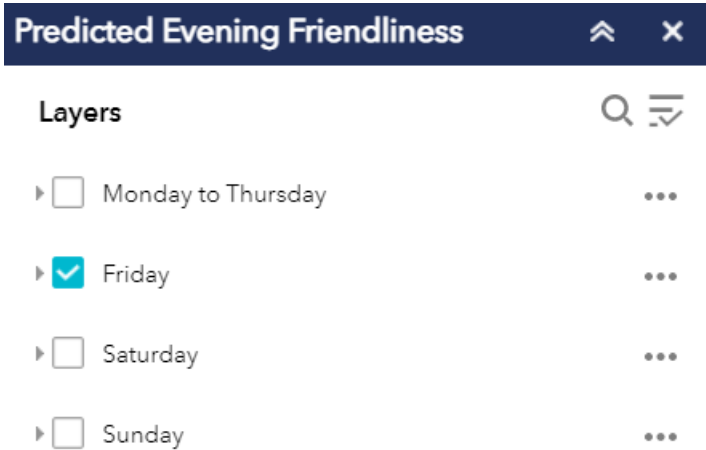


Friday 4am



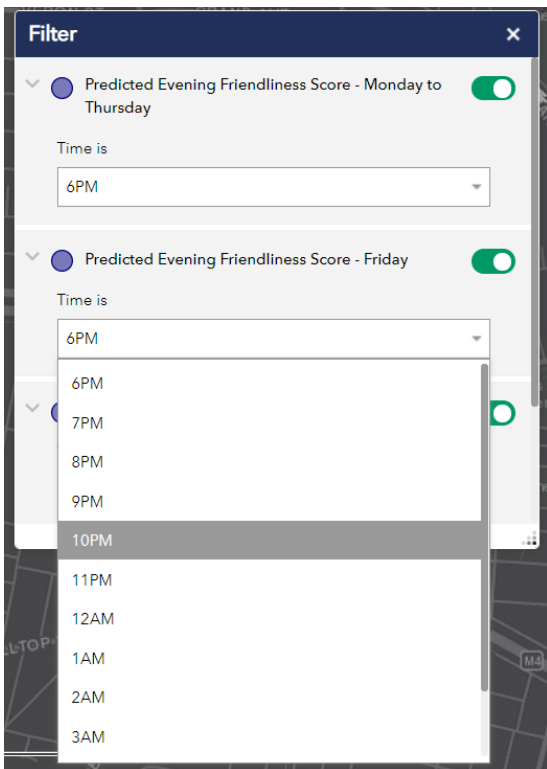
In the portal, the day of interest can be selected on the right hand side. **Figure 5-2** below shows the Predicted Evening Friendliness layer related to Friday selected.

Figure 5-2 Legend showing predicted evening friendliness layer for Friday selected



Once the day of the week is selected, the time of day can be selected using the filter within the portal. For example, **Figure 5-3** below shows how the evening friendliness score can be selected for Friday at 10pm.

Figure 5-3 Filter to select day and time to view evening friendliness score



6 Considerations for expansion

This section outlines the learnings from the project team during the production of the PSI, and improvements for potential expansion.

6.1 Customer experience

- > The index could be integrated with an existing or new wayfinding mobile application. The existing TfNSW Opal Travel App provides customers with a wayfinding service that includes the public transport Opal system cost as well as service and timetable information. The index could be integrated with the application to provide customers with the PSI service.
- > A feature could be added to the PSI allowing users to 'thumbs up' or 'thumbs down' the PSI on particular street segments or areas to continually validate the index. This information could be used to update and adjust the model.

6.2 Business data

- > Business data was a combination of HERE and Foursquare data. Duplicates between the two data sources were removed but ideally data should be from one source in order to remove this step. Also, it was found that some businesses from HERE or Foursquare were permanently closed when consulting Google.
- > Other issues with the HERE and Foursquare data was the accuracy of the business location. It was found that some business locations were actually on the wrong side of the road when comparing to Google or OpenStreetMap. More accurate location information would improve the accuracy of the index generation as it would inform which business should be connected to which footpath.
- > Many businesses had no opening/closing times and were attributed with assumed opening/closing times based on the type of business (restaurant, grocery store etc.). Ideally, more accurate information regarding opening/closing times would be needed which would better inform the time of day component of the index.

6.3 Survey

- > For the proof of concept, ten survey locations were used to feed into the prediction model. For expansion, a greater number of survey points across the study area will produce a more accurate result. Accuracy can also be improved by including a more diverse array of land uses/ street types.
- > Additional information for each survey point can be recorded, including the time of day, the number of businesses open (for example within a 50 metre radius), lighting level (captured with a light meter) etc. to be fed into the prediction model.
- > To help train the model better, photos could be taken in both day time and night time settings. For consistency, photos would be taken in the same location and focus on the same features for both times. The scores provided by survey participants would help train the temporal element of the model and produce a more accurate score.

6.4 Lighting data

- > Lighting data could alternatively be captured through sentiment, involving local residents or visitors to the area providing a rating of lighting for each street segment.
- > Lighting data could be measured in terms of the extent of the footpath that receives illumination. This could be done by creating a radius buffer around each light pole and determining its intersection with the footpath polygon layer. The proportion of footpath that is lit could be used as a criteria in the index.
- > An assessment of lighting luminosity in relation to the associated road hierarchy or path type in which it is located could provide a more nuanced result. For example road carriageway lighting may not be as luminous as footpath specific lighting and would provide a lower level of passive surveillance.

6.5 Other geospatial data

Other criteria to consider for the index are:

- > Levels and locations of vegetation. This data would give greater information on locations of higher or lower visibility, and potential locations for anti-social behaviour to occur eg bushes.
- > Building envelope data. This data provides information on site lines/ visibility around buildings, and would add to the measure of passive surveillance for pedestrians on street segments.
- > Level of activity on streets (GPS data to track movement of cars, people, food delivery users). This detailed measure of activity would provide insight on 'people busyness' of a location and provide an additional meaningful criteria to the index.

6.6 Other models

The tool used in the model is based on a machine learning technique known as random forests which is used for regression and classification problems. Other methods that could be trialled such as gradient boosted trees which potentially can improve the accuracy of results⁴.

⁴ More information can be viewed here: https://en.wikipedia.org/wiki/Random_forest

About Cardno

Cardno is a professional infrastructure and environmental services company, with expertise in the development and improvement of physical and social infrastructure for communities around the world. Cardno's team includes leading professionals who plan, design, manage and deliver sustainable projects and community programs. Cardno is an international company listed on the Australian Securities Exchange [ASX:CDD].

Contact

Level 9 - The Forum
203 Pacific Highway
St Leonards NSW 2065
Australia

Phone +61 2 9496 7700
Fax +61 2 9439 5170

Web Address
www.cardno.com

