Newcastle to Central Regional Population Projection: a Case Study

R. Ahmed, N. Iyer, S. Segal, J. Wang, R. Zhang

On behalf of Transport for New South Wales, statutory authority of the New South Wales government on matters relating to transport services

1 Project Background

In planning for the future transport needs of NSW, Transport for NSW needs to consider the impacts that transformative projects will have on communities and places. Currently, the NSW Government is working on developing a fast rail network, linking regional centres to each other and to Sydney. Four potential routes have been identified, including a Sydney to Newcastle corridor.

In order to estimate the potential benefits that a transformative project may bring, one of the responsibilities of Transport for NSW is to quantify the land use and development changes (including population and employment) that could be triggered by a transformative transport project such as Fast Rail. This is important because supporting services and infrastructure investment needs to be planned to anticipate growth of people and jobs at a local level. Land use changes may also not be constrained to the immediate walking catchment of the Fast Rail station and can extend further away along high frequency bus routes to areas of higher amenity (e.g. beaches, major retail centres, other transport hubs).

Transport for NSW requires the development of a methodology and model to calculate the potential change in population that may occur as a result of a transformative project such as Fast Rail for the Sydney to Newcastle corridor. The methodology and model should be guided by best practice from around the world in how population change is calculated for major infrastructure decisions.

The primary goal of this project is to develop a model to calculate the projected changes in population along the Sydney to Newcastle corridor as a direct result of the Fast Rail project. In addition, as specified by Transport for NSW in meetings, it is preferable that the model is able to spatially distribute the increases in population such that it is possible to see the differences in population growths depending on which region is being observed.

The problem statement is therefore as follows:

> Develop a methodology and a model that predicts the increases and/or decreases in population within smaller regions along the Sydney CBD to Newcastle corridor as a result (both direct and indirect) of the transformative Fast Rail project.

In solving this problem, the team was advised of the following:

1. Population projections from 2016 to 2056 will be provided, and this is to be used as a base case in the model. Note that provided projected population assumes that the Fast Rail project does NOT exist.

2. Proposed Fast Rail travel times once the transformative project is completed will be provided. The Fast Rail will change the frequency and speed of the trains servicing the corridor of interest; the details of these changes are fully provided by Transport for NSW.

2 Literature Review

This section gives a brief overview of existing knowledge regarding population projections and city modelling that will be used to develop a methodology to answer the problem.

A solution to the problem requires building a model from the base case, assuming that once the Fast Rail project is completed, the projection in population growth will change. The key question is how and why this change occurs. Once it is known the mechanism influencing population growths of a region, as well as the degree of influence, then it will be possible to assimilate those mechanisms into a single model describing this growth.

As such, the following key questions must be answered to allow a model to be developed:

1. When choosing a location to live on a permanent basis, what is/are the key factor/s an individual will consider?

2. For each such factor that an individual is assessing, to what degree does each influence their choice? If the effect of a factor is insignificant, it can be ignored.

3. How many of those factors influencing individual residential location choice relate to train service? Do changes in the train service (i.e. the existence of a new Fast Rail system) affect these factors?

Considerations regarding permanent living conditions have been well analysed as technology and societal norms have changed over time. Population density is a major indicator of features such as the commercial spread of economic hubs and living conditions of the overall populus. Hence, governments and corporations alike have taken to modelling population density changes due to a wide range of stimuli.

It is common to determine urban densities by first establishing a variety of factors that affect the concerning variable, and to use a mixture of linear and logarithmic relations to develop the final empirical analysis. (Alperovich, 1983, pg 289) However, many common measures such as environmental integration or number of notable tourist landmarks are irrelevant to the development of a new HSR, and hence have been discarded. The one significant indicator aside from the operation of the transport infrastructure itself is the agglomeration of industry within each particular area. More accurately, this encompasses the tendency for individuals to work in the local area and not have to travel long distances for their occupation.

In relation to the addition of a high-speed rail (HSR), the main direct factor affecting population density have been shown to be the frequencies of the HSR train rather than actual time savings due to the HSR. (Chen, 2014). Indeed, this highlights the notion of a 'perceived' waiting time (represented by frequency) affecting the utility of the traveler far more than actual travel time. For the calculation of population density, it was determined that the frequency of buses would similarly affect the utility of the average traveler, with the caveat that it was assumed bus operations would not change due to the addition of a HSR.

In order to determine the size of the area in Sydney that would be affected by a HSR

along the specified route, the maximum amount of time an individual would reasonably spend travelling to work , or travel time frontier (TTF) was considered. (Banerjee, et al, 2007). Here, it was shown that different societies and cultures output very different values for this TTF value. Since the analysis was between two first-world countries and one third world city, the research was flawed, yet provided the necessary information to develop a limit the range from the stations which would be considered in our model.

To determine to where travel times were taken, it was noted how bid-rent economic theory argues urban patterns are strongly influenced by the distance to a Central Business District (CBD). In this theory, jobs are assumed to be concentrated in a single CBD with residential communities surrounding that CBD. (e.g., Alonso, 1964; Muth, 1969).

3 Methodology

3.1 Identifying Relevant Variables

To develop a model that can predict the population size of a given spatial region in the future, the model must be able to take in a set of inputs values, process them, and then create an output value. Since the model is required to predict population size, the output value must therefore be the population size of interest. However, due to the nature of how Transport for NSW identifies Travel Zones (TZ) or sections of interest, the team decided that using population density is a much better measure of population (for example, it is difficult to tell if one million people is a lot; it is a lot for a city like Sydney, but not for a country like Australia as a whole).

The input variable items, however, are not as straight forward as the output variable. Whereas the model's output variable must be the population density in order to answer the problem, the input values chosen must be the factors that will influence population growth to a significant degree. For example, should the team believe that people care very much about having good access to public transport in the region they live, then access to public transport will be an input variable considered in the model in order to predict the future population in a region.

This example can be extended to a variety of other potential factors; individuals may consider the proximity to amenities such as parks, shopping centers etc. as important reasons to choose to live in an area, or that safety and health such as the presence of hospitals, police and fire services is of high importance to their residential location selection. The team had to identify what we believed to be the factors that most accurately predicted the choices of as many people as possible in order to model regional population growths in future years.

According to the review of literature as per section 2, the team has identified the following factors as being relevant inputs for this problem:

1. The number of job opportunities available within a region of consideration, termed 'j' for the remainder of this report.
2. The travel time to the nearest city CBD from the region of consideration, termed 'T' (capitalised) for the remainder of this report.
3. The frequency with which one can expect a public transport service within the region of consideration.

Now, after identifying the key factors, they must be expressed as variables that can be entered as inputs, which require each to be expressed mathematically as a number. The team decided to break the 3 factors into the following 5 variables as inputs into the model:

1. j is interpreted as number of job opportunities, and is easily quantifiable.

2, 3. T is broken down into T_{bus} and T_{train}, or the travel time an individual must experience on their journey on bus and train respectively. Since the Fast Rail is to do with upgrading the train infrastructure, this breakdown allows the model to improve train times while keeping bus times constant.

4, 5. t_1 and t_2 are introduced as headway variables, where t_1 is the headway of buses within a given region and t_2 is the same for trains. These two are a representation of how frequent buses and trains are. Note that high headway indicates more waiting time on average for customers and low headway indicates the opposite. Since more frequent buses and trains equates to less waiting time for customers, then this means there is a generally correlative relationship between frequency and headway.

3.2 Model Selection

As mentioned above, the model must be able to accept the above mentioned inputs and process them into an output. After looking at potential ways of doing this, the following possible statistical models were considered as candidates:

1. Simple linear regression model
2. Generalised linear regression model
3. Log-linear regression model
4. Support-vector machine model
5. Logistic regression model

The set of inputs and outputs are the same in each type of model; the differences lie in the method which the model treats the inputs with to obtain the output. The merits and drawbacks of each model is outlined below:

3.2.1 Simple linear regression model

Simplest model type available. The relationship between the output and the inputs is expected to be a linear one, meaning that the effect each input has on the output is constant (contrast with log-linear model below).

A few assumptions must be made for this model to be applicable. These assumptions are:

- The variance in population density of each region being inspected must be the same in all regions.
- The distribution of the error of the inputs relating to the output must be a normal distribution.
- The relationship between the input and output is indeed linear; if the relationship in reality is not linear, then the model will produce an inaccurate set of population density predictions.

Although there are a lot of assumptions made, they tend to be reasonable assumptions when discussing behaviours and tendencies of humans.

This type of model is very useful when a simple relationship is needed, computational resources are not abundant, or when a quick and rough idea of the relationship between two things are needed.

3.2.2 Generalised linear regression model

This type of model is very similar to the simple linear regression model. The key difference is that the generalised version of the model does not require the assumption of the error of the inputs to be normally distributed; the errors can have any distribution.

Since this type of model is a generalised version of the above, it has the same assumptions (except for the normality of the error).

This model is useful when the normality of the error of inputs cannot be assumed (i.e. simple linear regression is not valid).

3.2.3 Log-linear regression model

This type is also a type of linear regression model similar to the simple and generalised linear regression models. This one, however, looks at the logarithmic response of the output instead of the linear response.

This means that no matter the relationship between the output and the inputs, the output value will always be a positive number. This makes the log-linear model useful when it is known that the output value cannot be a negative (for example, there cannot be a negative number of people; population density therefore must be an output with a positive value).

3.2.4 Support-vector machine model

The support-vector machine model (SVM) is a completely new approach the team had considered in this report (the other models are all part of the family of linear regression models).

SVM is a machine-learning algorithm based method of developing a model. This means that existing data is needed to feed into an algorithm, which is programmed to empirically associate the required output with the defined input variables. After sufficient learning data has been processed, SVM will create a relationship based on machine prediction, unlike the other models that simply fit a model to the data with significant errors.

SVM is appropriate when there is large amounts of data and computational resources, and detail is the most important thing needed in the model.

3.2.5 Logistic regression model

Also a type of regression model, but unlike linear regression, a logistic regression model's output must have discrete options; that is to say, a logistic regression model's output must be value such as on/off, soft/hard, large/small etc. and cannot be generic numbers. Furthermore, logistic regression requires the output's choices to be binary (i.e. 2 choices only; red/yellow is acceptable, yellow/blue is acceptable, red/yellow/blue is not acceptable)

This type of model is useful when the goal is to measure how choices are made by indi-

viduals (human beings choosing pizza toppings, self-driving cars choosing a route, birds choosing sticks for their nest etc.).

3.2.6 Selecting the model type

The model selection process involved choosing which of the above 5 potential model types the team would use to develop the actual population prediction model.

Model 4 (support-vector machine) was very quickly disqualified as a possible model skeleton for the project. The non-trivial training time, large training dataset requirement, and foreseeable significant computational resource consumption during development and operation of the model meant that developing the model was impractical due to the time constraint, and operating the model will be much more complex than necessary.

Out of the remaining four model choices, model 5 (logistic regression) requires the outputs to be discrete. In the context of this project, as mentioned before, the output must be the projected population density in 2056. Population density is neither discrete nor binary; as such, the format of the output is incompatible with logistic regression.

The remaining three model choices are extremely similar to each other; all three operate under the same principle, except with differences pertaining to the assumptions made and the format of the output slightly. With that being said, the team initially considered using model 3 (log-linear) due to its output being purely positive (since population density can never be negative), but rejected it after realising that logarithmic relationships are harder to model and interpret results for.

From the remaining two models (simple linear and generalised linear), the team decided on using the generalised linear model approach. Its more generalised nature meant that less assumptions had to be applied to use it than the simple linear model. Additionally, the computational requirements and operational requirements of such a model were predicted to be low. It therefore suits the purpose of this project the best out of the five model choices.

3.3 Model Development

With the output, inputs, and model type all known, the next step is to develop the model itself. A generalised linear regression model is made by using existing data about the outputs and inputs to form a relationship that can be used to predict future values of the output given a different set of inputs. This means that despite knowing the outputs, inputs, and model type, a source of data relating them together is still needed.

In the context of this project, existing information about population sizes and each of the selected input variables listed above (t_1, t_2, T_{bus}, T_{train}, and j) must be known. Crucially, this information only needs to be about what currently exists along the Central to Newcastle corridor, and not about any future projections in future years. The source of this data can be selected from census data, open-source transport data etc. For this particular project, Transport for NSW has provided the necessary existing data to develop the model, but the data provided is too detailed to be directly put into the model; instead, the team had to aggregate and extract the required data from the provided sources into

a usable form.

The following outlines the data extraction procedure, and the development of the actual model in open-source software.

3.3.1 Range of Central to Newcastle corridor under consideration

While the aim of the team was to develop a general model that can be applied to any regional area just outside of a city CBD, the focus of this particular project is placed on the Sydney to Newcastle area. This large regional corridor contains significant portions of land not currently being used for any commercial or residential purpose and is also not likely to have any significant residential population in the coming decades (for example, in the Brisbane Water National Park west of Woy Woy).

As a result, the team decided it was inappropriate to consider the entirety of the land between Central station to Newcastle Station; instead, sections of the corridor that current have sizeable residential populations and/or the team expects to have population growths in the coming decades were selected for analysis, in order to ensure the model is as accurate as possible in depicting the conditions of regional NSW north of Sydney CBD.

After inspecting the corridor, the team decided that the areas immediately next to the train stations along the existing Central to Newcastle Main North railway line. For ease of identification, the team chose to focus on the SA2 regions containing the train stations from Central to Newcastle. SA2 is a method of categorising a geographical area used by the Australian Bureau of Statistics (ABS) for ease of regional identification. These regions are sizeable enough that most of the regional corridor that has residential potential is captured, but not so large that land not usable for residential purposes (e.g. national parks) are not included. The team also noted that Travel Zones (TZ) are already mapped to SA2 regions, meaning that Transport for NSW can easily identify which TZ lies in each SA2 the team has considered.

3.3.2 Obtaining Initial Data and Cleaning

The intention of this step is to obtain sufficient data to carry out the modelling process. The collected data was split into each individual SA2 being considered, and the population densities, t_1, t_2, T_{bus}, T_{train}, and j values for each found and documented.

Transport for NSW has informed the team that the required datasets involving existing population numbers can be taken directly from the Open Data hub set up by the organisation. This provides sufficient data about the population sizes along the Central to Newcastle corridor. These are then processed with data about the sizes of each SA2 (obtained from ABS) to find the population densities of each region.

For the variable j, the team obtained information about the number of existing employed persons along the Central to Newcastle corridor, and decided to use that value of employed persons as a substitute instead of the number of jobs in each region being looked at. The primary reason for this is due to the difficulty in estimating the exact number of jobs that may exist compared to a simple survey of total employed persons at any given time; as well as that, assuming there will generally be persons capable and willing to

fill any vacant jobs, then the number of jobs in a given region will be roughly equal to the total number of employed persons in that region, ignoring unemployment. The issue with unemployment is discussed in the end of this report under the 'Assumptions' section.

The remaining variables t_1, t_2, T_{bus}, T_{train} are easier to obtain data for than j. Since these four variables all relate to the travel times and frequencies of the buses and trains, the relevant bus and train timetables for each SA2 region considered was taken from the service company's website and analysed to obtain the needed times.

In considering the train travel time T_{train} and train headway t_2, since the only train line considered is the existing Newcastle - Central line, T_{train} is the average travel time between a given station relevant to each SA2 and Central (details in 'Assumptions') in both directions, and t_2 is the average headway of trains servicing the section between the SA2 region's station and Central.

For t_1 and T_{bus}, an average of each value for every SA2 region was taken and used. The trouble with these two variables, however, was that each SA2 region had a different number of bus routes servicing it, each with different headways and travel times, meaning that it was very difficult to identify the amount of travel and waiting time a random person living in the region would experience. As such, the team decided to select, for each SA2, the single bus route that serviced more of the region than any other service and used the t_1, T_{bus} values of that service as the data for that SA2. The assumption was that the highest coverage bus route in each SA2 is most representative of the experiences of a random person living in the area (details in 'Assumptions').

3.3.3 Development using R

The actual model was developed using the R language, in the RStudios development environment. Since the model being developed was a generalised linear model, built-in functions already exist to quickly create the model. Since the team already had data cleaned and ready to use, the development procedure was very quick, and took very little time.

However, an issue came up during development; while the team found that the resultant model was able to predict future residential populations along the corridor, it was not able to find how the Fast Rail project would affect the changes in residential population; the other variables had a much greater effect that invalidated the effect of the project. The team decided to tweak the inputs, and came up with several iterations of the model, all of which can be found in Appendix A. The most satisfactory of the models was then selected as the final model to be used to perform the remaining case study of population densities along the corridor by 2056. This final iteration of the model ignored the j variable entirely, but kept all others.

Details of the best model selection process is also documented in Appendix A.

4 Results

4.1 Results of the Modelling Process

The model with the best fit is outline below:

	Coefficient	p-value
(Intercept)	794.666	0.4723
t_1	-33.885	0.2205
t_2	-35.707	0.2219
T_{train}	24.624	0.0118
T_{bus}	-75.412	0.0582

In the table above, the 'Coefficient' column indicates the effect each of the row headers have on population density (for example, an increase in 1 unit of t_1 is expected to decrease population density in a region by 33.885 units). '(Intercept)' refers to how many people would live in a region even if there was no buses or trains at all.

The p-value column indicates the likelihood of the significance of each of the predictors, or how confident the model is that each of the given inputs actually affect population. A predictor's percentage significance is one minus the p-value (for example, t_1 has a $1 - 0.4723 = 0.5277 = 52.77\%$ chance of being significant). Target p-value for each input should be equal to or less than 0.05 (95% confidence). Given the low amount of data the team had to work with, the team decided that a p-value of 0.1 (90% significance) was also acceptable.

As a result, the team concluded that only the travel times by bus and train have a significant effect currently on the population densities on a region. The headways are not statistically significant.

4.2 Future Population Projections

The following table is the results of using the model to predict populations along the corridor by 2056:

SA2	Region Name (ABS)	Pop Density 2056 (persons/km^2)	Area (km^2)	Pop Count 2056 (persons)	Existing Population (persons)	Change (persons)
111031229	Newcastle - Cooks Hill	2743.92	3.9804	10921.89917	22379.61	-11457.71
111021219	Toronto - Awaba	653.1216	43.6626	28516.98717	16497.40026	12019.59
111011209	Glendale - Cardiff - Hillsborough	2292.4864	21.9175	50245.57067	29139.84325	21105.73
111011214	Warners Bay - Boolaroo	1838.322	12.5402	23052.92554	16445.27	6607.656
102021056	Warnervale - Wadalba	1871.17376	42.89	80254.64257	59908.12173	20346.52
102021057	Wyong	1848.08976	15.0334	27783.0726	15586.03227	12197.04
102021051	Ourimbah - Fountaindale	1608.61936	114.1172	183571.1372	6353.027854	177218.1
102021055	Tuggerah - Kangy Angy	973.80464	27.8463	27116.85615	11833.56924	15283.29
102011035	Narara	1669.38208	7.7021	12857.74772	9052.23	3805.518
102011032	Gosford - Springfield	830.278	16.9124	14041.99365	33102.73	-19060.74
102011042	Woy Woy - Blackwall	534.79	17.4225	9317.378775	18808.37	-9490.991
102011040	Umina - Booker Bay - Patonga	534.79	25.2284	13491.89604	30086.65946	-16594.76
102011036	Niagara Park - Lisarow	1637.819768	16.7316	27403.34523	10459.96	16943.39
111031224	Hamilton - Broadmeadow	2512.6406	6.7462	16950.77602	16364.65122	586.1248
111021215	Bolton Point - Teralba	1600.8488	21.9697	35170.16788	10551.78	24618.39
111031222	Adamstown - Kotara	2256.462085	8.0182	18092.76429	20331.74	-2238.976

The team observed a general tendency for population to increase, except in certain regions closer to larger train stations; the more regional/disconnected areas generally received an

increase in projected population.

Now, for the populations in the regions in 2016, just after the construction of the Fast Rail project, it is unknown the exact amount of population in each region. It is known that the end result will be the 2056 population table, but in 2016, only a fraction of the population will have migrated over. This unknown migration rate means that it is not possible to predict 2016 population using the model.

Instead, the team decided on a different approach: assuming that population migrates sufficiently fast, then the ratio of population in 2016 to population in 2056 will be the same in each region, both with and without the Fast Rail project. The additional population from between 2016 to 2056 are assumed to be a result of natural population growth (not migration), and therefore are discounted. Therefore, as a result, the following table is obtained:

		no Fast Rail	no Fast Rail		with Fast Rail	with Fast Rail
SA2	Region Name (ABS)	Pop Count 2016 (persons)	Pop Count 2056 (persons)	Pop Ratio	Pop Count 2056 (persons)	Pop Count 2016 (persons)
111031229	Newcastle - Cooks Hill	11435.14	22379.61	0.5110	10921.89917	5580.680185
111021219	Toronto - Awaba	13765.33573	16497.40026	0.8344	28516.98717	23794.40979
111011209	Glendale - Cardiff - Hillsborough	23987.33788	29139.84325	0.8232	50245.57067	41361.15183
111011214	Warners Bay - Boolaroo	13636.05158	16445.27	0.8292	23052.92554	19114.97238
102021056	Warnervale - Wadalba	15637.32798	59908.12173	0.2610	80254.64257	20948.21423
102021057	Wyong	9014.703919	15586.03227	0.5784	27783.0726	16069.27081
102021051	Ourimbah - Fountaindale	5014.340154	6353.027854	0.7893	183571.1372	144889.6724
102021055	Tuggerah - Kangy Angy	5441.789787	11833.56924	0.4599	27116.85615	12469.9681
102011035	Narara	6848.31	9052.23	0.7565	12857.74772	9727.309435
102011032	Gosford - Springfield	19380.78769	33102.73	0.5855	14041.99365	8221.222165
102011042	Woy Woy - Blackwall	14216.61	18808.37	0.7559	9317.378775	7042.691114
102011040	Umina - Booker Bay - Patonga	24234.60349	30086.65946	0.8055	13491.89604	10867.63226
102011036	Niagara Park - Lisarow	8201.18	10459.96	0.7841	27403.34523	21485.71953
111031224	Hamilton - Broadmeadow	12295.0304	16364.65122	0.7513	16950.77602	12735.39556
111021215	Bolton Point - Teralba	8721.99	10551.78	0.8266	35170.16788	29071.28964
111031222	Adamstown - Kotara	16022.7	20331.74	0.7881	18092.76429	14258.24521

5 Discussion of Results

5.1 General Discussion

The first thing to note is that the outputs show detail at the SA2 level, not Travel Zone level. This is because due to a lack of sufficient data, the team could not reduce detail of the model down to the Travel Zone level, only at the SA2 level. However, since each SA2 is made up of individual Travel Zones already, the average Travel Zone will have the features of its corresponding SA2 region. As a result, the above population prediction table can be extrapolated to the Travel Zone level.

The model results seem to suggest that the only factors contributing to predicting population growth in a region will be the travel time an individual will experience on bus and train. The headway of the services (and by extension frequency) are not nearly as effective or important (due to their high p-values). This seemingly counter-intuitive output may be a result of the inherent model instability that the model experiences due to the limited data points available, as well as a lack of consideration of other possible variables affecting population residential selection.

As for the predicted values of populations by 2056, by examining only the statistically significant variables T_{bus} and T_{train}, a general increase in the population of each region by 2056 can be observed as a result of the Fast Rail. The model indicates that an increase in bus travel time is associated with a decrease in population levels in a region, meaning that the Fast Rail improvements to travel time (by 20%) will increase the population levels of each region the Rail project affects. This is intuitively reasonable, as one can expect shorter travel times to and from work, leisure etc. and would take that into account when selecting a residential area.

However, strangely, the travel time by train seem to be positively correlated with population; that is, the higher the travel time by bus an individual expects to experience, the more likely they are to live in a region. This is completely opposite to expected wisdom; that people generally want to minimise their travel times. This is potentially the effect of model instability, as well as the fact that perhaps in the regions being examined by the model, the bus travel times are too short for people's likings (for example, individuals may prefer to have a longer bus travel time so they have more time to read on the bus).

The above point, however, must be taken with the additional condition that the association between bus travel time and population has a limit: according to literature, a total journey time of 176 minutes is the limit that an individual is willing to invest into daily travel as the absolute limit. Should train travel times increase to a point that the sum of T_{bus} and T_{train} exceed this daily limit, conceivably the population increase will reduce to 0. This may be a consequence of people deciding that the opportunity cost of spending time on public transport at this point is too much, and so stop considering a region as appropriate for residential purposes.

5.2 Assumptions

The following assumptions (and justifications) have been made in this project:

- The primary use of public transport will be for work purposes. Intuitively, the peak hour periods attract the most users of public transport, of which most users are using for work commuting. Significant amounts of weight, therefore, are placed on working populations during the analysis. Peak hours of a day are assumed to be between 6 and 10 am, and 3 and 7 pm.

- The train and bus timetables provided by Sydney Trains and relevant bus contractors are accurate.

- As per project brief, stations between Hornsby - Epping, Mt Kuring-gai - Asquith, Epping - Central were not considered in the analysis as they are serviced frequently by inner CBD trains.

- Housing is not a restriction on population increases. If this assumption is not made, then the model becomes needlessly complicated; it is reasonable to expect that the private sector will move to match demand in regional housing.

- The headway of the bus and train times are transformed into t_1 and t_2 by reducing each by 50%. In truth, the headways are not the deciding factor for residential selection; the waiting time of the customers are. Assuming that a random person in a region can potentially have to wait anywhere between no time to the headway time of a service, the average person is assumed to wait an average of these 2 times. The average waiting time is therefore obtained by taking an average of the headway.

- Data and numbers are considered at the SA2 level. Anecdotally, most people say things such as "I want to move to [suburb]", instead of defining a particular street in a suburb they are interested in. Practically, this means that having data at too detailed a level can mis-simulate the realities under which people make their choices. The team has decided that the SA2 statistical level is most appropriate as it is not too big to become too broad, but also not so small that it becomes too specific for people to consider.

- Although not ultimately used in this model, the effect of unemployment on j was not considered. While the Australian government publishes unemployment rates, these are ultimately a measure of employed persons, and NOT the availability of jobs. A mismatch between skill sets, desires etc. means that unemployed persons are not necessarily an indicator of available jobs (which is what j represents). To prevent this overcorrection on j, unemployment was therefore discounted.

- The bus route in each SA2 that services as much of the SA2 area as possible is taken to be representative of all bus routes in the region. t_1 and T_{bus} are taken from this bus route only. While this is clearly not the case for every person in the region, time constraints meant that the team needed a way to quickly extract a single representative value for headway and travel time in each SA2. The team reasoned that people tend to catch the bus closest to their house, rather than walk significant extra distance for a faster service. As a result, the bus that covers most of the region can be assumed to be the service and most people will take (since it will be "closer" to more houses for more time). In other words, a randomly chosen person in the region most likely takes the bus that covers most of the SA2 region.

- The population of each SA2 can be evenly distributed among its travel zones. This assumption is made out of necessity; a lack of detailed data at the travel zone level for employment forced the team to consider SA2 regions instead (one size larger than TZ). While employment was eventually not considered during the model development process, it does not resolve the issue of too little geographic spatial data at the travel zone level. If more detailed data could be provided in a future project (or if more time is given), travel zone-level detail could potentially be achieved by the model.

5.3 Limitations of the Model

The time constraints imposed by the deadline of the project meant that a very simplified model of population was made. As such, due to the numerous assumptions and technical limitations regarding computational resources that the team experienced, the following are a summarised list of limitations to the model:

- Very high p-values in the model's output. This model indicates that travel times are significant predictors of population size in a region, but it has difficulty in actually reliably predicting the exact number of people due to the high p-values of the other predictors.

- Model allows population size to become negative. Theoretically, if the travel time by bus T_{bus} was increased to thousands of minutes, then the model will determine that a negative number of people will end up living in a region. This has no real meaning, and, although an extreme edge case, means that the model is not a perfect candidate. Other, more technical models such as log-linear regression may be a better candidate instead.

- The model does not give indications to any other factors that may influence population that the team missed. For example, the number of/quality of facilities such as hospitals, schools etc. within a region may be of interest to potential residents. Such factors were not considered as part of the model.

- The team dismissed the j variable due to it skewing the analysis of the Fast Rail related variables. The j variable, however, had the greatest influence on population (see Appendix A). This does NOT mean, however, that the Fast Rail project has no potential effect on population; instead, it simply means the team did not investigate second-order effects that the Fast Rail project may have. For example, it is not possible at this stage to say whether or not the Fast Rail project will contribute to an increase in jobs in regional areas, which in turn affect the population size as a whole. More work must be done on this area in future analyses.

6 Conclusion

This report detailed the development of a population prediction model using public transport accessibility as a key predictor metric. A study of the Newcastle – Central corridor serviced by existing Sydney interregional trains using this model was also conducted in accordance with project brief. The study concluded that a Fast Rail transformative project along the corridor will serve to shift population sizes in the surrounding regions; a general increase was observed for most of the regions studied, with declines in population in Newcastle, Gosford, and Woy Woy regions.

The model developed as a result of this project can be used in future public transport projects; further improvements in buses and trains can be quantified in terms of the travel time experienced by customers and used to predict general changes in population sizes.

However, the model considers only public transportation as predictors of population size; second order effects such as employment, access to amenity resources such as sporting centers, schools, hospitals etc. were not considered. As such, the model is highly limited in scope and unsuitable for detailed analysis of exact numbers; only the rough general trend in increase/decrease should be used. A more detailed model can perhaps be developed that takes into account these additional factors in a future project.

The project team has no specific recommendations regarding the findings; the team does have general recommendations regarding upgrading infrastructure local for the benefit of the community.

7 References

Alonso, W. (1964). *Location and land use: toward a general theory of land rent*, Harvard University Press, Cambridge.

Alperovich, G. (1983). "Determinants Of Urban Population Density Functions," *Regional Science and Urban Economics*, Vol. 13 (2), pp 287 - 295.

Australian Bureau of Statistics 2019, *ABS Maps*, published by Australian Bureau of Statistics, accessed 14 Sep 2019.
<https://itt.abs.gov.au/itt/r.jsp?ABSMaps/>

Banerjee, A., Pendyala, R.M., Ye, X. (2006). "Understanding Travel Time Expenditures Around the World: Exploring the Notion of a Travel Time Frontier," *Transportation*, Vol. 34 (1), pp 51 - 65.

Chen, M. (2014). "Impacts of French High-Speed Rail Investment on Urban Agglomeration Economies" Publicly Accessible Penn Dissertations. 1234.

Muth, R.F. (1969) *Cities and housing: the spatial pattern of urban residential land use*, University of Chicago Press, Chicago.

Transport for New South Wales (2019), *Open Data Hub*, general open source data, published by Transport for New South Wales, accessed 11 Sep 2019.
<https://opendata.transport.nsw.gov.au/>

8 Appendix A

The following is a list of all models that the team generated. The format of each model is identical to the one as presented in 'Results'. The p-values for each item should ideally be less than or equal to 0.05. e indicates multiplying by a power of 10 (e.g. e03 means multiply by 10^3). Statistically significant p-values at the 90% level are affixed with a * superscript.

AIC is an indication of the quality of the model; whilst p-value indicates how significant a predictor might be, it gives no indication to how good the actual model is. In the case of the following models, the lowest AIC indicates the most ideal model.

Model 1:

	Coefficient	p-value
(Intercept)	730.21561	8.50e-05*
j	0.57588	5.63e-08*
AIC	245.55	

Model 2:

	Coefficient	p-value
(Intercept)	890.80849	0.127001
t_1	-12.11452	0.393806
t_2	-7.94266	0.598825
T_{train}	-28.60664	0.175207
T_{bus}	5.80802	0.291777
j	0.47832	0.000169*
AIC	248.86	

Model 3:

	Coefficient	p-value
(Intercept)	-536.568	0.6555
T_{train}	22.794	0.0377*
T_{bus}	-73.545	0.1132
AIC	275.67	

Model 4:

	Coefficient	p-value
(Intercept)	2957.51	0.00373 *
t_1	-47.56	0.12221
T_{bus}	-49.89	0.28874
AIC	278.14	

Model 5 (selected model):

	Coefficient	p-value
(Intercept)	794.666	0.4723
t_1	-33.885	0.2205
t_2	-35.707	0.2219
T_{train}	24.624	0.0118*
T_{bus}	-75.412	0.0582*
AIC	270.5	

Note that the models above with the lowest AICs all have j as the only significant predictor, the team decided to use model 5 (the model with the lowest AIC without j as a predictor).