Optimising Roadwork Project Timing - Methods Statement

Data 422 (Distance) - Group A

Contributors: Alistair Richardson, Margarita Grishechkina, Deanna Manuel

Code Repository (Github)

https://github.com/DATA422-24S2/distance_group_a.git

Contents

This document outlines the methods used in the project, including explanations of key decisions made when processing the data as well as limitations and future developments. This is covered in the following sections.

1. Processing Mobile Device Data

An essential part of this project was to take the raw data provided in *sp_data.csv.gz* and *vf_data.parquet* and transform it in a way that would make it suitable for analysis. This section outlines the key decisions that were made in this part of the pipeline and resulted in the final code *'00_process_cell_data.R'*, details of which can found in the ReadMe documentation.

Time period

The time period covered by the Spark and Vodafone (now One NZ) data sets initially appeared to be different. A closer examination revealed that the Spark data was showing UTC (Coordinated Universal Time) while the Vodafone data was 12 hrs ahead suggesting it was in New Zealand Standard Time (NZST). By adjusting both datasets to NZST using the lubridate package, the time periods covered for both datasets aligned to the same 14-day period.

Missing data

Preliminary checks of the Spark and Vodafone data sets showed a significant amount of missing device connections data that would need to be addressed. Further exploration of this data found that missing data fell into one of two categories:

1. Data is missing for all time points within an specific sa2_code
2. A single data point is missing within an sa2_code (always at the same time point (2024-06-13 06:00:00))

To determine the best way to deal with these missing data points the sa2_codes effected were examined. In the case of 1. (an sa2_code missing all data), it was found that these areas were all in the Rotorua district. For the second case (single data point missing), it was found that the effected regions were all oceanic sa2_codes that represent areas at sea. Since Rotorua was not an area of interest for this project, and oceanic regions are not relevant to roadwork considerations, then it was decided that all sa2_codes that contained missing data could be deleted.

Duplicate data

Another major factor to explore in the data was the possibility of replicate data. In both datasets we expect that each combination of datetime and sa2_code should be unique, meaning that there is only one data point for every sa2_code at each time point. However, when this was investigated it was found that many replicate rows existed. Closer analysis of these rows showed a clear pattern.

1. Replicate rows were from only three distinct sa2_codes (Twizel, Takapuna Central, and Hamilton Lake North)
2. All replicates appeared 8 times
3. Across the 8 replicates 4 rows had very low device counts and 4 rows had 'normal' counts

Examination of these replicated rows suggested that the row with the highest device counts was most probably the 'correct' counts and that other rows containing the same data, or very low counts, could be deleted. This was addressed by grouping replicate data, selecting the row with the maximum counts, deleting other rows in the group and then ungrouping.

Merging data

Spark and Vodafone datasets were merged on datetime and sa2_code using a full join to include all values from both datasets. The full join was chosen because of the known extra rows in the Spark dataset and the need to address these later (see imputation section below).

The combined device data was also joined to a dataframe derived from sa2_ta_concord_2023.csv so that columns for sa2_name, ta_name, and ta_code could be added. This was achieved by reading select columns from the sa2_ta_concord_2023.csv file, ensuring that columns were in the correct format (string vs. int), and then performing a left join so additional rows were not added to the device data.

Imputation

Prior to merging the Spark and Vodafone datasets it was noticed that the Spark dataset contained an additional rows. This was investigated by comparing the two datasets and selecting combinations of datetime and sa2_code that did not appear in both datasets. From this analysis it was found that the Vodafone data for Wellington Central was incomplete as records stopped before the end of the 14 day time period. This was the only difference found between data sets.

Wellington Central is a key sa2 area in this project as it is part of Wellington CBD. If we were to only take the Spark data for this location we would underestimate the number of people, and if we delete this entire area we lose data highly relevant to the project goals. Therefore we decided to impute the missing values in the Vodafone data set.

To perform the imputation we fitted a linear model to predict Vodafone data based on the Spark data. Using this model we then added the predicted values to the Vodaphone dataset to form a complete data set.

Calculate total device connections

A column for total devices was added by taking the sum of the Spark and Vodaphone connections at each time point. The number of connections in this column will be relied on in further analysis of the data set.

2. Modelling People Counts From Device Connections

Once raw data had been processed the most important decision to make was how 'device counts' could be used to model 'people counts' in each area. The only information regarding the number of people in each area we have access to is the population estimates obtained from subnational_pop_ests.csv. This section will describe how this data was tied together in order to model people counts in each sa2 area.

Process population estimates

The population document supplied appeared to contain multiple series of information. This included population estimates for 0-14 yr olds, an estimate of the full population labelled 'NZTA', and additional population estimates for area codes that were not SA2 codes, TA codes or UR codes. For the model we had chosen (see below) it was only necessary to extract a total population estimate for each distinct SA2 area. This was achieved using code shown in *'02_model_people_from_devices'* by reading select columns and filtering for the desired data. The sum of these population estimates was approx. 5.2M suggesting that our interpretation of the data was correct and the sections we removed represented counts for alternate area classifications. With a clean set of population estimates for each sa2 code we could join this data to the combined_cell_data dataframe (from *'00_process_cell_data.R'*).

Calculating people counts

To estimate the number of people in each SA2 area, we assumed a fixed ratio of 'people' to 'connected devices'. This is calculated using the population estimates provided for each SA2 code. This method assumes that the average number of devices connected at any given time reflects the typical number of devices present when the "people count" equals the population estimate.

For instance, if Wellington Central has an average of 1,000 connected devices per hour and the population estimate for that area is 5,000 people, we infer that there are 5 people per connected device. This "people per device" ratio is then used to estimate the number of people in each SA2 area at any given time by multiplying the device count by the ratio specific to that area.

This model was chosen because while it is relatively simple, it minimizes assumptions about device usage behavior. More complex models, while potentially providing better insights, would require us to make assumptions about factors such as:

- The number of devices an individual might own.
- The percentage of people who may not carry or connect a device at all.
- The likelihood of individuals using their devices at different times of day.

Without additional behavioral data or individual device identifiers, introducing these factors would not necessarily increase the accuracy of the estimates and could introduce greater uncertainty. Therefore, a model based on average device counts and population estimates provides a straight forward means to estimate people in each sa2 code at each time point.

NOTE: Limitations of this model are discussed below in section 5.

Clean dataset

With estimates of people count now calculated for each time point we could filter the data to provide the clean dataset for the client using the code in *'04_clean_deliverable.R'*.

3. Defining CBD Boundaries

The client has asked for insights specifically relating to performing road works in the central business districts (CBDs) of Auckland, Wellington, and Christchurch. In order to provide this information we must collate data relating to these areas and must therefore define the geographic boundaries of each CBD.

The data provided includes information about statistical areas, which territorial authority they belong to, and a classification of the area using an urban-rural indicator. The relevant information from this documentation was combined and summarised using *'01_process_area_data.R'*. None of the classifications in this documentation defines a 'CBD'. Additional research indicated that Stats NZ do not have an official CBD definition for reporting purposes. While the urban rural indicator can distinguish between locations that are more developed than others, it does not have a classification that is appropriate to define a CBD. 'Major Urban Settlement' is the most relevant classification, but at this level virtually all suburbs of major cities are included. Therefore a more precise definition of CBD is required.

While formal definitions may be lacking, each city has a generally accepted area that is referred to as the CBD. For example, in Auckland it is often defined as being bounded by several major motorways and by the harbour coastline in the north.Both Christchurch and Wellington can also be defined in a similar manner - Wellington by 'Wellington Central' bounded by the harbour, the motorway and the end of the 'golden mile'. And Christchurch by the area encompassed by the four avenues (Bealey Ave, Moorhouse Ave, Fitzgerald Ave and Deans Ave). In this project we have selected the sa2_areas that make up these geographic regions and used data from these regions to model people within the CBD.

The following figures represent the geographic areas we have chosen to define as CBDs, and the subsequent table lists the sa2_codes that make up these areas. The script *'03_cbd_data.R'* filters the existing dataframe to give data relating to the CBDs of relevance to this project. A column containing the NZ Stats population estimates was added in the resulting data when the code is run to provide a 'baseline' when visualisations are implemented.

NOTE: Geographic considerations for the CBDs are discussed in the stakeholder report and limitations are mentioned in section 5 below.

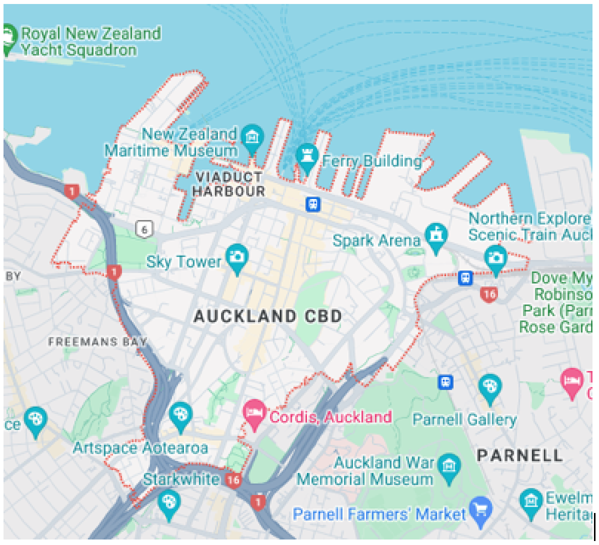

Figure 1: Auckland CBD Boundaries

Table 1, Areas codes and names for Auckland CBD

Table 2, Areas codes and names for Christchurch CBD

Table 3, Areas codes and names for Wellington CBD

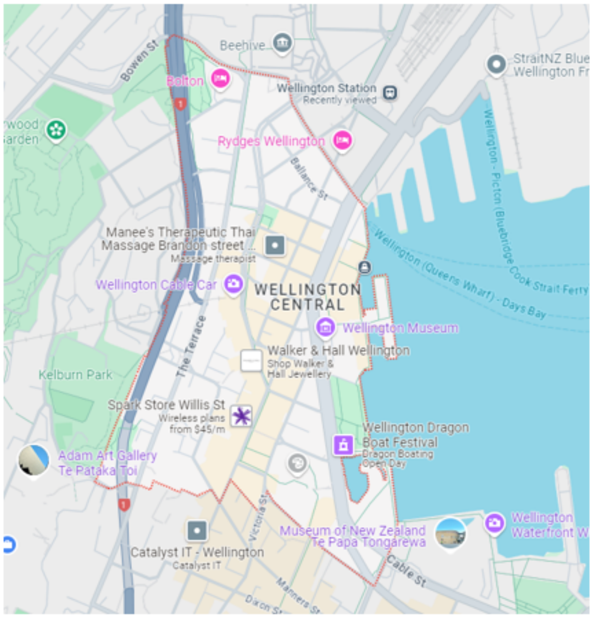

Figure 2: Wellington CBD Boundaries

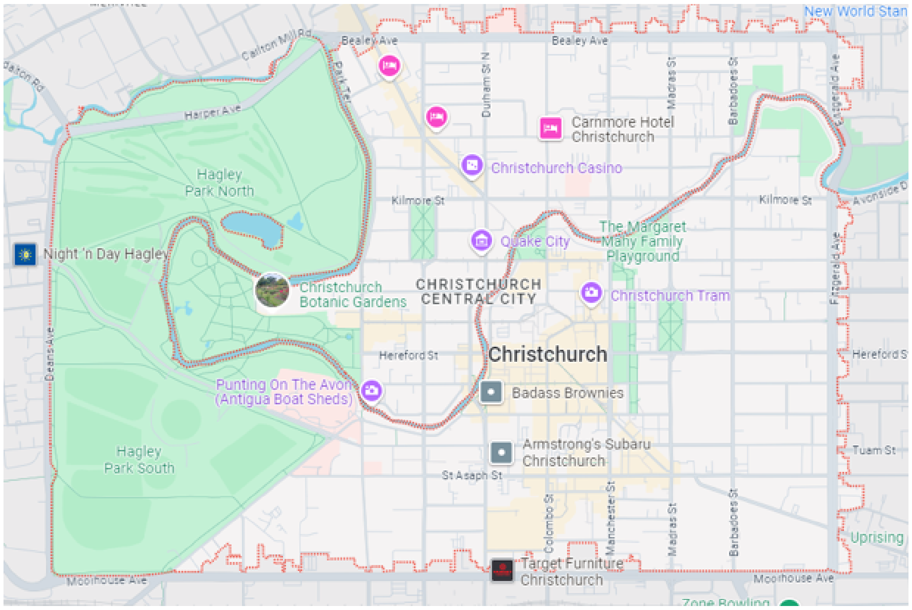

Figure 3: Christchurch CBD Boundaries

Table 1: SA2 Names and Codes for Auckland CBD

No	Auckland SA2 Name	Auckland SA2 Code
1	Anzac Avenue	134500
2	Auckland-University	134800
3	Hobson Ridge Central	133400
4	Hobson Ridge North	132700
5	Hobson Ridge South	133800
6	Karangahape East	134302
7	Quay Street-Customs Street	133301
8	Queen Street	133200
9	Shortland Street	133700
10	Symonds Street East	135900
11	Symonds Street North West	135100
12	Symonds Street West	135300
13	The Strand	135700
14	Victoria Park	132400
15	Wynyard-Viaduct	131300

Table 2: SA2 Names and Codes for Christchurch CBD

No	Christchurch SA2 Name	Christchurch SA2 Code
1	Christchurch Central	326600
2	Christchurch Central-East	327000
3	Christchurch Central-North	325800
4	Christchurch Central-South	327100
5	Christchurch Central-West	325700
6	Hagley Park	324900

Table 3: SA2 Names and Codes for Wellington CBD

No	Wellington SA2 Name	Wellington SA2 Code
1	Wellington Central	251400

4. Analysis and Visualisations

Having processed and calculated the required data and defined the geographic areas of interest the final step of this porject was to analyse the data and create visualisation that would allow use to provide insights to the client. The goal of these visualisations was to compare the population count during a school holiday week vs. a non-holiday week across three different CBDs of interest, namely Auckland, Wellington, and Christchurch. This grpahics illustrate how the estimated number of people changes throughout the week (Monday to Sunday) for each area.

Base code generation and flexibility

The core of the visualisation was designed around a flexible function (`comparison_plot`) that generates a comparison between holiday and non-holiday weeks for any given CBD dataset. This function can be applied to any subset of data through the use of `lapply`, allowing for the automatic generation of plots for each CBD area (Auckland, Wellington, Christchurch) without manually writing code for each individual plot.

By splitting the dataset (`cbdall`) by the territorial authority code (`ta_code`), the `lapply` function efficiently applies the `comparison_plot` to each subset of data. This approach makes the code scalable and adaptable to any number of CBDs. If additional CBD data is included, the same base code can be reused without modification, making it a highly flexible and efficient solution for visualising data across multiple regions.

This method ensures that the same logic and structure are applied consistently across all areas, reducing redundancy and enhancing maintainability.

Line plot to show trends over time for both periods

The main visual element is a line plot that shows the population count (y-axis) against time (x-axis). The x-axis is broken into segments representing the hours of the week (Monday to Sunday). Line plots makes it easy to see fluctuations in population throughout the week. By plotting both holiday and non-holiday data on the same graph, it's simple to make a direct comparison between the two.

Colour differentiation for Holiday and Non-holiday weeks

Two contrast colours are used for the different week types: red for holiday weeks and blue for non-holiday weeks. A dashed black line represents the population estimate given by Stat NZ.

X-axis labelled with days of the week

The labels are shifted to represent the corresponding day of the week from Monday to Sunday.

Alpha transparency to show focus on holiday week A slight transparency is applied to the non-holiday line (alpha = 0.5), while the holiday line remains fully opaque (alpha = 1). This subtle transparency ensures that the holiday data, which might be of primary interest, stands out more clearly.

Vertical grid lines to show day transitions Vertical grid lines are drawn at 24-hour intervals. These grid lines help the viewer align the data with specific days of the week and reinforce the idea that the data represents continuous time over a weekly period. This improves readability and makes the transitions between days visually clear.

Population estimate as a reference line A dashed horizontal line represents the population estimate for the area, serving as a reference point. Adding the population estimate as a reference allows the viewer to easily compare the actual population count to an estimated baseline. This helps contextualise whether the recorded counts are above or below what might be expected.

5. Limitations, Considerations, and Future Developments

This section outlines some of the limitations and considerations that were made in each section of the project, highlights some of the implications this may have had on results, and suggests possible developments should further work on this project be requested by the client.

The dataset

Access to Additional Telecommunication Providers: Access to device connection data used in this report were from two major providers (Spark and formerly Vodafone, now OneNZ) only. Data from other major providers (e.g. 2 Degrees) was not included and could have influenced the estimation of people counts and therefore the conclusions drawn.

Time Period Covered by Data: The data provided covered a two week time period, one week of school holiday and one wee in the lead up. If we are trying to make a general conclusion of whether road works should be performed in holidays or or term time it would be preferable to have data for additional time periods, both holidays and non-holidays. Since only one week of each condition was provided it is possible that other factors confounded this data. For example, bad weather may have reduced the number of people visiting the CBD during the holiday week when in a normal holiday there would be an increase in people. Similarly a large event such as a concert, show, protest, or graduation ceremony may have been held during one of the time periods and artificially increased people counts in the CBD. By examining more data from other holiday periods and other years we could develop a more reliable prediction of people movements.

No Individual Device Data: Since data was provided as a sum of total device connections for each provider we are unable to identify individual devices and track people movements more accurately. This identifying information could have facilitated the use of a more complex model to more closely monitor the number of devices per person, the specific locations they were travelling between, and other behavioral insight that are not currently accounted for due to the summarised nature of the dataset. However, this data may not be able to be utilised given the obvious privacy concerns.

The model

Perhaps one of the largest considerations in this project is the decision of how to model people counts from devices, which has a major impact on the resulting data and subsequent interpretation. As described in section 2 we chose to use a simple ratio based model to avoid making too many assumptions about the dataset. However, there are still several important considerations when taking this approach.

Device Usage Patterns: People may not always have their devices connected, particularly during nighttime when device usage is lower. As a result, the chosen model may underestimate the number of people present at certain times. However, this is not a significant issue if the goal is to estimate people counts for use cases like roadwork planning, which are likely more relevant during higher device usage times.

Varying Ratios Across Different Areas: The ratio of people to devices may differ between areas based on factors like urbanisation and the nature of the location. For example, a central business district (CBD) may have a higher number of connected devices relative to its residential population compared to suburban areas. This could lead to underestimating the number of people in CBDs. Since the analysis does not compare different types of areas directly, this limitation should have minimal impact on the conclusions drawn.

The CBD definition

There are several significant differences between the CBDs as we have defined them.

There are notable differences between the CBDs of Auckland, Wellington, and Christchurch as we have defined them for this analysis. These differences could influence traffic patterns and the ability to predict optimal times for roadworks.

Size: Auckland and Christchurch CBDs are significantly larger than Wellington's, affecting how traffic and population density might impact roadworks planning.

Through Traffic: Auckland and Wellington CBDs include major motorways that serve as critical routes to and from the city, as well as towards transport hubs like airports and ferries. These motorways may show higher traffic volumes that aren't representative of local CBD movement, making it essential to distinguish between motorway-related roadworks and those on smaller roads. In contrast, Christchurch's major traffic routes bypass the CBD, reducing the influence of motorway traffic on CBD data.

Parking and Public Transport: Wellington's compact CBD has limited parking and a heavy reliance on public transport (trains, buses, bikes, etc.), meaning that fewer cars on the road could lead to less impact from roadworks. In comparison, Christchurch and Auckland have more sprawling CBDs, where people are more likely to drive, resulting in greater road traffic and a higher impact from roadworks.

University Presence: Auckland's CBD includes a major university campus, while Wellington and Christchurch do not. School holidays overlapping with university holidays could significantly reduce the population in Auckland's CBD, potentially confounding traffic predictions.

Hagley Park (Christchurch): Hagley Park, consisting largely of parkland without roads, includes facilities like a hospital and sports areas. We have included it in this analysis as these faculties can attract significant numbers of people who most probably used vehicles to enter the CBD.

Future developments

The sections above have highlighted some limitations of the approach that was taken in this project and how this may have influenced the insights and analysis provided to the client. Some of these points, such as the choice of model or the scope of the telecommunications data used, could be altered and developed in future work to refine the estimation of people and produce more accurate insights. However, it is also worth considering larger changes to the project should the client wish to pursue this project on a larger scale or translate the insights to different regions. These larger changes may include reconsidering the type of data used for the analysis. While telecommunications data and device connections are a suitable way to monitor the total number of people in a specific area there may be better options when looking at roadworks specifically. For example, many people in a CBD probably arrive via public transport and then travel by foot, whereas roadworks are most likely concerned with traffic levels. Repeating the project with a dataset relating specifically related to traffic, perhaps records from parking buildings or vibration data from earthquake centers could be used to estimate the traffic in a specific region over holiday vs. non-holiday time periods.