Urban Transport Consultancy Stage 2

Commonwealth Grants Commission

Stage 2 - Final Report

IA174500 - Stage 2 Final Report | D

25 October 2018

Urban Transport Consultancy Stage 2

Project No:	IA147500
Document Title:	Stage 2 - Final Report
Document No.:	IA174500 - Stage 2 Final Report
Revision:	D
Date:	25 October 2018
Client Name:	Commonwealth Grants Commission
Client No:	
Project Manager:	Phillip Truong
Author:	Jacobs and Synergies Economic Consulting
File Name:	J:\IE\Projects\04_Eastern\IA147500\21 Deliverables\Stage 2\IA147500 - Stage 2 Final Report Rev D.docx

Jacobs Australia Pty Limited

Level 7, 177 Pacific Highway
North Sydney NSW 2060 Australia
PO Box 632 North Sydney
NSW 2059 Australia
T +61 2 9928 2100
F +61 2 9928 2444
www.jacobs.com

Document history and status

Revision	Date	Description	By	Review	Approved
A	05/08/2018	Preliminary Draft	Simon Sagerer	Paul McLeod	David Lowe
B	27/09/2018	Draft report	Simon Sagerer	Paul McLeod	David Lowe
C	17/10/2018	Final report	Simon Sagerer	Paul McLeod	David Lowe
D	25/10/2018	Final	Simon Sagerer	Paul McLeod	David Lowe

Contents

Executive Summary

Jacobs and Synergies Economic Consulting ('the Team') have been tasked with assisting the Commonwealth Grants Commission (CGC) to review its modelling of State urban transport recurrent and infrastructure expenditure requirements. The objective of the project is to develop an alternative model that could inform future allocation of GST funds to the States and Territories.

Analytical Framework

The relevant literature shows that expenditure on public transport provision will vary across cities based on the transport task to be undertaken, the characteristics of the transport system and of the specific characteristics of the city itself. A recurrent expenditure model has been developed in this study consistent with this proposition. Variables representing each of these factors have been assembled in a comprehensive data base covering 101 Significant Urban Areas (SUA) defined by the Australian Bureau of Statistics. A number of models have been estimated that are consistent with the basic proposition and tested to determine which recurrent expenditure model best explains the expenditure variations between these cities.

The easiest way to explain the modelling framework we have adopted is to illustrate the underlying concept by focusing on a single mode such as bus transport. The cost of operating a bus system would depend on volume (number of bus passenger kilometres) and input prices (e.g. prices for vehicles, fuel, maintenance, labour etc.). The relevant equation would be;

$$E_i = F(V_i, F_i,) \qquad \text{(eq. ES.1)}$$

Where E_i is expenditure (or cost), V_i is volume and F_i is factor prices.

If all areas had identical population distributions, identical economic activity patterns and identical geography (spatial and topographical), a simple equation of this form would work well. Estimating such an equation would allow a determination of the way bus costs vary with volume and factor prices. It would show how costs compare between high volume and lower volume areas and what role factor price differences (e.g. differential wage rates) play. It would also reveal the nature of economies of scale. Preferred functional forms for explaining expenditure would be tested consistent with economic theory. Many studies have investigated how costs vary and the shape of the cost function for bus travel using this approach. If the only mode in the jurisdiction was bus, then the equation would also account for the way public transport costs vary across jurisdictions.

The current study deviates from this ideal in three important ways. These are:

1. The areas (SUAs) vary greatly along a number of dimensions relevant to costs of operating and providing public transport such as size, demography or topography. These city specific variables shift the cost function and therefore influence costs in ways unrelated to volume per se.

2. The available data is limited. In particular, volume (passenger kilometres) is only available for the eight capital cities.

3. The areas (SUAs) covered are multi modal. The expenditure of interest covers all public transport combined (bus, rail, ferry).

Recurrent Expenditure Model

We tested a wide range of explanatory variables to identify those that produced the best statistical fit using the available data, and that were consistent with the above principles. The outcome of this testing is the finding that:

- density is a suitable measure for the demand variable,
- bus and train passenger counts are robust proxy variables for supply or network related variables such as congestion or the cost of provision; and
- distance and mean slope are SUA- specific variables that capture spatial and topographical differences between the SUAs and that influence the cost of provision across SUAs.

Specifically, the preferred model uses density ($dense_i$) to depict demand, distance to work ($dist_i$) to represent network complexity, passengers by public transport mode ($pax_{i,mode}$) to represent availability and congestion, and mean land slope ($slope_i$) to account for topography. Formally the model can be specified as:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 \ln(pax_{i,train)} + \beta_5 \ln(pax_{i,bus}) + \varepsilon_i \quad \text{(eq. ES.2)}$$

The estimated coefficients in the model follow intuition as the results suggest that net expenses per person

- increase with urban density (representing demand);
- increase with the distance to work (representing network complexity);
- increase with mean land slope (depicting topographical complexity); and
- increase with train and bus passengers.

The model incorporates passenger mode numbers in a linear-log functional form (the name arises because the independent variables have been transformed by a logarithm, while the dependent variable has not). The linear-log relationship implies that per capita expenses increase as the network becomes more complex but the rate at which this occurs decreases as passenger volumes increase. This holds for buses and rail. Specifically, the linear-log relationship implies that for every 1% increase in passenger mode numbers, per capita expenses increase by a dollar amount equal to the respective estimated coefficient divided by 100.

This means the linear-log form of the model can be interpreted as indicative of scale effects in the wider sense as it suggests that growth from additional passengers becomes less substantial as total volume increases.

The preferred model is consistent with the theoretical framework and also performs well in statistical tests. As the model captures all key relevant (theoretical) drivers, its forecasts can be considered a relevant benchmark for appropriate expenses under each SUA's specific attributes. Hence, it can be applied to derive a policy neutral benchmark per capita expense level for all SUAs.

Capital Expenditure Model

The availability of capital expenditure data is very limited but the observations that could be obtained cover a wide range of values. It appears unlikely that a single-variable model will be able to establish a meaningful functional relationship for the entire range and there are too few observations to estimate (and test) a multi-variable model with confidence. Taking the available theoretical and empirical evidence into account, we therefore recommend using a single expense model that accounts for all key cost factors as the sole basis of the funding allocation mechanism:

- From a theory perspective, there is good reason to expect that operating and capital costs are correlated for a system in "equilibrium" – maintaining services, maintaining utilisation, meeting demand growth as required etc. There is evidence supporting this. For example, the American Public Transport Association annual factbooks report on a large number of public transport cost and performance indicators for US and Canadian public transport systems. They show a very stable relationship between total operating costs and capital costs over time and systems. Between 2001 and 2015 annual urban bus system operating costs varied from 78% to 82% of total operating and capital costs. For heavy rail systems the range was 49% to 60%. For light rail is was 26% to 32%.
- The empirical evidence suggests that distributions of capital expenditure values and that of expenses are very similar. In fact, with a correlation coefficient of 0.98 are very highly correlated with their corresponding values from the expense dataset. Considering the close correlation, it appears very likely that, if a sufficiently large dataset were available for both, an investment and an expense model would generate very similar results.

Treatment of Satellite Cities

The degree to which satellite cities are labour market integrated with their capital cities was assessed by an approach that focusses on revealed travel preferences measured as the self-sufficiency of employment. Based on this approach, an SUA should be considered a satellite to a capital city if:

- it has a relatively high outside SUA dependency index value.
- it has a relatively high dependency to the capital city index value.

Satisfying these two tests would mean that a high proportion of the resident workforce travel outside the SUA to work, and of those travelling outside the SUA to work, a relatively high percentage go to the capital city.

The analysis of Australia's eight capital city regions found the following:

- The SUAs of Gisborne-Macedon, Melton and Bacchus Marsh could be considered labour market integrated satellites to Melbourne based on their self-sufficiency index values. However, as expense data is unavailable, these SUAs cannot be included in the analysis and Melbourne has to be treated on its own.
- Sydney's surrounding SUAs are not satellites to Sydney and should be treated separately.
- Brisbane's surrounding SUAs are not satellites to Brisbane and should be treated separately.
- Neighbouring Perth, Yanchep shows a high capital dependency index of 56% and a very high outside SUA dependency of 81%. Therefore, it should be considered a satellite to Perth.
- Adelaide does not have neighbouring SUAs.
- The ACT consists of a single SUA.
- Hobart does not have neighbouring SUAs.
- Darwin does not have neighbouring SUAs.

1. Introduction

Jacobs and Synergies Economic Consulting ('the Team') have been tasked with assisting the Commonwealth Grants Commission (CGC) to review its modelling of State urban transport recurrent and infrastructure expenditure requirements. The objective of the project is to develop an alternative model that could inform future allocation of GST funds to the States and Territories.

This report represents Stage 2 of the project. Stage 1 was completed in 2017 by Jacobs[1] (this previous report is henceforth referred to as the 'Stage 1 Report'). The Stage 1 report built on the findings of a similar review conducted in 2015[2] ('the 2015 Review') that explored the historical background and consideration of urban transport recurrent and infrastructure models and formulated a first model aimed at quantifying the key drivers of net per capita operating expenses to calculate States' GST requirements.

The concept developed in the 2015 Review rests on the view that larger cities need much more stock per capita than smaller cities. It came to this conclusion by identifying a high correlation between the annual cost of capital charges and the population of each of the cities. Specifically, it specified an recurring expenses model in which per capita expense s increase with the natural logarithm of the associated city and a capital expenditure model which uses the square of urban population as a proxy for asset needs. It explained these findings based on the following two key reasons:

- The number of trips per capita and trip length rise as city population increases and more assets are needed to carry the greater number of users.
- Diseconomies of scale mean larger cities need more capital than smaller cities to undertake the transport task. For example, more buses may be needed because of the slower average travel time in larger cities, or rail systems may be required to meet high levels of demand. Such effects, however, may be partly offset by greater productivity of the assets in larger cities, for example with higher average vehicle occupancy.

This current report refines this modelling by identifying variables that explicitly incorporate these principles and that can be included as relevant quantifiable measures in an econometric model.

The relevant literature shows that expenditure on public transport provision will vary across cities based on the transport task to be undertaken, the characteristics of the transport system and of the specific characteristics of the city itself. This report develops a recurrent expenditure model consistent with this proposition[3]. Variables representing each of these factors have been assembled in a comprehensive data base covering 101 Significant Urban Areas (SUA) defined by the Australian Bureau of Statistics. A number of models have been estimated that are consistent with the basic proposition and tested to determine which recurrent expenditure model best explains the expenditure variations between these cities.

Extending the expenditure dataset from the 42 SUAs with a population of more than 20,000 used in 2015 to the 101 Significant Urban Areas (SUA) defined by the Australian Bureau of Statistics, has enabled a wider range of explanatory variables to be considered and a more versatile and robust model to be developed consistent with the basic theoretical propositions regarding expenditure. The new analytical framework is described in detail below.

1.1 Analytical framework

The Stage 1 report found that the provision of urban public transport can be considered in three components:

- public transport demand;
- public transport service provision with given network capacity, and
- network capacity provision.

[1] Jacobs (2017): Modelling of urban transport recurrent and infrastructure expenditure requirements: Stage 1 report to the Commonwealth Grants Commission

[2] Commonwealth Grants Commission (2015), Report on GST Revenue Sharing Relativities 2015 Review – Volume 2 – Assessment of State Fiscal Capacities.

[3] As presented in section 4, there are insufficient observations to develop a robust infrastructure expenditure model.

Public transport demand and public transport service provision relate to conventional demand and supply. Public transport service provision relates to the availability and accessibility of public transport services at any point in time. The public transport system accommodates demand by providing services within the network capacity it has available. This service capacity will be determined by the management of the existing fleet and road/rail networks and hence determine recurrent expenditure.

Network capacity provision relates to infrastructure investment to allow expansion of the network capacity in the public transport system. This occurs in the long run when network capacity will be expanded (expanded bus fleets, rail rolling stock fleet, track kms etc) to accommodate growth in demand.

The analysis of recurrent expenditure across jurisdictions/areas (e.g. cities, urban centres, and Significant Urban Areas or SUAs) where there are units of varying size is effectively a long run cost analysis. Each jurisdiction is at the same point in time but at a different point on the underlying long run cost curve for the provision of public transport. The point each jurisdiction is at on the underlying long run cost curve depends on demand, as this variable determines the volume of service to be provided. This is illustrated in the following diagram, refer Figure 1.1.

If the production function for the provision of public transport exhibits economies of scale as volume expands and network capacity and operation are optimised at each volume, area 2 in the diagram (Figure 1.1) has lower costs than area 1 by virtue of having a larger volume. The larger volume delivered in area 2 reflects the size of the market in area 2 which is a direct reflection of the position of the demand curve. The demand curve for area 2 is shifted to the right reflecting for example, a higher population or a higher number of employed persons.

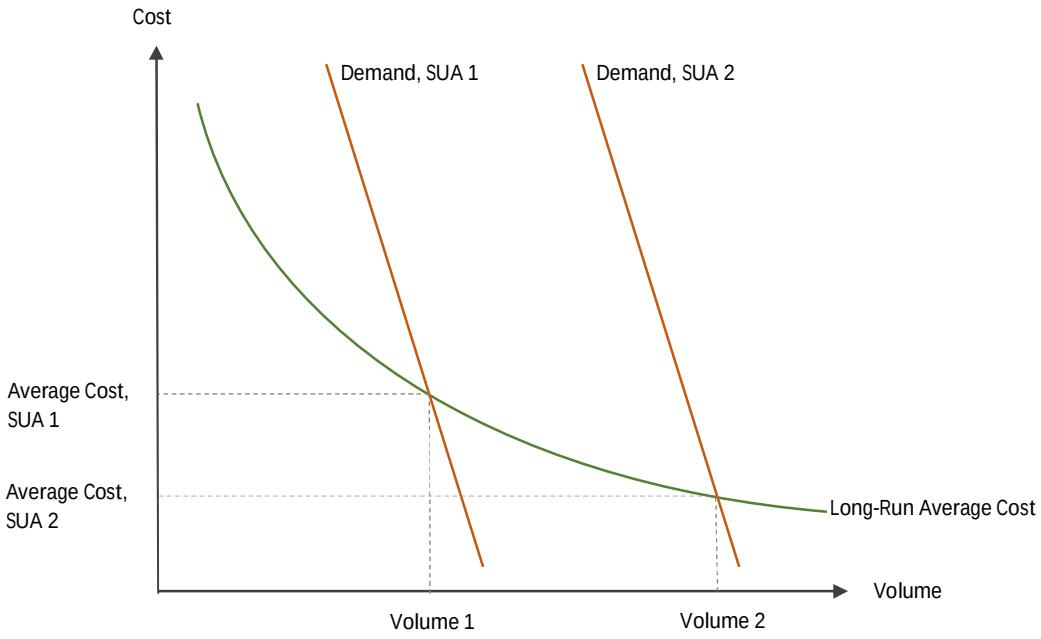


Figure 1.1: Illustrating economies of scale

The easiest way to approach this as a modelling task is to illustrate the underlying concept by focusing on a single mode such as bus transport. The cost of operating a bus system would depend on volume and input prices. The relevant equation would be;

$$E_i = F(V_i, F_i,) \qquad \text{(eq. 1.1)}$$

Where E_i is expenditure (or cost), V_i is volume and F_i is factor prices.

If all areas had identical population distributions, identical economic activity patterns and identical geography (spatial and topographical), a simple equation of this form would work well. Estimating such an equation would allow a determination of the way bus costs vary with volume and factor prices. It would show how costs

compare between high volume and lower volume areas and what role factor price differences (e.g. differential wage rates) play. It would also reveal the nature of economies of scale.

Preferred functional forms for explaining expenditure would be tested consistent with economic theory. Many studies have investigated how costs vary and the shape of the cost function for bus travel using this approach. If the only mode in the jurisdiction was bus, then the equation would also account for the way public transport costs vary across jurisdictions.[4]

However, the current study deviates from this ideal in three important ways. These are:

1. The areas (SUAs) vary greatly along a number of dimensions relevant to costs of operating and providing public transport such as size, demography or topography. These city specific variables shift the cost function and therefore influence costs in ways unrelated to volume per se.

2. The available data is limited. In particular, volume (passenger kilometres) is only available for the eight capital cities.

3. The areas (SUAs) covered are multi modal. The expenditure of interest covers all public transport combined (bus, rail, ferry).

These added complexities have been given careful consideration in this study. Our treatment of each is discussed in turn below.

Item 1 – City-specific characteristics

Recognising that area specific variables can influence costs causes the cost/expenditure equation to be of the form:

$$E_i = F(V_i, F_i, C_i) \qquad \text{(eq. 1.2)}$$

Where V_i is volume, F_i are factor prices (or price indexes) and C_i are area specific variables.

These area or city specific variables are potentially many and varied. Candidates include; terrain, congestion level, average trip distance, passenger density, population density, employment density, economic activity dispersion, age of the city.

Strictly speaking some of these variables may not be independent of volume. For example, volume as measured by vehicle or passenger kilometres will be correlated with some of these city or jurisdiction specific variables. This collinearity introduces potential bias in the estimated coefficient on volume, leading to erroneous conclusions about scale economies. Careful attention needs to be paid to these area specific variables with appropriate statistical testing of functional form and of the statistical significance relationship to ensure an unbiased estimate of the coefficient on volume and the analysis of economies of scale. However, such an analysis is not possible for this study because, as noted above and developed further in Section 2.1, the dataset for this study does not have volume data available for all areas. In this case city specific variables included are partly acting as proxies for volume and partly capturing area specific effects.

Item 2 – Limited data

The majority of the expenditure data to be used in this analysis were reported directly by the States to the CGC, which then provided it to the Team. Not all data have been derived in the same way. As a result, data

[4] Not surprisingly linear and quadratic forms have always been candidates for testing for the presence of scale economies. Early applications of this approach include: Wabe, J and O. Coles. "The Short and Long Run Cost of Bus Transport in Urban Areas". *Journal of Transport Economics and Policy.* Vol. 9. No 2. (1975), pp. 127-140. . An early attempt to be more specific about the production function and the resulting cost function is: Tauchen, H. Fravel, F, and G, Gilbert. "Cost Structure of the Intercity Bus Industry". *Journal of transport economics and Policy*. Vol. 17. No. 1. (1983). pp. 25-47. They apply a translog cost function derived from a multiproduct cost function in order to be make the underlying production and cost conditions consistent with economic theory. However, studies are often limited by data and the nature of the bus systems being studied. On balance the question of economies and diseconomies of scale in bus systems is still an open question. (remove extra full stop) For more on this see the short but informative literature review in: L. de Grange. Tronsco, R. and I. Briones. "Cost, production and efficiency in local bus industry: An empirical analysis for the bus system of Santiago". *Transportation Research Part A: Policy and Practice*. Vol. 108. Dec. (2017). pp. 1-11.

consistency for the dependent variable (that is, recurrent transport expenditure) is one of the major challenges of this assignment. Since independent variable data are readily available or require only minor modifications, this report builds up the proposed modelling dataset starting with the dependent variable.

Following on from item 1, the problem of limited data requires the use of proxies to account for volume and, by implication, scale effects. No one proxy will work. Volume reflects both the demand that has been satisfied and the supply provided. Measured as passenger kilometres it encompasses both demand and network variables. Demand variables directly connected to volume include population, employment and education enrolments. Network variables related to volume include average trip length, congestion, and network density. By definition, in the absence of a single volume measure, no simple interpretation of economies of scale is possible. Scale effects are measured in relation to the proxy variables and are not subject to the same expected values as for a single volume measure. Lack of volume data and associated pricing data for the areas in this study (SUAs) means that separate estimation of demand and supply curves is not feasible using the available observations.

Item 3 – Multiple modes

In an ideal world we might estimate an expenditure/cost function for each of the modes and generate an aggregate expenditure for each area (SUA) by aggregating the expenditure for that area (SUA) across the modes at the relevant modal volumes. There would be different underlying cost functions for each mode and varying economies of scale. There would also be separate demand curves for each mode and modal share would have to be modelled for each area. However, as already noted, volume as passenger kilometres is not available meaning that such an analysis is not possible. Proxies for volume are required making such an analysis logically impossible. For example, if the proxy was population it would imply the same volume for each mode and if the coefficients were different it would imply that the same population has differential effects on public transport costs for an SUA.

In the absence of volume data, it is preferable to assume that the demand and network proxies drive costs across all modes, that jurisdictions have optimised modal mix, and that the aggregate cost for public transport will be influenced by the mix of modes. In this way we can account for the inherent cost differences in carrying passengers by rail, bus and ferry. Candidate variables would be dummy variables for mode existence (e.g. heavy rail) and measures of the relative size or importance of modes in an area.

In essence, without volume the ideal equation to explain costs across areas is replaced by a version that substitutes volume with relevant demand and network proxies and includes additional area or city specific variables, and variables that relate directly to the costs of provision to the extent that they can be shown to affect costs over and above the volume proxies.

This means estimating a single equation of the form:

$$E_i = F(D_i, S_i, C_i) \qquad \text{(eq. 1.3)}$$

Where E_i is expenditure (as net expenditure), D_i are demand variables (essentially proxies for volume) S_i are supply or network related variables, some of which are proxies for volume and some of which capture cost of provision and factor price effects and C_i are city specific variables that capture differences between or SUA specific variables that influence differences in cost of provision across SUAs.

Estimating the cost function equation (equation 1.1) would require that volume be included as a matter of theoretical accuracy. As soon as we move away from the ideal form (equation 1.1) we lose this requirement because we are dealing with proxies for volume and a range of city or area specific variables some of which are related to volume and some of which capture cost of provision impacts. Exactly what functional form works best, and which variables are included becomes essentially a pragmatic exercise[5]. However, there are two points worth noting.

[5] The idea that city specific variables to be included is essentially a pragmatic exercise goes back a long way in the literature. It was first argued by Miller in Miller, D. R. "Differences Among Cities, Differences Among Firms, and Costs of Urban Bus Transport." *Journal of Industrial Economics*, Vol. 19 No. 1 (1970), pp.22-32.

First, based on the theory we expect volume to be a variable that should be included in any expenditure or cost function. Therefore, in the current case where volume data is unavailable for all areas and proxy variables for volume must be used, we expect the model form used to have variables from the set of proxy variables for volume along, and with, a set of area specific variables. This will keep the model consistent with the theoretical framework underpinning equation 1.3.

Second, the expected functional form is an open question to some extent. This is because, out of necessity, the volume variable has been replaced with appropriate proxies and because different modes with different cost characteristics are combined into aggregate expenditure. However, we do expect an element of non-linearity based on findings for bus and rail costs reported in the literature. The recent study by Graham et al of the rail costs and productivity across 17 rail systems in cities around the world looked specifically at the role of returns to scale. Their estimates reveal constant returns to scale but increasing returns to density for rail.[6] A study of Swiss bus and trolley bus systems found increasing returns to scale.[7] Significant economies from expanding bus vehicle miles were found by Williams in a study of US systems[8]. Density and scope economies for urban bus transport have been reported by Giacomo and Ottoz in their study of urban and intercity bus systems.[9]

Based on these insights one would expect to at least see economies of scale across both key modes in a wider sense: The associated coefficients do not necessary have to be negative – this would imply lower costs per capita or passenger – but could just show slowing growth for additional passengers.

1.2 Structure of this report

The report is structured into the following sections:

- Section 2 presents key concepts that were applied when deriving the data used in the subsequent analysis.
- Section 3 provides an overview of the recurrent expenditure data, explores challenges and identifies remedies for these issues.
- Section 4 provides an overview of the infrastructure expenditure data and establishes linkages between this and the recurrent infrastructure both theoretically and empirically.
- Section 5 summarises inputs and outputs of the econometric analysis and applies the preferred models.
- Section 6 presents conclusions.
- Section 7 contains the table of references and data sources.
- Appendix A presents and assesses candidate demand variables.
- Appendix B presents and assesses candidate supply variables.
- Appendix C presents and assesses candidate cost variables.
- Appendix D presents and assesses the correlation between candidate independent variables.
- Appendix E contains the technical details of the econometric analysis.
- Appendix F contains a table with expense data availability and quality by SUA.
- Appendix G contains a table with self-sufficiency index values by SUA.

[6] Graham, D.J. et al. "Economies of scale and density in urban rail transport: effects on productivity." *Transportation Research Part E: Logistics and Transportation Review.* Vol. 39. No. 6. (2003), pp. 443- 458.

[7] Farsi, M. Fetz, A. and Massimo Filippini, "Economies of Scale and Scope in Local Public Transportation" *Journal of Transport Economics and Policy.* Vol. 41, No. 3 (2007), pp. 345-36.

[8] Williams, M. "Firm Size and Operating Costs in Urban Bus Transportation". *Journal of Industrial Economics.* Vol. 28. No 2. (1979), pp. 209-218.
[9] Giacomo, M. and Ottoz, E. "The relevance of Scale and Scope Economies in the Provisions of Urban and Intercity Bus Transport." Journal of Transport Economics and Policy. Vol. 44. No 2 (2010), pp 161-187.

2. Concepts

This section examines topics that have been the subject of ongoing discussion in the context of allocating funds for recurring infrastructure expenditure:

- How can the selected model ensure policy neutrality of allocation?
- What is the most appropriate geographic basis for measuring expenditure?
- How can potentially important drivers, for which only very limited data is available, be included in the modelling framework?

These matters are important to address at the outset, as they constitute overarching considerations that will affect all candidate variables tested. Moreover, they can ultimately also affect all considered functional forms of the model.

2.1 Policy neutrality

The Team understands the important principle and imperative that funding allocations to States and Territories are independent of policy factors that may otherwise drive the apparent 'need' for more or less funding. The challenge is to develop an econometric model that recognises the influence that policy factors have on expenditure levels, and to subsequently remove the effect of these factors on funding shares.

The Stage 1 report articulates this principle of 'policy neutrality':

"The principle of the CGC's advice on GST revenue distribution among states and territories (States) is horizontal fiscal equality (HFE). Therefore, the recurrent expenditure model must be independent of the policy of individual governments (policy neutral) and reflect what States do on average."

In our opinion, policy neutrality and a reliable model can only be ensured following a two-step modelling process:

1) Estimate a model that includes variables accounting for **both** policy-related and policy neutral cost drivers.
2) Use this model to adjust the expenditure observations to policy neutral levels by removing the effect of policy variables on expenditure. Funds can then be allocated based on the relationships of these standardised expenditure levels.

If policy-related independent variables are excluded from the outset, their effect on the dependent variable (expenditure) will not be excluded from analysis. Rather it will be attributed to other variables included in the model with which they are correlated or through the error term. This leads to what is referred to as 'omitted variable bias'. It means that potential policy influences on the dependent variable will be reflected in the coefficient estimates associated with the (policy neutral) independent variable(s). In other words, any allocation mechanism based on the resulting model will be subject to policy effects, but these effects will not be visible in the model specification and therefore cannot be adjusted for. Clearly, in situations where transport expenditures are heavily influenced by policy factors, 'other' variables such as population may be statistically good fits to the data, but they are unlikely to be good explanators of the dependent variable as their coefficients will be biased.

The implication of omitted variable bias can be illustrated by way of example: Assume there are three SUAs with similar populations. They are similar in their public transport layout, but one has a light rail, which is considered a policy decision. Light rail (LR) can be introduced into the model as a dummy variable, where LR = 1 signifies that light rail is present, while LR = 0 indicates that there is no light rail.

Table 2.1 presents this illustrative dataset. It shows that the locality with the light rail has significantly higher expenditure than the two others.

Table 2.1: Ensuring policy neutrality: illustrative dataset

SUA	Expenditure	Population (*pop*)	Light rail present (*LR*)
	$ million	million persons	
A	1,200	4.600	0
B	1,210	4.606	0
C	1,500	4.670	1

Figures for illustrative purposes only

If policy related variables are to be excluded at the outset of the regression analysis, population is the only independent variable and the estimated (linear) regression model is:

$$y = 4{,}386.6pop - 18{,}986 \quad \text{(Model 1)} \quad \text{(eq. 2.1)}$$

This equation reproduces the data reasonably well (see Table 2.2 below). It would suggest that after a certain point represented by the negative intercept, a locality should receive $4.39 billion in funding for every one million persons.

If the presence of the light rail is included as a dummy variable, the estimated (linear) regression model is:

$$y = 1{,}666.7pop + 183LR - 6{,}467 \quad \text{(Model 2)} \quad \text{(eq. 2.2)}$$

Table 2.2 compares the results of the two models.

Model 2 reproduces the data well too. It is superior to Model 1 because it isolates the expenditure of the light rail. Model 2 suggests that after a certain point represented by the negative intercept, a locality should receive (a much lower) $1.67 billion in funding for every one million persons. Furthermore, Model 2 can be used to show that the expenditure associated with the light rail in SUA C amounts to approximately $183 million. This (policy driven) expenditure can now be subtracted and thus a policy neutral expenditure level can be derived.

Table 2.2: Ensuring policy neutrality: illustrative estimates

SUA	Model 1	Model 2	
		Total estimate	Policy neutral estimate
A	1,192	1,200	1,200
B	1,218	1,210	1,210
C	1,499	1,500	1,317

Figures for illustrative purposes only

The results in Table 2.2 demonstrate that the policy neutral estimate produced by Model 2 for SUA C is much lower than that produced by Model 1. In contrast, the total estimate of Model 2 and Model 1 are almost identical indicating that Model 1 estimates are still policy influenced.

The much lower population coefficient (reflected in the lower slope of the blue line in Figure 2.1 below) estimated by Model 2 is further testament to the inaccuracy of Model 1. By excluding the presence of light rail from the regression analysis, its effect on expenditure was attributed to the (relatively small) difference in population between SUA C and the other two and thus this model over-estimates the magnitude of this effect. Including the light-rail variable enabled us to correct for this expenditure driver and thus ensure policy neutral results. Figure 2.1 illustrates this correction.

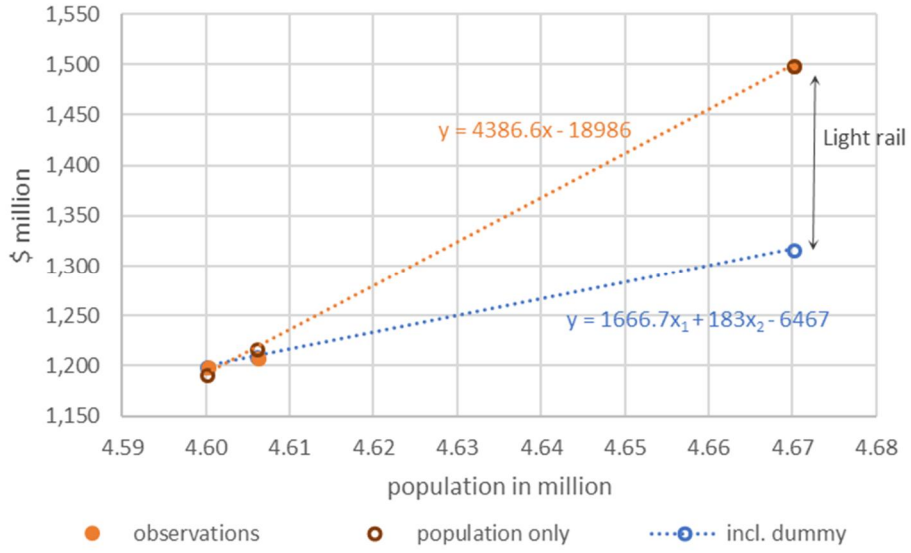


Figure 2.1: Ensuring policy neutrality: graphical illustration

Figure for illustrative purposes only

Another example of an application of the above principle could be the level of service offered measured by the number of stops. The number of bus/train/light rail stops in a State is directly under the control of States. If the number of stops was included in the regression, the model could be run with an appropriate number of stops – e.g. the average per head across all jurisdictions or even an international benchmark – to estimate expenditure levels under average policy.

Applying these principles to the actual data will not be as straightforward as the worked example above. The distinction between policy neutral and policy influenced variables can be grey at times and hence setting a policy neutral value for the policy driven variables will be a less trivial task. Nevertheless, including (at least some) ability to control for policy measures will ensure more robust and transparent estimates than introducing a bias to the model by omitting potentially key explanatory variables.

2.2 Urban self-sufficiency

The Stage 1 report identified the ABS' Statistical Urban Area (SUA) as the preferred geographic definition, and data from the States has been collected as this level. The ABS defines a SUA as follows:

The regions of the SUA structure are constructed from whole SA2s. They are clusters of one or more contiguous SA2s containing one or more related Urban Centres joined using the following criteria:

- *they are in the same labour market*
- *they contain related Urban Centres where the edges of the Urban Centres are less than 5km apart defined by road distance*
- *they have an aggregate urban population exceeding 10,000 persons*
- *at least one of the related Urban Centres has an urban population of 7,000 persons or more.*[10]

[10] ABS cat 1270.0.55.004 - Australian Statistical Geography Standard (ASGS): Volume 4 - Significant Urban Areas, Urban Centres and Localities, Section of State

This means the ABS has in effect already made some economic judgments about the relationship between SA2s when aggregating them to form SUAs. For this study, the key question is therefore whether any SUAs should be combined. That is, are there any SUAs that should be considered as having a sufficiently integrated labour market with the neighbouring capital city? We will refer to such cases as labour market integrated satellites.

In our view, the best way to proceed here is to develop and apply criteria that assesses if SUAs exist that could be considered labour market integrated satellites to a capital city. The criteria are, as much as is possible, to be independent of any consideration of what their application means for model estimation and results. Their purpose is to test whether any meaningful satellites can be found. The underlying approach would only be considered fit for this purpose if it could be systematically and equally applied to all Australian cities.

Rationale

The Stage 1 report suggests applying a travel time threshold for assessing whether or not a satellite city forms part of a greater capital city. The logic is that beyond a certain travel time threshold an area could be deemed independent. It notes that in international literature thresholds range from 40 to 180 minutes. This is a wide range and suggests that it is highly likely that the acceptable commute time depends on the specific characteristics of each individual city and the nature of the transport network. There is no agreed commute time to use in Australian cities. Additionally, the urban form of cities is dynamic and what may be an unacceptable commute time today could be considered normal in five years. Consequently, travel time does not meet the above requirement.

We therefore propose to adopt an approach that focusses on revealed travel preferences (rather than hypothetical benchmarks) measured as the self-sufficiency of employment. Examples of the application of such a measure can be found in academic literature[11] as well as government planning publications[12]. It can be applied equally to all urban areas and since it is dynamic its results can change as a city develops.

The index is limited to employment related commuting and does not include students. In addition to the absence of reliable comprehensive data on their commuting patterns, students tend to be constrained in their choices as they do not have access to all available transport modes and the choice of (public) schools is often mandated by pre-determined school catchment areas. This means that, as the associated commuting patterns are the most flexible, an employment-based index is likely to constitute the upper bound of capital city dependency. In other words, if a potential labour market integrated satellite is not considered part of a capital city under the framework set out below, it is even less likely to be considered a dependent satellite if student commutes were to be included.

The idea is that the travel patterns themselves reveal how well two areas are interconnected from the perspective of their residents. Many planners focus on employment self-sufficiency when planning for city growth. The best example is planners attempting to relocate jobs to a middle or outer suburb. The objective is to have people able to work closer to home, thereby reducing travel times.

There are different ways to think about and measure employment self-sufficiency. One approach looks at the proportion of local resident workers who work within an area. It indicates the extent to which local residents seek employment outside the area in which they live. Generally speaking, this containment of employment will be higher in regional towns and lower in areas within a wider urban area. It is in part a function of separation (transport distance and time), in part a function of the transport systems, and in part a function of the skills of the residents and how well they match available jobs. Where the share of contained employment is low, residents are commuting to an outside area for work.

Simple examples can illustrate this point: In Western Australia, a distinct regional city, Albany, has 86.6% of its resident workers employed locally, while only 13.4% leave the area for work. The City of Armadale, an LGA (local government area) at the end of suburban rail at the south-eastern fringe of the Perth metropolitan area,

[11] For example: Bierman, S. and Martinus, K. (2017) *Boundary Objects as Tools for Integrated Land Use Planning*. In Bierman, S. Olaru, D and P. Valeria, editors. Planning Boomtown and Beyond. Perth: UWA Press; or Kirsten Martinus & Sharon Biermann (2018) *Strategic Planning for Employment Self-Containment in Metropolitan Sub-Regions*, Urban Policy and Research, 36:1, 35-47.

[12] For example: Western Australian Planning Commission (2018), *Perth and Peel and 3.5 million*.

has only 23.8% of its resident workers employed locally, with 76.2% leaving the area to work. The inner ring LGA of the City of Vincent which is a direct neighbour of the Perth CBD has just 15.9% of its resident workers employed locally, with 84.1% leaving the area to work.

Index definitions

We defined three indices covering three geographic dimensions to capture key aspects of extended labour market integration through commuting. The index definitions encompass our approach to using the concept of self-sufficiency to help define the appropriate geography for the study, in particular, determining whether there is a case for defining an area as a labour market integrated satellite city. In using the self-sufficiency index measures we consider both the percentage of workers leaving an area to work and also the percentage of these workers going to the adjacent capital city.

The below formulae use the following variables:

R_{SA2} ~ Resident workforce (number of persons residing and working in a SA2)

LF_{SUA} ~ SUA workforce (number of persons residing in a SA2 and working in the assigned SUA less the resident workforce)

LF_{CC} ~ Capital city workforce (the number of persons residing in each SA2 and working in the assigned capital city SUA less the resident workforce if the SA2 is part of a capital city SUA)

LF_O ~ number of persons travelling to a non-SUA area.

The three indices are:

- *Outside SA2 dependency*
  is calculated as the share of population working outside the SA2. As a result of the SA2's relatively small area, most persons are likely to work outside the SA2 they live in. It can therefore be expected to be relatively high in most instances. It is always larger than the outside SUA dependency index. Formally:

$$I_i^{SA2} = 1 - \frac{R_{SA2}}{R_{SA2}+LF_{SUA}+LF_{CC}+LF_O} \qquad \text{(eq. 2.3)}$$

- *Outside SUA dependency*
  is calculated as the share of population working outside the assigned SUA. A high value means that many people travel to areas outside their resident SUA for work. It can thus be a first indicator for an urban centre being a labour market integrated satellite to a capital city. This would be the case if an SUA neighbouring a capital city shows a high index value. Formally:

$$I_i^{SUA} = \begin{cases} 1 - \frac{LF_{CC}+LF_{SUA}+R_{SA2}}{R_{SA2}+LF_{SUA}+LF_{CC}+LF_O} & if\ SA2\ in\ capital\ city \\ 1 - \frac{LF_{SUA}+R_{SA2}}{R_{SA2}+LF_{SUA}+LF_{CC}+LF_O} & else \end{cases} \qquad \text{(eq. 2.4)}$$

- *Capital city dependency*
  is calculated as the share of population working in the assigned capital city. A high index value means that a large proportion of the people living in an SA2 work in the associated capital city. It can therefore be the definitive indicator for an urban centre being a satellite to a capital city: if an SA2 shows a high outside SUA dependency and a high dependency to the capital city that is not its SUA, it can be considered a labour market integrated satellite to this capital city. Formally:

$$I_i^{CC} = \begin{cases} \frac{LF_{CC}+R_{SA2}}{R_{SA2}+LF_{SUA}+LF_{CC}+LF_O} & if\ SA2\ in\ capital\ city \\ \frac{LF_{CC}}{R_{SA2}+LF_{SUA}+LF_{CC}+LF_O} & else \end{cases} \qquad \text{(eq. 2.5)}$$

All three indices can be calculated using ABS Census 2016[13] data extracted as a cross tabulation of place of usual residence (PUR) to place of work (POW) by SA2. This serves as a proxy origin-destination matrix (O-D matrix) of commuters from/within each SA2. The data were accessed by state which means that interstate travel is not captured. For the analysis they were combined with spatial files of ABS boundaries for SA2 and SUA also published as part of the 2016 Census.

The three indices can be derived from this data in the following steps:

1) Match each SA2 to an SUA using a spatial query and add this SUA as a second destination identifier to the O-D matrix.
2) Assign a capital city (SUA boundaries) to each SA2 based on the State as shown in Table 2.3 below.
3) Add population and labour force data to the SA2 (spatial) data set and calculate population density as population/area in km^2
4) Query the O-D matrix to extract:
    a) The resident workforce (R_{SA2}) calculated as the number of persons residing and working in each SA2
    b) The SUA workforce (LF_{SUA}) calculated as the number of persons residing and working in the assigned SUA less the resident workforce. The SUA resident workforce is set to zero if the SUA assigned to the SA2 is a capital city SUA.
    c) The capital city workforce (LF_{CC}) calculated as the number of persons residing in each SA2 and working in the assigned capital city SUA less the resident workforce if the SA2 is part of a capital city SUA.
    d) The number of persons travelling to a non-SUA area (LF_O) calculated as the difference between the total labour force and the sum of the resident workforce, the SUA workforce and the capital city workforce.

Table 2.3: Capital cities by state and number of associated SA2

State	Capital city	Number of SA2 in state
New South Wales	Sydney	576
Victoria	Melbourne	462
Queensland	Brisbane	528
Western Australia	Perth	252
South Australia	Adelaide	172
Australian Capital Territory	Canberra	131
Tasmania	Hobart	99
Northern Territory	Darwin	68

Source: Synergies analysis of Census 2016 data

2.3 Geographic definitions

In order to identify the most appropriate definition of urban areas we constructed the suite of employment self-sufficiency indices presented above. Based on the principles established above, an SUA should be considered a satellite to a capital city if:

- it has a relatively high outside SUA dependency index value.
- it has a relatively high dependency to the capital city index value.

[13] All Census data used in this report is referenced as Census 2016. The specific datasets used are listed in the references section.

Satisfying these two tests would mean that a high proportion of the resident workforce travel outside the SUA to work, and of those travelling outside the SUA to work, a relatively high percentage go to the capital city.

The following provides an assessment of the areas around Australia's eight capital cities based on these criteria. All three indices defined in the previous section were calculated for all SA2s in Australia. They were assigned to the spatial file and plotted as heat maps below, which also highlight SA2s with population densities of between 500 and 1,000 and more than 1,000 persons per square kilometre as an indication of built-up areas. In addition to the maps, a table showing the numeric values for the employment self-sufficiency index for each SUA is presented in Appendix G.

The criteria have been applied at the level of 60% working outside and 40% working in the capital city.

The analysis of Australia's eight capital city regions is laid out in detail on the following pages. It finds the following:

- The SUAs of Gisborne-Macedon, Melton and Bacchus Marsh could be considered labour market integrated satellites to Melbourne based on their self-sufficiency index values. However, as expense data is unavailable, these SUAs cannot be included in the analysis and Melbourne must be treated on its own. The sensitivity of the estimates to a higher expense level in Melbourne will be tested in Appendix F.
- Sydney's surrounding SUAs are not satellites to Sydney and should be treated separately.
- Brisbane's surrounding SUAs are not satellites to Brisbane and should be treated separately.
- Neighbouring Perth, Yanchep shows a high capital dependency index of 56% and a very high outside SUA dependency of 81%. It should be considered a satellite to Perth.
- Adelaide does not have neighbouring SUAs.
- The ACT consists of a single SUA.
- Hobart does not have neighbouring SUAs.
- Darwin does not have neighbouring SUAs.

Melbourne

The maps in Figure 2.2 below show that in the greater Melbourne area the SUAs of Gisborne-Macedon, Melton and Bacchus Marsh could be considered satellites to Melbourne. For all associated SA2s, over 60% of the workforce work outside these SUAs and at least 40% (in some SA2s more than 60%) work in the Melbourne SUA. However, as expense data is unavailable these SUAs cannot be included in the analysis and Melbourne must be treated on its own. The sensitivity of the estimates to a higher expense level in Melbourne will be tested in Appendix F.

Other potential candidates for satellite cities such as Geelong, Ballarat, Bendigo, and Drouin-Warragul show very low capital city dependency values and are therefore to be treated separately.

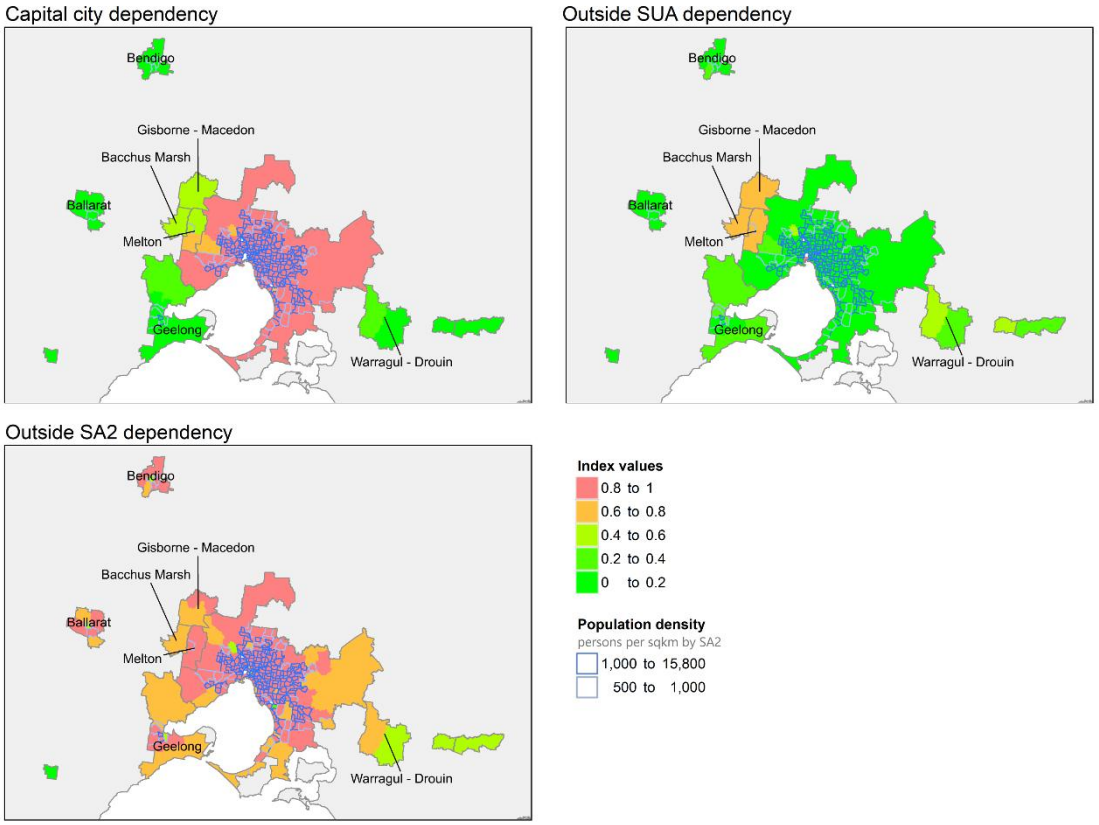


Figure 2.2: Melbourne indices

Data source: Synergies map

Sydney

In Figure 2.3, the maps illustrate that the northern (low population density) part of Wollongong exhibits a relatively high capital dependency index of more than 60%. However, since the southern part (which accommodates the majority of its population) shows low index values and the SUA wide average capital dependency index is 15%, Wollongong can be classified as self-sufficient.

Based on the principles above, all other SUAs are not satellites to Sydney and should be treated separately.

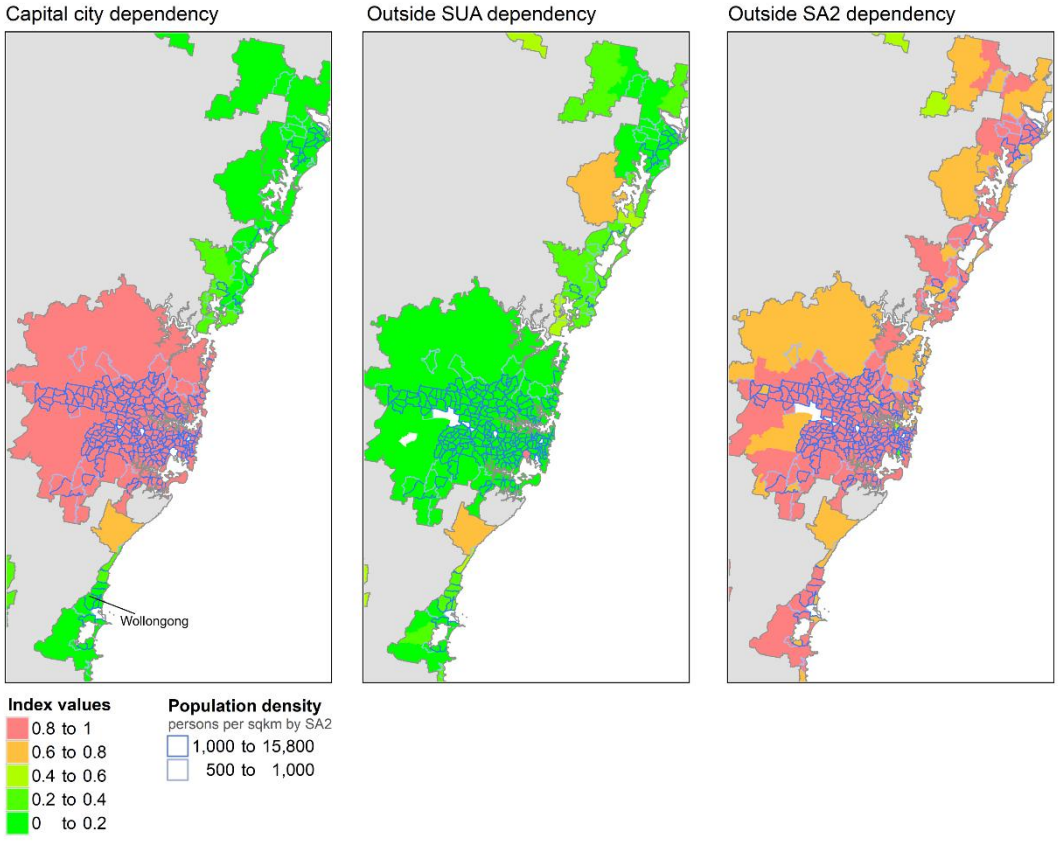


Figure 2.3: Sydney indices

Data source: Synergies map

Brisbane

Based on the principles above, all SUAs around Brisbane are not satellites and should be treated separately, as shown in Figure 2.4. The two neighbouring SUAs of Sunshine Coast and Gold Coast both have capital city dependency index values below 20%.

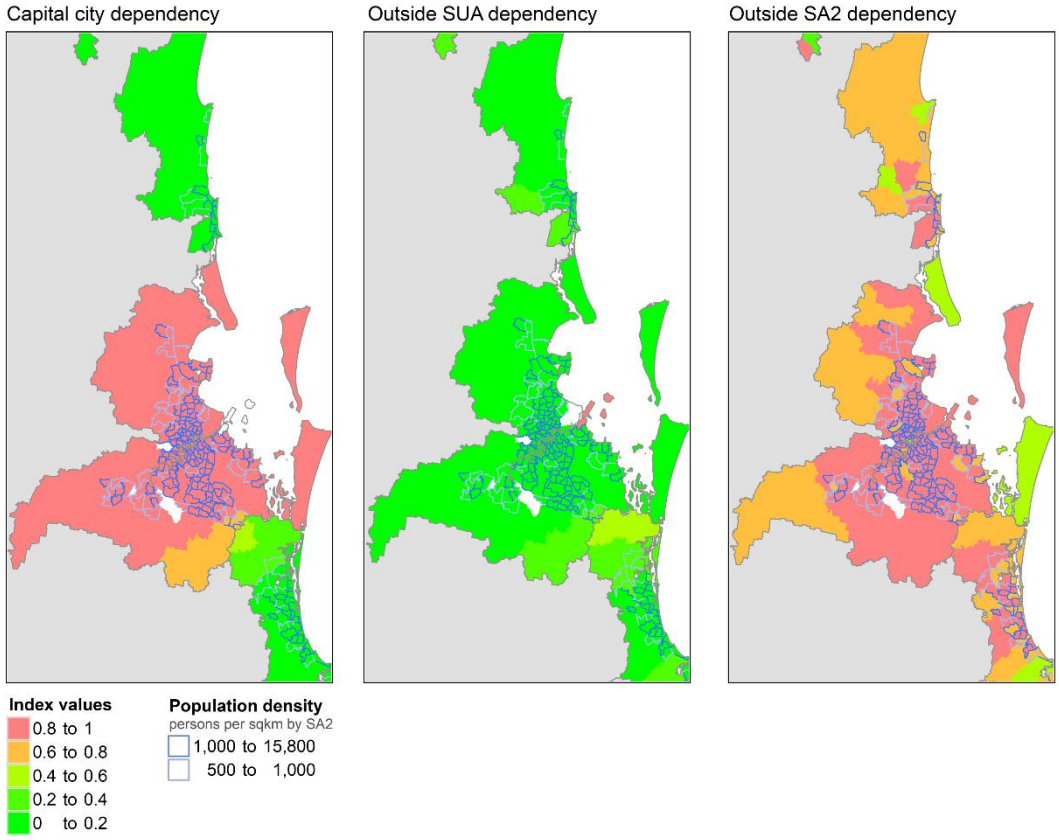


Figure 2.4: Brisbane indices

Data source: Synergies map

Perth

The maps in Figure 2.5 show that the southern (high population density) part of Yanchep shows a high capital dependency index of 56% and a very high outside SUA dependency of 81%. Therefore, it should be considered a satellite to Perth.

Yanchep is Perth's only neighbouring SUA.

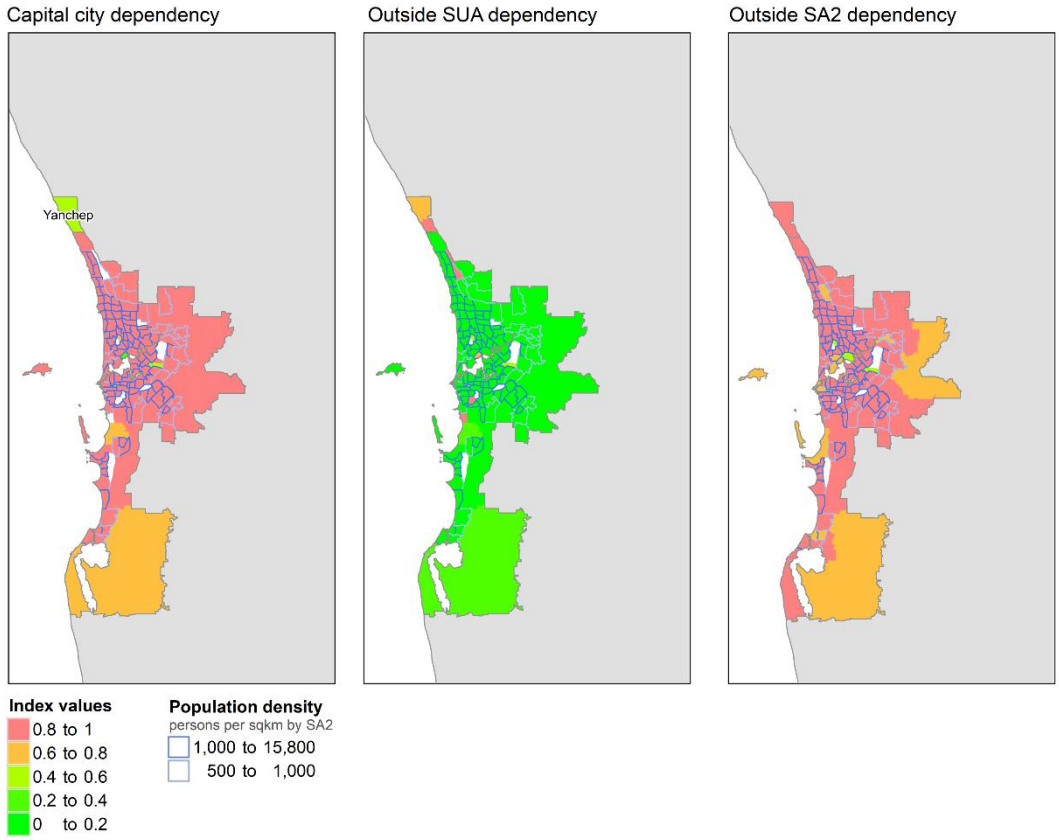


Figure 2.5: Perth indices

Data source: Synergies map

Adelaide

Adelaide does not have neighbouring SUAs, as per Figure 2.6.

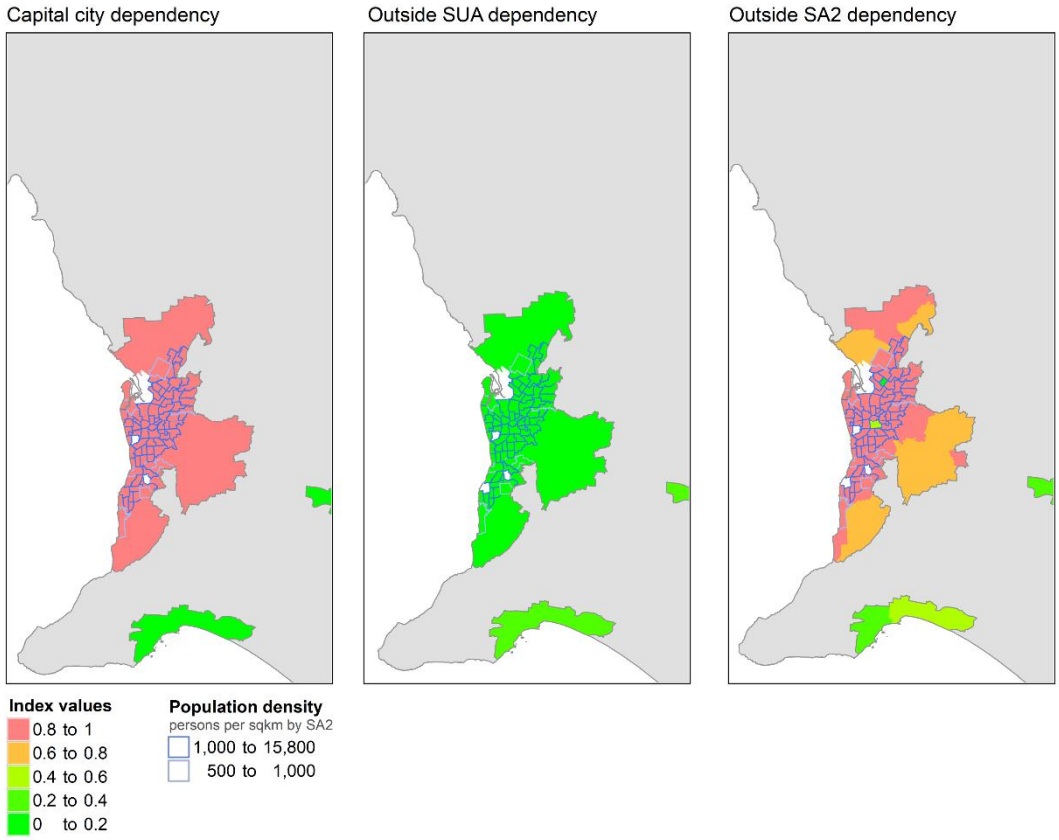


Figure 2.6: Adelaide indices

Data source: Synergies map

Canberra

Canberra consists of a single SUA, as shown in Figure 2.7. Neighbouring SUAs are to be treated separately as they are outside the State border.

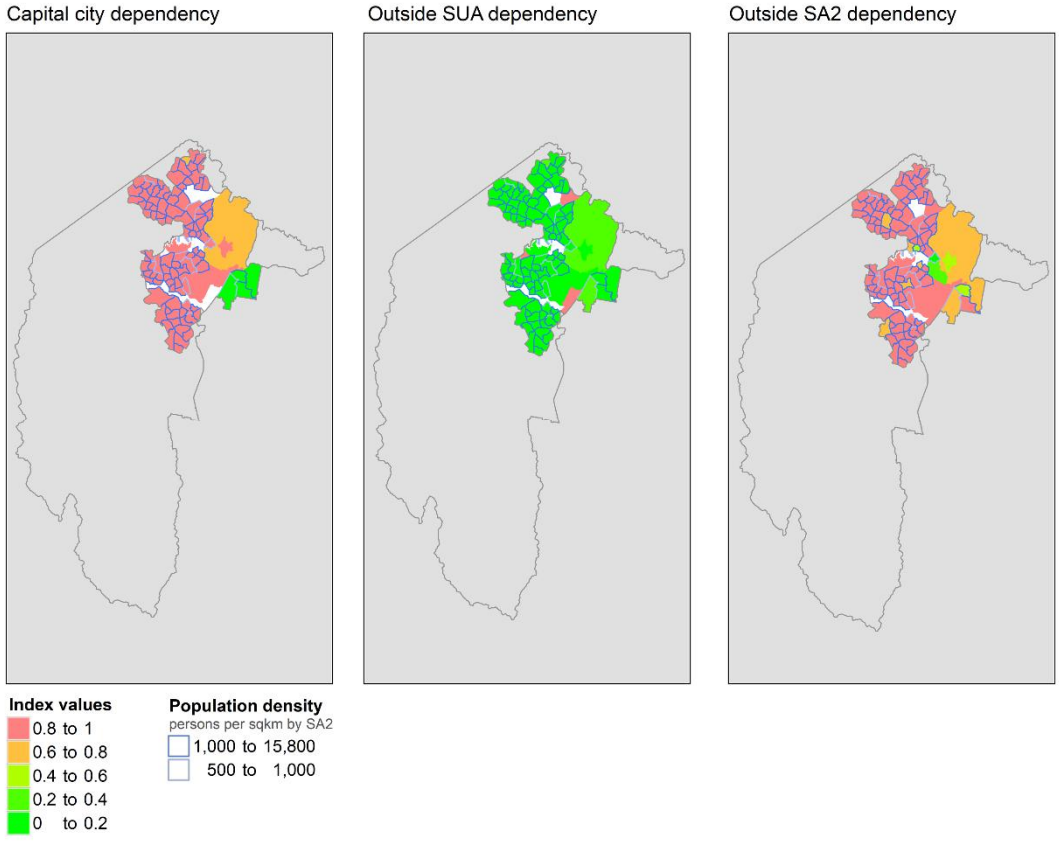


Figure 2.7: Canberra indices

Data source: Synergies map

Hobart

Hobart does not have neighbouring SUAs, as per Figure 2.8.

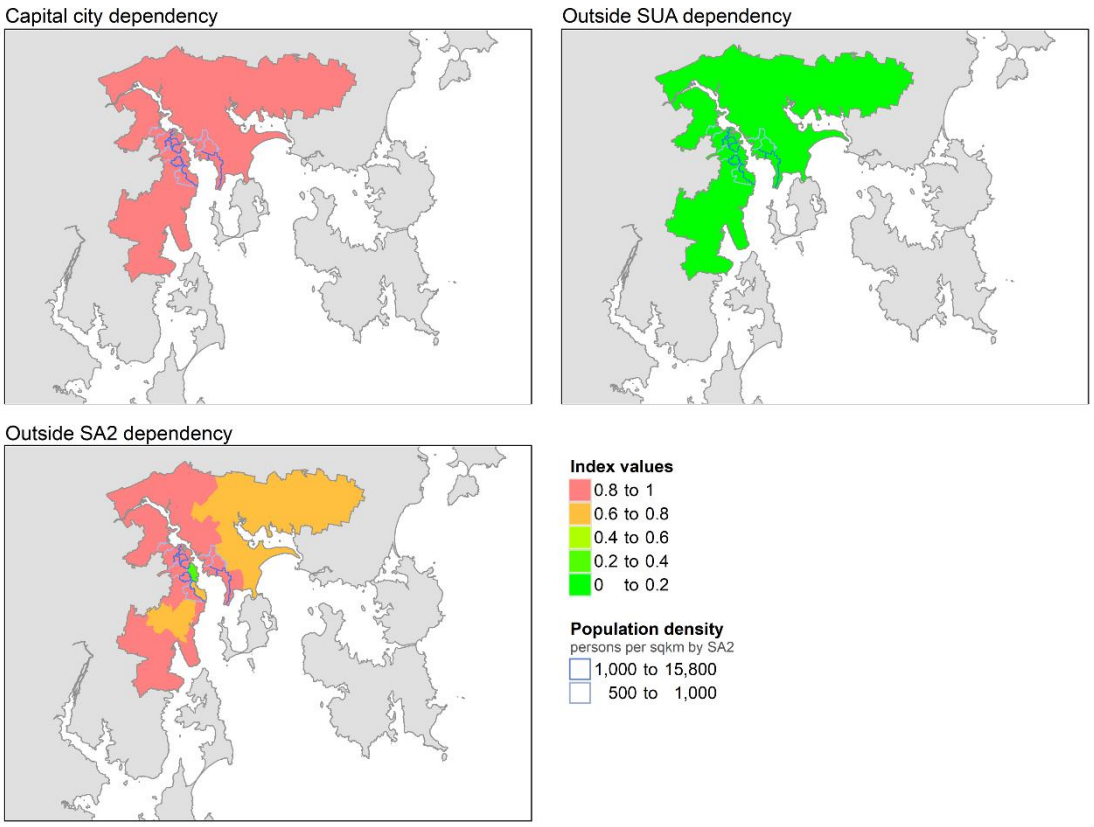


Figure 2.8: Hobart indices

Data source: Synergies map

Darwin

Darwin does not have neighbouring SUAs, as per Figure 2.9.

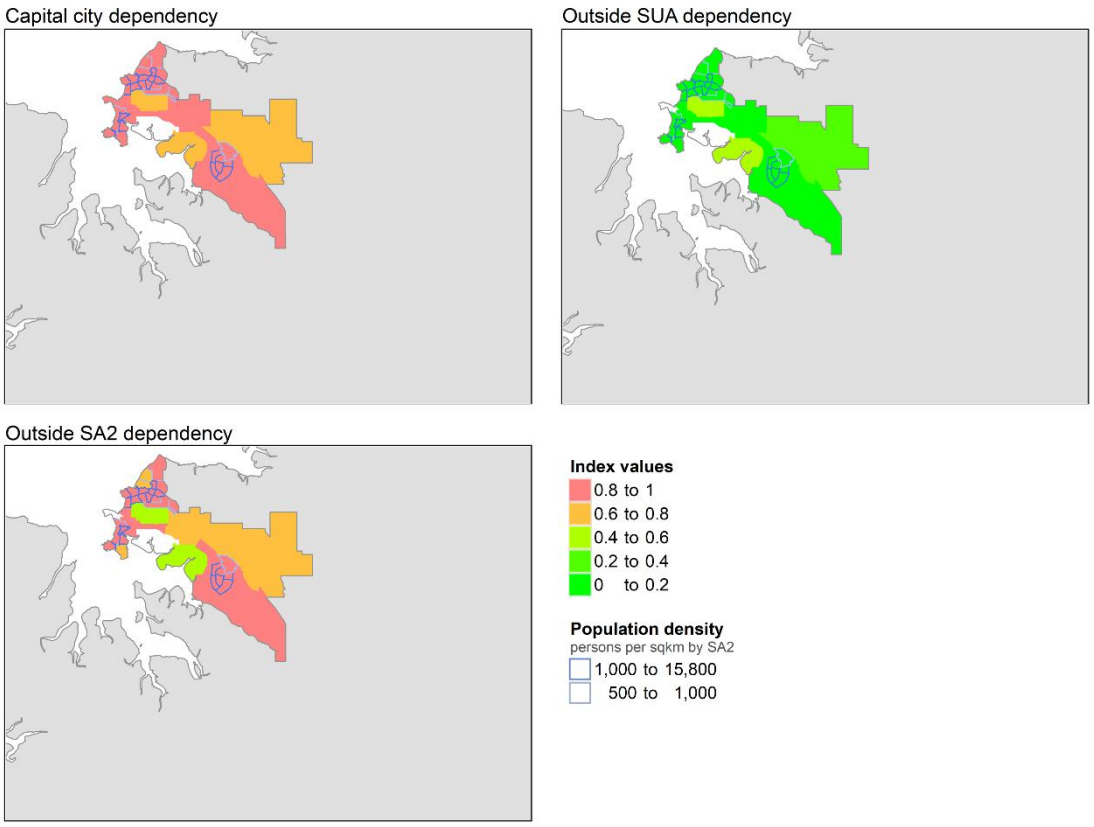


Figure 2.9: Darwin indices

Data source: Synergies map

2.4 Proxy variables

In regression analyses, proxy variables can replace a primary variable in cases where the primary variable cannot be quantified, or insufficient observations are available. The idea is that a variable (or combination of variables) that can be expected to be highly correlated with the primary variable (or a certain key aspect of it) can be used to depict the effect of the primary variable in a regression analysis. This way it effectively still enters the regression.

Traffic volumes

The 2015 Review found that net expenses per capita tend to increase as city size increases. It considered the reasons for this were the greater quantity of travel per capita made by public transport. Figure 2.10 illustrates this relationship for the eight capital cities.

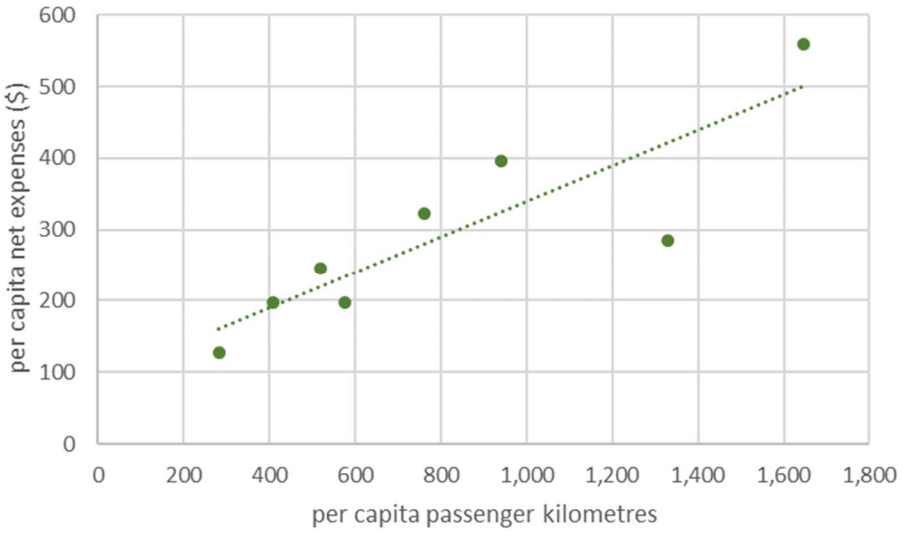


Figure 2.10: Net expenses vs. transport task by capital city, per capita average of 2009-10 to 2011-12

Data source: 2015 Review of State Government Subsidised Urban Public Transport Services

The above relationship uses three primary variables: net expenses, passenger kilometres and population, where population enters the analysis as a scaling parameter for the other two variables. The finding suggests that a model using population and passenger kilometres should be able to fit the net expense data well. In other words, population and passenger kilometres should be key candidate variables. The Stage 1 report refines this point and suggests including passenger kilometres by public transport mode as an explanatory variable because different modes tend to have different cost structures.

While population is readily available from the Census 2016 for all 101 SUAs[14] in Australia, passenger kilometre estimates are only available for the eight capital cities[15]. This means that using passenger kilometres as a candidate explanatory variable would reduce the maximum sample size from 101 to 8. This would make a multivariate regression analysis effectively meaningless as the number of variables would be too similar to the number of observations. Therefore, passenger kilometres as published by BITRE are not a viable candidate explanatory variable.[16]

The variable *passenger kilometres* by mode covers three components of the public transport task:

- It is an indicator for supply as it indicates the presence of a certain mode.
- It quantifies demand as the number of passengers is one of the inputs to its calculation.
- It captures aspects of the complexity and length of the network as it measures the (average) distance travelled by each passenger.

A set of proxy variables able to comprehensively depict the effects of passenger kilometres travelled on net expenses consequently needs to depict the demand for a certain mode and the distance travelled. The supply aspect is less relevant in this context as it is fair to assume that if a reliable source indicates demand for certain mode, this mode is actually present in the respective area. The Census 2016 contains data that can be used to depict both of these aspects:

- Demand for a certain mode in an SUA can be extracted from place of usual residence database.

[14] The ABS defines 101 SUAs. There are five SUAs that cross state borders, but these are each counted as a single SUA.
[15] *Australian Infrastructure Statistics Yearbook 2016*; Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2016
[16] At the time this report was prepared, the Team was not aware of any other sources that publish reliable and publicly available estimates of passenger kilometres travelled at a more granular level than the BIRTE reports.

- As an indicator for trip length, the average distance to work by SUA can be obtained from the same database.

As both components can be extracted from the Census, they will be available for all 101 SUAs and hence their inclusion will not result in the loss of observations. Furthermore, the effects of demand and network length can now be tested separately using these proxy variables which will make the modelling results more transparent. This means that the use of this set of proxy variables not only allows for the inclusion of the effect of passenger kilometres in the regression, but it could also improve the quality of the model.

For the purposes of this study, we will test the explanatory power of the above set of proxy variables on net expenses.

Congestion

Being a key contributor to bus operating costs, congestion should be considered as a candidate variable. However, just like passenger kilometres travelled, estimates for congestion are only available for the eight capital cities.[17] This means that using a congestion measure as a candidate explanatory variable would reduce the maximum sample size from 101 to 8. This would make a multivariate regression analysis effectively meaningless as the number of variables would be too similar to the number of observations. Therefore, the congestion measure as published by BITRE is not a viable candidate explanatory variable and a proxy needs to be identified.[18]

Congestion occurs when roads approach their capacity, i.e. when many people use similar routes at the same time. Hence, one likely trigger for congestion could be that many people live in a certain area and leave their homes at a similar time. This means that one potential simple proxy variable for congestion could be population density. In particular, we have examined population-weighted density (see Section 1.1) given its advantages over more conventional measures of density. Density has the potential to solve the data availability issue associated with congestion as it can be easily derived from the Census 2016 which contains data for all 101 SUAs in Australia.

Figure 2.11 plots the population density of the eight capital city SUAs derived from the Census 2016 against the avoidable congestion costs per capita published by BITRE.

[17] *Information sheet 74: Traffic and congestion cost trends for Australian Capital Cities*; Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2015

[18] At the time this report was prepared the Team was not aware of any other sources that publish reliable and publicly available estimates of congestion at a more granular level than the BIRTE reports.

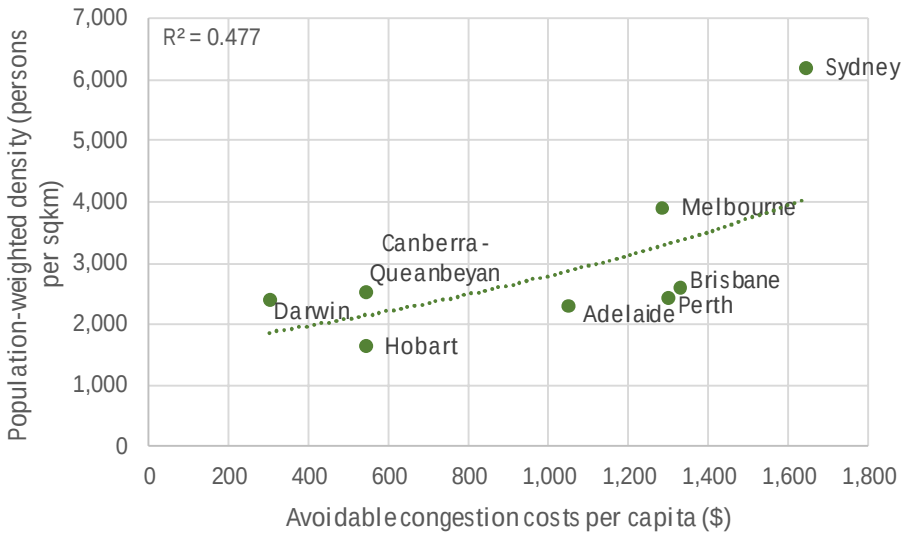


Figure 2.11: Avoidable congestion costs per capita vs. population density

Data source: ABS Census 2016 and BITRE

The figure visually suggests a weak relationship between the two variables, confirmed by the low R^2 of 48%. This means that there might be better proxy variables for congestion than population density warranting further investigation.

It appears reasonable to expect that total (avoidable) congestion costs also increase with population. The plot above suggests that this is also the case for per capita congestions costs as the datapoints of avoidable congestion costs seem to follow the population sizes of the cities: Canberra, Hobart and Darwin are on the left and Sydney and Melbourne on the right extremes of the plot. This suggests that there might be a close correlation between population and congestion.

Figure 2.12 compares population of the eight capital city SUAs with the avoidable congestion costs per capita published by BITRE.

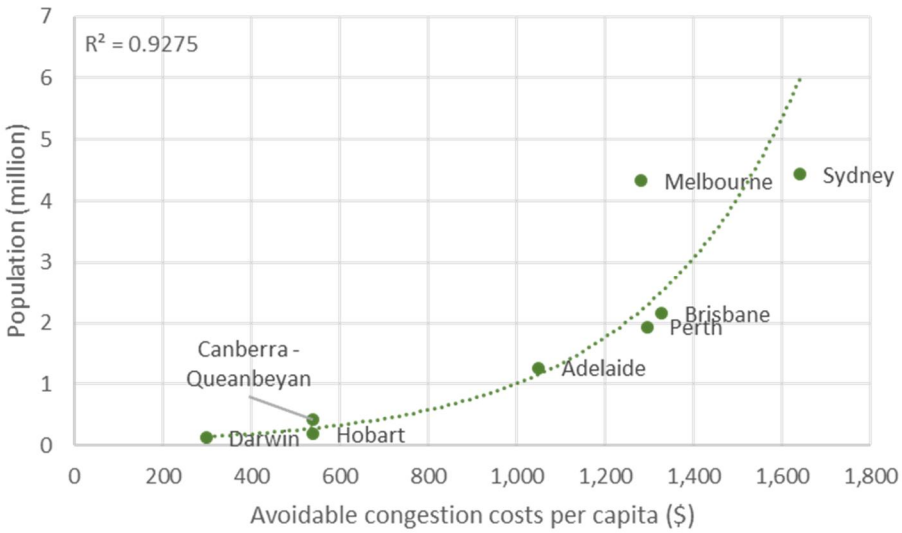


Figure 2.12: Avoidable congestion costs vs. population

Data source: ABS Census 2016 and BITRE

The plot shows that all points lay on or close to the trend line indicating a close correlation between population and per capita congestion. The R^2 of 93% statistically confirms this visual impression. This finding means that in a regression analysis, population will accurately depict the effects of congestion and vice versa. However, we will test models in which population enters the regression as part of the dependent variable and can therefore not also appear in the regression as an explanatory variable (see Section 3.2 for a detailed discussion).

Based on the definitions presented in the analytical framework in Section 1.1, population is a volume measure. Alternative volume measures that could appear as explanatory variables in the regression analysis are the count of employees (by place of work) or the public transport passenger discussed above. Intuitively, both variables could be closely correlated with congestion:

- Just like the population variables, employment by place of work is a measure of the number of persons using similar routes at the same time and hence an indicator for road utilisation.
- The number of public transport passengers is likely to increase as people (commuters) seek alternative transport means as a reaction to congestion on the roads. While, for simplicity, Figure 2.13 below presents the total number of public transport passengers, this measure will enter the regression analysis as four mode specific variables, that is as bus, train, tram and ferry passengers.

Figure 2.13 compares the count of employed individuals (top) and public transport passengers (bottom) in the eight capital city SUAs with the avoidable congestion costs per capita published by BITRE.

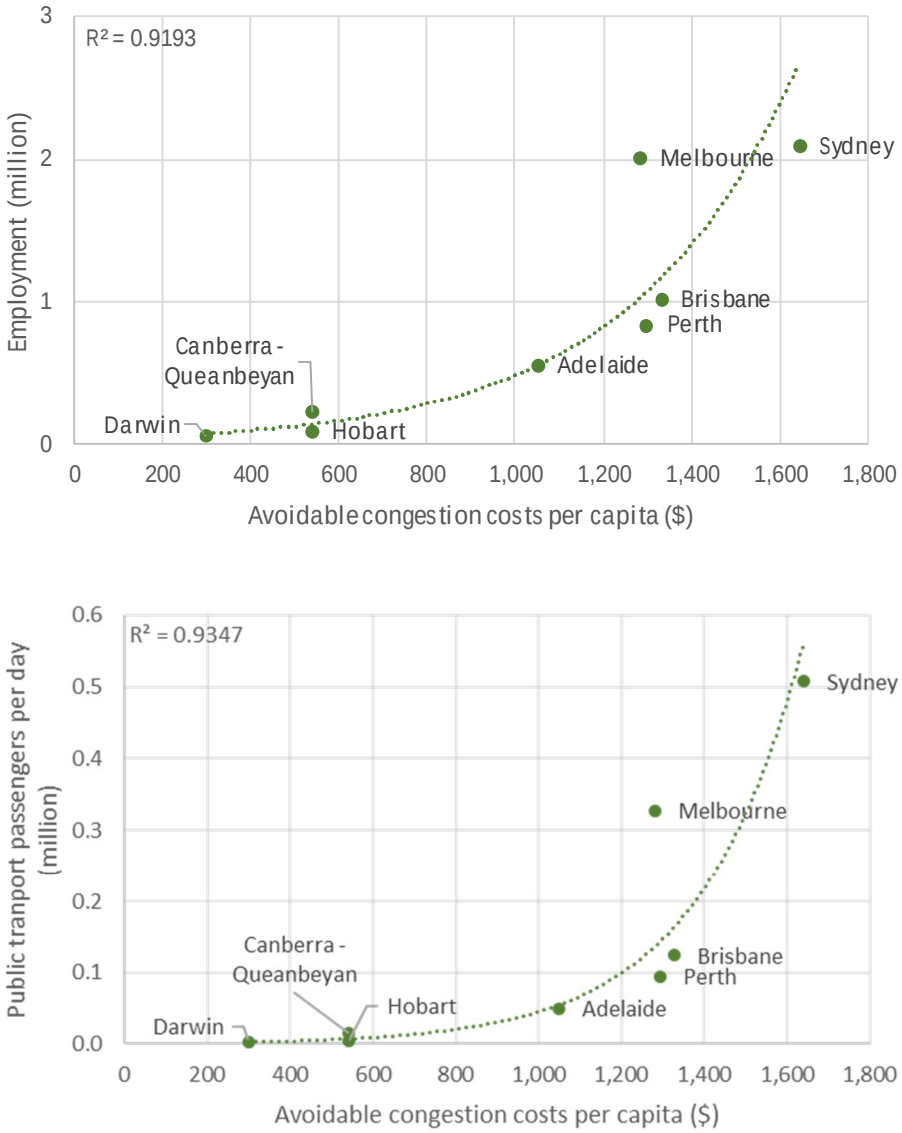


Figure 2.13: Avoidable congestion costs vs. employment (top) and public transport passengers (bottom)

The plot shows that for both variables all points lay on or close to the trend line indicating a close correlation between them and per capita congestion. The R^2 of, 92% and 93% respectively, statistically confirm this visual impression. This finding means that in a regression analysis, both the number of employed persons in an SUA and a count of public transport passengers can serve as proxies to depict the effects of congestion.

Since all three variables are available for all SUAs and congestion only for the eight capital cities, we consider the inclusion of congestion as a candidate explanatory variable of expenses unnecessary. In the case where population indirectly enters the regression, such as when expenses per capita is the dependent variable, then one of the proxy variables presented above could represent congestion.

3. Recurrent expenditure

Net expenses will be used as the dependent variable representing recurrent expenditure. If not explicitly stated otherwise, *expenses* and *net expenses* are used interchangeably in this report, as each term is taken to refer to the same thing.

The majority of the expenditure data to be used in this analysis were reported directly from the States to the CGC who then provided it to the Team. In jurisdictions where public transport is (partly) operated by private contractors, it has been difficult to obtain the total operating expenses. Since private operators consider their cost and revenue figures commercial in confidence, only the contractor fee (i.e. the government expense) can be directly observed. Therefore, to achieve a consistent basis for the expense estimates, in the jurisdictions that do not rely on contractors, revenue needs to be deducted from the total expenses. The resulting “net-expenses” will be broadly equivalent to the contractor fee paid by government to contractors and hence constitute an adequate basis for comparison across all jurisdictions.

Not all data have been derived in the same way. We understand that some States extract their data from annual reports while others use internal financial databases or ABS data. Using different data sources could result in inconsistent estimates across States. This means that data consistency is one of the major challenges of this assessment and we consider developing a thorough understanding of the expenditure dataset a crucial first step towards estimating a robust model.

For this purpose, this section first presents an overview of the ways in which data were collected and the potentially associated issues. It then discusses the key challenges of working with derived data from a statistics point of view before suggesting ways of overcoming these challenges. The section concludes by proposing and describing the dataset to be used for the dependent variable.

3.1 Overview of expense data

The expense data collected from the States for this project are the result of extensive consultation by the CGC with the States. We understand that in several instances data could not be collected at the SUA level. When this was the case, a range of techniques were applied to apportion regional expenditure to the SUA level.

While deriving data is often unavoidable, it comes with its challenges. In some cases, it is sufficient to be aware of the shortcomings and their expected effects while in others it can improve the quality of the analysis if the associated datapoints are excluded. How to deal with the issues depends on how exactly the data were derived and on the independent variables selected for the regression analysis. The following summarises our understanding of how SUA data were derived by/for each State.

- **New South Wales**
  Expense data were reported by SUA. The Team understands that rail expenses for the SUAs of Sydney, Newcastle, Central Coast and Wollongong were derived by the State from two expenses figures for NSW Trains, which services Newcastle and Central Coast, and Sydney Trains which services Sydney and Wollongong. CGC staff reallocated this estimate to the four SUAs using data from the household travel survey.

- **Victoria**
  Bus expense data were provided at State level. Where available, they were split across SUAs based on their subsidy shares. For the remaining SUAs, values were estimated by the CGC using the results of a population-based regression analysis. Metropolitan train expenses were provided at SUA level.
- **Queensland**
  Bus expense estimates are available for eight SUAs in Queensland. For the remaining SUAs, values were estimated by the CGC using the results of a population-based regression analysis. Metropolitan train expenses were estimated for the entire South East Queensland network that – like in NSW and Victoria – crosses SUA boundaries. Here, Census Journey to Work data was used to allocate expenditure to SUAs.
- **Western Australia**
  Bus expenses were provided by SUA. Metropolitan train expenses were estimated for the entire Perth network which already includes expenses associated with the Yanchep SUA. As Yanchep constitutes a satellite to Perth, these SUA will be treated as one and expenses do not have to be split.
- **South Australia**
  All expenses were provided at SUA level.
- **Tasmania**
  Expenses (bus only) were provided at State level and allocated to SUAs based on the number of boardings for Hobart, Launceston and Burnie and using population shares for Devonport and Ulverston. These data were derived in different ways because services in Hobart, Launceston and Burnie are provided by the (government operated) PNFC METROTAS while the Devonport and Ulverston operations are managed by privately owned Merseylink.
- **Northern Territory,**
  Expenses (bus only) were provided by SUA.
- **ACT**
  Expenses (bus only) were provided by SUA.

The next two sections explore the implications of the presented estimation methods from a statistical point of view. The final section will combine the findings of all three sections to develop a proposed dataset for modelling.

3.2 Challenges with derived data

Some of the expenditure data has been derived by allocating regional/state estimates to SUAs using population-based regression analysis. This can create issues in further econometric modelling if population is included as an explanatory variable.

If population is first used to derive expenditure data (or any other dependent variable) and then it is used as an explanatory variable for the thus derived data, the regression effectively regresses population on population. While this will produce a statistically significant model, the knowledge gain is somewhat limited. Furthermore, if only a subset of the expenditure data is estimated using population, the estimated relationship will be skewed towards the regression used for estimating these points as – since they are derived by one - they will be well represented by a functional relationship. This will particularly cause issues if a linear function is used to derive datapoints and alternative (non-linear) functional forms such as a linear-log form are used in the regression model.

The second issue relates to deriving values for the dependent variable in general. Any regression model will not perfectly reproduce the data used to estimate it. (If it did, there would be no need for statistical estimation). In other words, it will generate an estimation error. Formally:

$$\hat{y}_i = \widehat{\alpha_0} + \widehat{\alpha_1} x_i + \varepsilon_i \qquad \text{(eq. 3.1)}$$

Where x_i is the independent variable, $\hat{\alpha}$ is the estimated coefficients and ε_i is the estimation error. If $\hat{y}_i$ is only an intermediate estimate to fill a dataset and z the variable that is to be modelled with it, it follows that:

$$\hat{z}_i = \widehat{\beta_0} + \widehat{\beta_1}\hat{y}_i + \delta_i \qquad \text{(eq. 3.2)}$$

Since $\hat{y}_i$ is itself an estimated variable, it can be replaced with its regression equation. The resulting term includes two error terms (ε_i and δ_i):

$$\hat{z}_i = \widehat{\beta_0} + \widehat{\beta_1}\widehat{\alpha_0} + \widehat{\beta_1}\widehat{\alpha_1}x_i + \widehat{\beta_1}\varepsilon_i + \delta_i \qquad \text{(eq. 3.3)}$$

Crucially, the second error term cannot be measured as it is hidden in the estimates for $\hat{y}_i$ and hence the statistical tests on $\hat{z}_i$ will be misleading. Therefore, it is always preferable to directly formulate a regression equation with non-derived data. In this case this would be:

$$\hat{z}_i = \widehat{\alpha_0} + \widehat{\alpha_1}x_i + \varepsilon_i \qquad \text{(eq. 3.4)}$$

This equation has only one error term that can be measured. This means all statistical tests work and the results can be assessed for any potential inaccuracies and biases with confidence.

For these reasons we will attempt to use as little derived data as possible. The next section will show that, as long as a robust model can be estimated, using a data sample will enable us to generate robust estimates for most if not all SUAs.

3.3 Using a representative sample as dependent variable

This section explores the possibility of using the SUAs for which the robust data are available as a sample of the population of all SUAs. The model estimated with this sample can then be used to generate estimates for the SUAs for which data are not available. This is a standard statistical approach. Its appropriateness depends on the representativeness of the available sample, in this case the SUAs for which robust information is available.

Broadly, a (regression) model can generate three types of estimates:

- Within sample predictions
- Out of sample within range predictions
- Out of sample and out of range predictions

The quality of within sample predictions can be assessed simply by comparing them to the corresponding observed values. This is why within sample predictions are often used to assess the quality of the model's fit.

Out of sample within range predictions refer to values that do not have a direct correspondence in the dataset but for which the independent variable values for the estimate fall into the range of the observations used for the estimation. For example, if population is the only independent variable and the SUAs used for the regression estimate have population sizes between 20,000 and 4.5 million inhabitants, an expenditure estimate for an SUA with 50,000 inhabitants that had not provided expenditure would be out of sample but within range.

Since they can be related to observed values, the quality of out of sample within range predictions can be assessed to a certain degree even though no directly comparable data is available. Potential outliers can of course not be appropriately reproduced. However, the ability to anticipate outliers is of limited importance in this context as the model is intended to generate the appropriate average expenditure levels. Therefore, out of sample within range prediction can be used with confidence.

Out of sample out of range predictions refer to values that do not have a direct correspondence in the dataset and for which the independent variable values for the estimate do not fall into the range of the observations used for the estimation. For example, if population is the only independent variable and the SUAs used for the regression estimate have population sizes between 20,000 and 4.5 million inhabitants an expenditure estimate for an SUA with 10 million inhabitants that had not provided expenditure would be out of sample and out of range.

Out of sample out of range predictions can be reliable if the underlying functional relationships is linear. Otherwise, accuracy can be very low as the following example will demonstrate.

Assume the illustrative dataset presented in the previous chapter has been adjusted for policy neutrality and two smaller SUA were added to the sample. Table 3.1 presents this dataset. The dataset suggests that there are significant scale effects: For the two small SUAs, expenses are just less than half those of the three large SUAs while population is only about one tenth.

Table 3.1: Representative sample: illustrative dataset

SUA	Expenditure (*exp*)	Population (*pop*)
	$ million	million persons
1	1,200	4.600
2	1,210	4.606
3	1,320	4.670
4	500	0.500
5	350	0.300

Figures for illustrative purposes only

A linear-log regression fits the dataset well as the high R^2 and the predicted values presented in Table 3.2 below demonstrate. To illustrate the effects of sampling, the regression was repeated seven times where the first five iterations drop one observation and the last two drop two observations.

The top part of Table 3.2 shows the estimated coefficients and the associated $R^{2.}$ The models' predictions are presented in the bottom part. The estimated values for the dropped values are highlighted.

Table 3.2: Representative sample: models and predictions

	All SUA	Without SUA 1	Without SUA 2	Without SUA 3	Without SUA 4	Without SUA 5	Without SUA 1 & 2	Without SUA 4 & 5
Intercept	721	729	727	708	713	748	748	-10,942
Coefficient	341	350	348	326	346	332	372	7,956
R^2	0.99	0.99	0.99	0.99	0.99	0.98	0.99	>0.99
SUA	**Expenses ($ million)**							
1	1,242	**1,263**	1,259	1,206	1,242	1,255	**1,316**	1,200
2	1,243	1,264	**1,259**	1,206	1,242	1,256	**1,316**	1,210
3	1,247	1,268	1,264	**1,211**	1,247	1,260	1,321	1,320
4	485	486	486	482	**473**	518	491	**-16,457**
5	310	308	308	315	296	**348**	301	**-20,521**

Estimated model: $\boldsymbol{exp_i = \beta_0 + \beta_1 * ln(pop_i)}$

Figures for illustrative purposes only

The table shows that all five models based on a sample with only one dropped observation predict this value reasonably well. Even if two of the three large SUAs are dropped, the predictions are not too far off. The coefficient estimates are similar for all six models. Only the example in which both small SUAs are dropped produces very different regression results and diverging predictions. Figure 3.1 illustrates the two diverging trend lines.

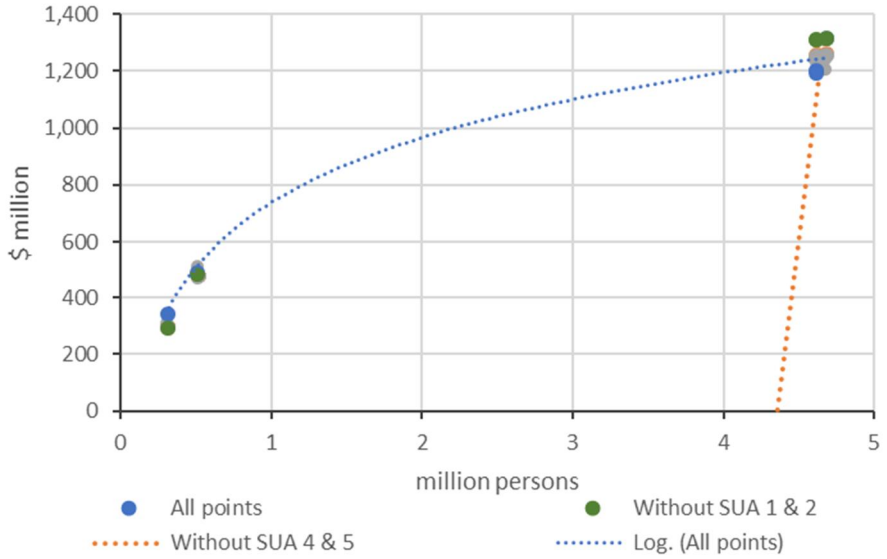


Figure 3.1: Representative sample: graphical illustration

Figure for illustrative purposes only

This example illustrates that a non-linear model is likely to accurately reproduce and predict values that lie out of sample and within range. However, results can be misleading if points at one end of the spectrum (e.g. both small SUAs) are missing. For the modelling, this means that as long as the independent variable values for SUAs with available data cover (a large share of) the variance of the independent variables values across all SUAs, predictions of the missing expense values can be expected to accurately represent typical levels. In other words, a complete dataset is not required, because with a sufficiently diverse sample of SUAs a regression model can be estimated that can be expected to reliably predict the missing values.

The following section will propose an expense dataset to be used for the modelling.

3.4 Proposed dataset for modelling

In total, the ABS defines 101 SUAs. The five SUAs that persist across State boundaries are treated as a single datapoint by the ABS, and we follow this approach in the modelling. The estimates can then be apportioned to the States based on population or travel shares.

A comprehensive datapoint for an SUA must represent expense data for all key public transport modes present in the SUA. In this context, key public transport modes comprise buses, trains, ferries and light rail. If a mode is not present in an SUA, the expenses associated with this mode are zero and the SUA total will only consist of those that are present. A datapoint is also considered comprehensive if total expenses are reported as zero.

Table 3.3 presents an overview of the expense data by state. A table showing data availability and quality by SUA can be found in Appendix F.

Table 3.3: Data overview by state

State	SUAs	Complete datapoints	Excluded datapoints	Reason for exclusion
NSW	35	31	4	Expense data not reported for 4 SUAs
Vic	21	7	14	Derived using population
Qld	18	8	10	Derived using population
WA	11	10	1	Yanchep is a satellite to Perth
SA	8	8	0	
Tas	5	3	2	Derived using population
ACT	1	1	0	
NT	2	2	0	
Total	**101**	**70**	**31**	

Source: Synergies analysis of State data collected by the CGC

Table 3.3 shows that for 26 of the 101 SUAs, expense data for at least one mode was derived using population. Since population (or a closely related variable) is expected to be one of the key variables in the regression analysis, we consider the usability of these values low and will exclude the associated datapoints. For 4 SUAs in New South Wales, expense data was not reported, and therefore these SUAs have been excluded. Finally, in one case, an SUA (Yanchep) is considered a satellite to a capital city. This means that, in total, data is available for 70 SUAs. Together these SUAs cover 96.50% of Australia's urban population (see Appendix F) indicating that the sample of the 70 SUAs is very likely to be representative for all 101 SUAs. Figure 3.2 presents the 70 datapoints as boxplots on a linear and on a logarithmic scale (see Appendix A for a summary of what such plots show and how they can be interpreted).

The right boxplot in Figure 3.2 shows that on a logarithmic scale the values are spread relatively equidistantly over the entire range. Under this transformation they are relatively evenly distributed.

Values range from $20,000 for the SUA of St Georges Basin-Sanctuary Point to $3.6 billion for Sydney. As could be expected, the five major capital cities dominate the expenses. With $1.3, $1.1, $0.7 and $0.3 billion for Melbourne, Brisbane, Perth and Adelaide respectively, the other four cities reported significantly lower expenses than Sydney, however. Overall, the distribution is somewhat skewed by these five values as indicated by a median of $0.8 million and an average of $114 million.

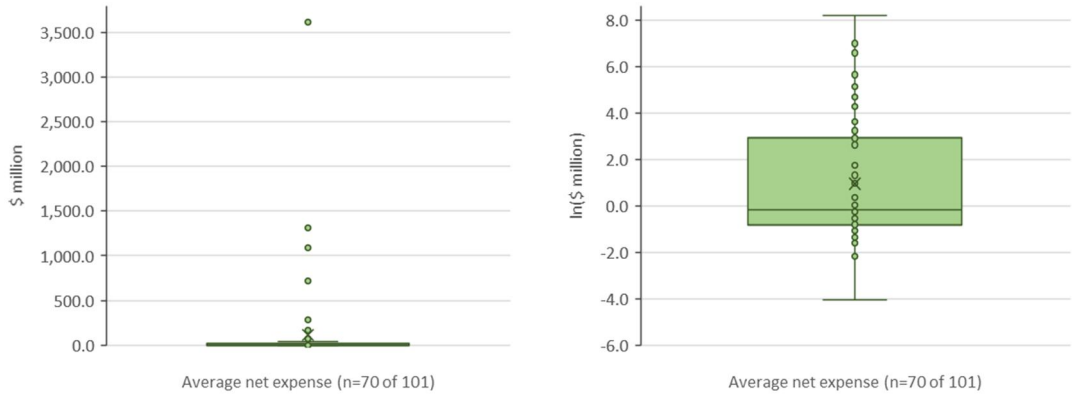


Figure 3.2: Distribution of average net expenses

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

As mentioned in Section 2.4, the 2015 Review based its analysis on expenses per person; i.e. SUA net expenses were divided by the SUA population. In order to produce comparable results, we will replicate this approach and consider expenses per person[19] as a dependent variable. This might also reduce the skew of the distribution as there is clearly a correlation between expense and the population size of the SUA.

Figure 3.3 presents the 70 datapoints as boxplots on a linear and on a logarithmic scale. Values per person range from $1 for the SUA of St Georges Basin-Sanctuary Point to $815 for Sydney. While this is still a considerable range and the sample is still dominated by the major capital cities, the maximum expense per person is only about 800 times larger than the minimum with a relative standard deviation of 146. The maximum net expense is 180,000 times larger than the minimum with a relative standard deviation of 478.

Bringing the extremes of the sample closer together will make fitting a function using regression analysis statistically less complex and is thus likely to produce a more robust model. The trade-off, however, is that

[19] In order to ensure consistency with the expense estimates, population was estimated as the average of the same three years as expenses.

population becomes redundant as a candidate explanatory variable as it would appear on both sides of the equation (see Section 3.2).

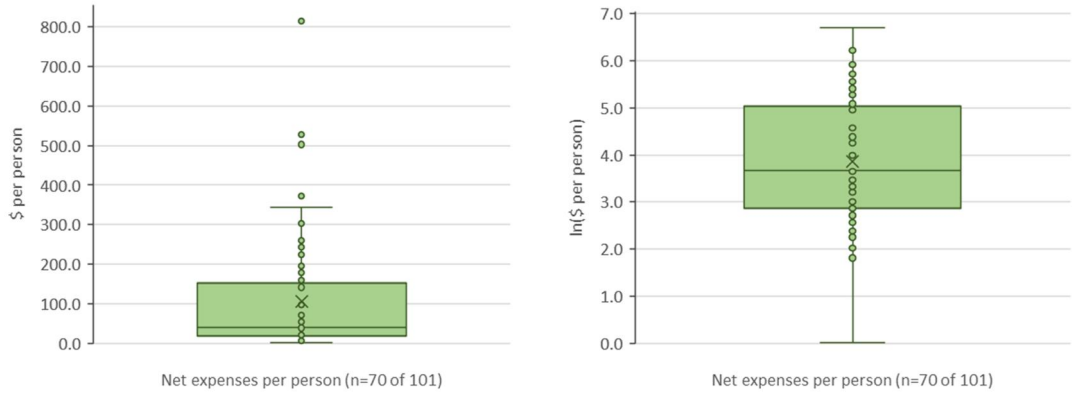


Figure 3.3: Distribution of average net expenses per person

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

4. Infrastructure expenditure

The Team understands that most States expressed reservations with or opposition to using the current investment model. The key reasons are:

- The model is considered too simple.
- The data are considered not sufficiently reliable.
- There were considered to be too few observations.

The dataset provided to the Team contained datapoints for 19 of 101 SUAs. Half of these were SUAs in Western Australia. We therefore share the States' concerns and will suggest an alternative modelling framework for infrastructure expenditure in this section.

4.1 Key Stage 1 report findings

The Stage 1 report suggested developing a modelling framework based on a sound theoretical underpinning as a way of addressing these concerns. The development of a statistical model to represent this model could then deal with the lack of data (both observations and variables) in a transparent way that would make necessary deviations from the theoretical model clear.

The Stage 1 report suggested that this theoretical framework should consider the following:

- **Maintenance of the appropriate capacity to meet ongoing travel demand**
  This reflects the fact that existing physical assets need to be maintained, refreshed or replaced at the end of their economic life. Indicators for this need may include population or value of the existing asset base.
- **Ability to build new capacity to meet the increased demand**
  Increased transport demand is driven by population growth, employment growth and employment patterns, economic activities and household income. For example, higher employment means more journeys to work, and people with higher incomes tend to travel more than those with lower incomes.
- **Ability to meet peak hour demand**
  Transport capacity is designed to meet peak demand. Employed persons typically travel in the AM or PM peaks which puts additional demand on urban transport capacity. Reliable time of day travel data is only available in Sydney, Brisbane, Melbourne and Perth. Thus, the peak demand data is not likely to be used in the Stage 2 model.
- **Capture of construction cost adjustment**
  The construction cost for one track kilometre in different capital cities and urban centres will be different due to a range of factors such as terrain, land value and utility adjustment. Ideally, the unit cost in each urban centre can be used as an indicator of the construction difficulty factor, but such data is unlikely to be available. In recent years, tunnels and bridges have been selected as engineering options. The indicator of transport construction unit cost can be the intensity of structures, bridges and tunnels. CGC has obtained urban waterways and bridge statistics that could potentially be used as a variable representing urban construction cost.

As noted above, having a theoretical framework allows for a better understanding of the way data limitations influence the ultimate "model" recommended. The Stage 1 report pointed this out, anticipating that the "recommended model will be different to the theoretical model". In particular, the Stage 1 report appears to anticipate that reliable investment data cannot be collected from the States.[20]

The remainder of this section will show that given the unresolved data availability issues, it is neither possible nor necessary to estimate a second separate model for investment that follows the above framework. Recurrent expenditure and investment are highly correlated and hence all these variables can be incorporated in a model of recurrent expenditure.

[20] See Stage 1 Report Table 3.2: Dependent and explanatory variables for an urban transport infrastructure expenditure model

4.2 Data availability

The investment dataset provided to the Team contained datapoints for 19 of 101 SUAs, of which nine are from Western Australia. Data is available for all of these SUAs for the financial years between 2013-14 and 2016-17. While some States appear to not have provided data for 2017-18, others provided forecasts out to 2020-21.

Figure 4.1 below presents the distribution of the average investment values between 2013-14 and 2016-17 for the 19 SUAs. Similar to the expenses dataset, the investment values are dominated by the major capital cities. Overall, values are spread between $100,000 and $3.3 billion with a mean of $368 million and a median of $22 million.

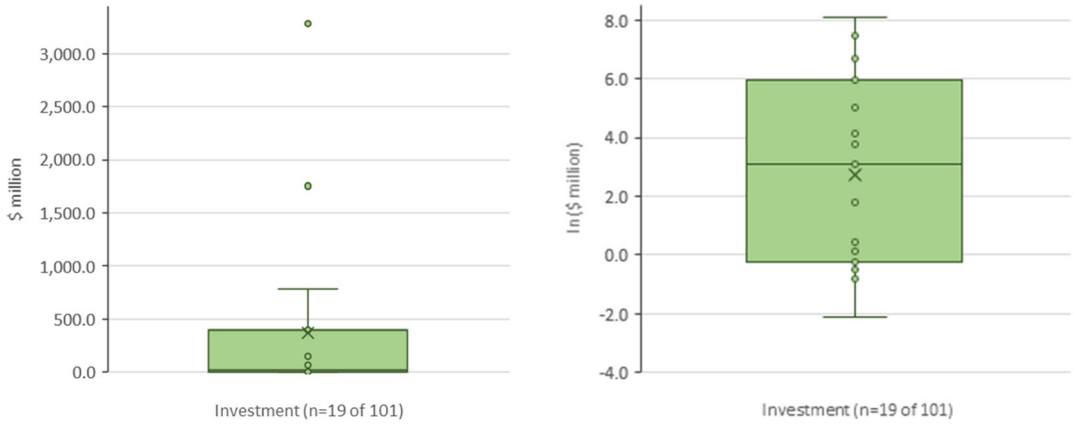


Figure 4.1: Distribution of average investment

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Based on the very wide range of values, the Team considers that the number of observations might be insufficient to estimate a robust model: It appears unlikely that a single-variable model will be able to establish a meaningful functional relationship for the entire range and there are too few observations to estimate (and test) a multi-variable model with confidence.

Considering the relationship between expenses and investment could lead to a solution to the issue: If robust evidence can be found that investment and recurrent expenditure are structurally correlated, it would be sufficient to estimate a single model based on recurrent expenditure that accounts for all key cost factors affecting State transport expenditure.

The next section presents such evidence both from a theoretical and empirical perspective.

4.3 Linkages between investment and recurrent expenditure

Investment in public transport can be generally thought of as addressing three broad objectives:

- Maintenance of capacity
- Expansion of capacity
- Quality and efficiency improvements.

Investment in maintaining capacity allows the network to continue to handle the existing passenger demand. The higher the network capacity and demand, and the greater the utilisation, the higher we expect annual maintenance investment costs to be. Capacity, patronage and utilisation will be reflected in the passenger and

vehicle kms produced on the network. Operating costs can be correlated with passenger and vehicle kms produced.

Expansion of capacity can include investment in new rolling stock, new systems for handling passengers, expanded track, bus lanes etc. Some investments allow greater passenger volumes to be handled leaving the underlying network (e.g. rail track kms) unchanged. Other investments allow for greater patronage by expanding the underlying network (new rail track, extended bus road network). In the former, passenger and vehicle kms will increase. In the latter, passenger and vehicle kms will also increase. We expect that as passenger and vehicle kms increase, total operating costs increase.

Quality and efficiency improvements can be associated with enhancing the existing network, the expanded network or both. To the extent that investment is aimed at improving efficiency it will, all other things equal, tend to lower per unit costs. To the extent that it is aimed at quality improvement (air conditioning, timetable reliability, reduced waiting times, frequency) it will tend to increase per unit costs.

Overall there is good reason to expect that operating and capital costs are correlated for a system in "equilibrium" – maintaining services, maintaining utilisation, meeting demand growth as required etc. Hence, in the absence of comprehensive data on capital investments by SUA, an approach based on an association between capital and operating costs is a likely way forward.

There is evidence supporting this. For example, the American Public Transport Association annual factbooks report on a large number of public transport cost and performance indicators for US and Canadian public transport systems. They show a very stable relationship between total operating costs and capital costs over time and systems. Between 2001 and 2015 annual urban bus system operating costs varied from 78% to 82% of total operating and capital costs. For heavy rail systems the range was 49% to 60%. For light rail is was 26% to 32%.[21]

This is also supported in the data provided by the States. The distributions of the investment values (Figure 4.1) and that of the expenses (Figure 3.2) are very similar. In fact, with a correlation coefficient of 0.98, the 19 data points for which investment data are available are very highly correlated with their corresponding values from the expense dataset. Figure 4.2 illustrates this relationship by juxtaposing the two datasets in a scatterplot.

[21] *2017 Public Transport Factbook 68th Edition.* American Public Transport Association. Washington 2018. Appendix A. Historical Tables

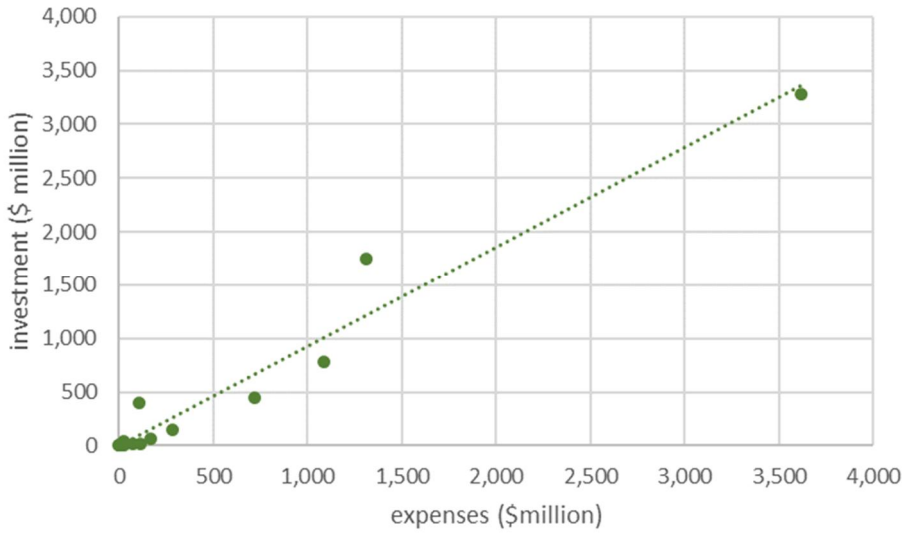


Figure 4.2: Average investment volumes vs. average recurring expenses by SUA, 2013-14 to 2016-17

Data source: Synergies analysis of State data collected by the CGC

Considering the close correlation, it appears very likely that, if a sufficiently large dataset were available for both, an investment and an expense model would generate very similar results. They would consequently lead to similar funding allocation outcomes. Therefore, the Team is of the view that in this instance one model is better than two. It appears – because of the issues discussed at the start of this section – that the investment model would be more likely to generate misleading results than to add insights.

In short, taking the theoretical and empirical evidence presented here into account, we recommend using a single expense model that accounts for all key cost factors as the sole basis of the funding allocation mechanism.

5. Econometric analysis: summary

Using the theoretical framework developed in the introduction, to this point the report has used statistical analysis to identify the most suitable measures for the influencing factors identified in the framework. Having also developed an understanding of their value ranges and data quality, we can now be confident that the coefficient estimates produced by the subsequent regression analyses can be thoroughly tested for their significance and hence the estimated relationships can be expected to be causal and not coincidental.

From a data perspective the limiting factor in the regression analysis is the availability of observations for the dependent variable. While most independent variable data are readily available or only require minor modifications, in several instances reliable expenditure data has simply not been collected at the required level of detail. We have identified a dataset covering 70 of the 101 SUAs that contains net expense estimates that can be used with confidence and – more importantly – is likely to be representative. As discussed in Section 4, since investment data is only available for 19 SUAs and analysis suggests that a separate investment model should lead to similar funding allocation outcomes as one for expenses, we only estimate an expense model that will form the sole basis of the funding allocation mechanism.

The assessment of the candidate independent variables found that where comprehensive data is available the distribution of values indicates that the sample of the 70 SUAs for which reliable expense data is available is very likely to be representative for all 101 SUAs.

In the course of the analysis so far, we have identified the key issues or methodological challenges and developed the resolutions presented in Table 5.1.

Table 5.1: Identified issues and proposed resolutions

Topic	Issue	Resolution	Details in
Policy neutrality	Remove the effect of policy related factors on subsequent funding shares	Include policy related variables in the modelling as the ability to control for policy measures will ensure more robust and transparent estimates than introducing a bias to the model by omitting potentially key explanatory variables.	Section 2.1
Geographic framework	At least partly due to the misalignment between the currently used statistically motivated boundaries (SUA) and the administrative boundaries of transport networks, the amalgamation of principal and satellite cities has been an ongoing discussion topic.	The self-sufficiency indices are based on revealed travel preferences and have provided key insights regarding the extent to which an area depends on other areas to provide jobs for its workers. The indices help determine whether there is a case for defining an area as a labour market integrated satellite city.	Section 2.2
Economies of Scale	2015 Review findings provide evidence for diseconomies of scale (a larger population base requires more spending per capita). Meanwhile, States have provided a case for economies of scale (a larger population base requires lower spending per capita).	Economies of scale can be modelled with specific proxy variables, such as congestion or mode shares, which describe expense patterns. Alternatively, it can be accommodated through the choice of functional form by using logarithms. We have applied both approaches in our analysis.	
Consideration of passenger kilometres travelled and congestion measures as candidate variables	Insufficient observations of these two variables	Consideration of passengers by public transport mode and average distance to work as proxy for mode specific passenger kilometres travelled and congestion	Section 2.4 and the Executive Summary

Topic	Issue	Resolution	Details in
Expenses data availability	When data could not be collected at the SUA level, a range of techniques were applied to apportion regional expenditure to the SUA level.	Analysis indicates that the sample of the 74 SUAs for which robust data is available is very likely to be representative for all 101 SUAs. We will base the modelling on this sample.[22]	Chapter 3
Investment data availability	Most States expressed reservations with or opposition to using the current investment model	Since investment data is only available for 19 SUAs and analysis suggests that a separate investment model should lead to similar funding allocation outcomes as one for expenses, we will only estimate an expense model that we recommend will form the sole basis of the funding allocation mechanism.	Chapter 4
Asset quantity data quality and availability	The quantity of some assets is either understated or missing for certain SUAs.	Where asset quantity data is questionable, we will model these factors as dummy variables, which simply indicate the presence or absence of a particular mode of transport for each SUA. In light of the relative paucity of the data, we will only use basic asset quantity measures such as the number of stops or vehicles that can be easily verified. Such measures could also be a way of ensuring policy neutrality as the model could be rerun with appropriate benchmark levels.	Appendix B
Correlation between variables	Using correlated vectors as independent variables in one regression can lead to mis-estimates of the effects of these variables	The correlation analysis suggests that in several instances variables are different measures for the same factor allowing us to group the candidate independent variables. In the econometric modelling we will not use more than one variable from each group.	Section 5.2 and above

This section first motivates our choice of candidate variables and develops the theoretical framework before presenting the results of the regression analysis in the form of a set of preferred models. It concludes by suggesting an approach to estimating benchmark expense levels. The technical details of the econometric analysis can be found in Appendix E.

5.1 Candidate explanatory variables

This section investigates which measures could be used on the right-hand side as demand generating variables alternative to urban population.

To the extent that employment and population patterns differ, employment may provide a robust measure of peak demand. As per the Stage 1 report, commute trips drive the AM and PM peak travel demand. Consequently, employment patterns may be a more accurate driver of transport requirements. For instance, average operating cost per passenger km may be lower during peak periods when transport occupancy is higher. Thus, employment patterns are relevant for informing expenses. School enrolments is another source of transport demand, but this may be potentially too narrow in its representativeness of the broader population's transport needs.

Density is another candidate variable that is also likely to capture geographic factors. In addition to this, the correlation matrix in Appendix D shows that population-weighted density, when compared to raw population, is less strongly correlated with other potential explanatory variables, thereby mitigating the risk of multi-collinearity.

A detailed assessment of the variables that we have considered for the regression analysis can be found in Appendices A to C. A brief summary of our main findings is presented below in Table 5.2. The variables are

[22] At a later stage – or if it proves too complex a task as an addition to this project – identifying a nationally available proxy data series that follows similar trends as are observed in the State expenditure data could provide a reference point that further helps assess the appropriateness of the States' expenditure data return.

classified according to the demand, supply and cost categories identified in Section 1.1. Further evidence on candidate variables can be inferred by looking at the correlation between independent variables. This is the subject of Section 5.2 and Appendix D.

Table 5.2: Summary of data assessment findings

Category	Variables	Main findings
Demand variables	· Employment · School enrolment · SEIFA · Income · Population-weighted density	SEIFA and income are both candidates to represent socio-economic status, but only one will be required. Distribution of employment and enrolments are heavily skewed due to the dominance of capital cities. In addition to being considered as a cost variable (see below), density is also relevant as a demand driver because it counts the number of people within a particular area. Population-weighted density is favoured over standard density, as it overcomes some of the weaknesses that potentially arise from more basic measures. Boxplots demonstrate that the sample of 70 SUAs is representative of all SUAs for these variables.
Supply variables	Transport infrastructure: · Heavy rail track km · Light rail track km · Busway lane km · Ferry wharves Transport vehicles: · Heavy rail cars · Light rails cars · Buses · Ferry vessels · Consolidated revenue km Public transport mode use indicators: · Mode use levels (tram/train/ferry bus) · Mode use dummies (tram/train/ferry/bus) · Census Journey to work	Although theoretically sound, use of the transport infrastructure and transport vehicle variables is limited by the number of observations available. Consolidated revenue km data is also incomplete To remedy this, public transport mode use indicators that could act as proxies, along with Census journey to work data. Mode use indicators can be expressed in levels, although there is the potential for correlation with population. Alternatively, mode use can be expressed as dummies (which take a value of zero or one depending on the presence of a particular mode)
Cost variables	· Population-weighted density Land slope indicators: · Zero slope land area · Land slope mean · Land slope SD Road and railway bridge indicators: · Road bridge line (count) · Road bridge point (count) · Road bridge line dimension · Railway bridge line (count) · Railway bridge point (count) · Railway bridge line dimension Railway segment indicators · Railway segment slope degree · Railway segment rise positive · Railway segment rise length	In addition to being relevant as a demand driver, density is likely to contribute to cost. Expenditure in higher density areas may entail greater complexity with regards to infrastructure provision and network management. As for the supply variables, several of the cost variables are constrained by a lack of observations. Land slope indicators report observations for most SUAs in the expense sample. Due to correlation between variables, not all three indicators will be required. Road and railway bridge data are theoretically sound, but lack observations. The railway segment data is less incomplete relative to the bridge data, but its use would nevertheless require several SUAs to be removed from the regression analysis entirely. Consequently, land slope indicators may more successfully account for topographical factors.

5.2 Variable groups

Using correlated vectors as independent variables in one regression can lead to mis-estimates of the effects of these variables. In order to avoid misspecifications resulting from uncontrollably inflated standard errors (multi-collinearity), combinations of highly correlated variables should not be included in the same regression equation. A correlation matrix provides important insights about which variables may and may not be compatible with each other when included as explanatory variables. As laid out in detail in Appendix D, we observe high correlation between variables that are likely to explain similar variation in expenses. Where two such alternatives exist, it may be appropriate to experiment with both variables to ascertain which one is more suitable for the econometric analysis.

The correlation analysis suggests that in several instances variables are different measures for the same factor. Together with the developed proxies, this allows us to group the candidate independent variables as follows:

1) Income and SEIFA: socio-economic status
2) Population density, employment, and school enrolments: demography
3) Transport infrastructure (e.g. track km), road/rail bridge variables, and transport vehicles (e.g. rolling stock): asset quantity
4) Land slope mean and land slope standard deviation: topography
5) Proxy variables:
    a) Public transport mode passengers and average distance to work for mode specific passenger kilometres travelled
    b) Along with employment, public transport mode passengers may also be able to proxy for congestion, whereas the relatively high correlation of congestion with population suggests that its effect might be partly covered by measuring expenses on a per capita basis.

In the econometric modelling below, we will not use more than one variable from each group[23] in a single specification because i) due to the relatively small sample of 70 observations, only a limited number of candidate variables should be included to ensure robust results and ii) introducing highly correlated variables to a single equation causes technical issues (multi-collinearity). In some cases, there may be a reasonable theoretical foundation for including more than one variable from a particular group. For instance, we have considered models that simultaneously include both density and employment. Even though they are both classified as demography variables, we have discussed throughout the report how density may account for variation in expenses not already captured by employment.

[23] In some instances, especially in the transport assets group, this could be a set of dummy variables representing for example the presence or absence of the four key modes.

Table 5.3 summarises the candidate independent variables arranged into the groups described above. For each variable, we provide a brief overview of its suitability for analysis and indicate whether it was tested in the final models. In some cases, a variable may have been examined in a preliminary regression. However, as discussed in Appendix E, in this report we present results for a set of final models that we deem most promising for informing future funding shares.

Table 5.3: Candidate independent variables, by group

Group	Variable	Selected for final models	Comments
1) Socio-economic status	Income	No	Was investigated as an alternative to SEIFA, but not included in final models. Expenses appear to be more closely related to SEIFA
	SEIFA	Yes	Suitable proxy for socio-economic status
2) Demography	Population-weighted density	Yes	Likely to be key driver of expenses, more suitable than basic density
	Employment	Yes	Important demand driver
	School enrolments	No	Relatively correlated with employment, some concerns about data
3) Transport infrastructure and transport vehicles	Heavy rail track km	No	Insufficient observations
	Light rail track km	No	Insufficient observations
	Busway lane km	No	Insufficient observations
	Ferry wharves	No	Insufficient observations
	Heavy rail cars	No	Insufficient observations
	Light rails cars	No	Insufficient observations
	Buses	No	Insufficient observations
	Ferry vessels	No	Insufficient observations
4) Topography	Zero slope land area	No	High correlation with employment compared to other geographical indicators
	Land slope mean	Yes	Geographical indicator with strong theoretical foundations
	Land slope SD	No	Too correlated with land slope mean to warrant inclusion
	Road bridge line (count)	No	Insufficient observations
	Road bridge point (count)	No	Insufficient observations
	Road bridge line dimension	No	Insufficient observations
	Railway bridge line (count)	No	Insufficient observations
	Railway bridge point (count)	No	Insufficient observations
	Railway bridge line dimension	No	Insufficient observations
	Railway segment slope degree	No	Theoretically sound, but data is missing for almost 20% of SUAs in expense sample
	Railway segment rise positive	No	Theoretically sound, but data is missing for almost 20% of SUAs in expense sample
	Railway segment rise length	No	Theoretically sound, but data is missing for almost 20% of SUAs in expense sample
5) Proxy variables	Mode use levels (tram/train/ferry bus)	Yes	More robust alternative to asset / transport vehicle data with more observations available – although it may exhibit high correlation with employment
	Mode use dummies (tram/train/ferry/bus)	Yes	More robust alternative to asset / transport vehicle data with more observations available
	Journey to work	Yes	Key driver of transport needs

5.3 Statistical model selection criteria

In selecting preferred models, we have had regard to theoretical soundness, goodness of fit indicators (such as R^2 and information criteria), the statistical significance of explanatory variables, and visual inspections of the fitted values and residuals.

R^2 and Adjusted R^2

R^2, which measures the proportion of variation in the dependent variable explained by the given model, is one of the primary goodness of fit indicators. As such, R^2 is valuable in illustrating the improvement of our chosen models over the reference model. However, there is also an important trade-off between the fit of the model and its parsimony. The principle of parsimony stipulates that additional explanatory variables should only be included to the extent that they materially improve the fit of the model. In other words, where a simple model and complicated model offer similar explanatory power, the simpler model should be adopted. One drawback of R^2 is that it cannot decrease as additional variables are added to the model. This means that even the inclusion of irrelevant variables can increase the R^2. To address this, a supplementary indicator is the adjusted R^2, which effectively penalises the inclusion of variables that do not materially improve the explanatory power of the model.[24]

Information Criteria

As a supplement to R^2 and Adjusted R^2, other indicators that strike a balance between model fit and parsimony include the Akaike Information Criterion (AIC) and Bayesian (or Schwarz) Information Criterion (BIC).[25] In regard to model selection, lower information criteria are favoured to inform the preferred model.

Statistical significance

With regards to the statistical significance of variables, we would recommend placing less weight on p-values relative to other model and/or variable selection indicators. In the present setting, the size of the sample makes heavy reliance on p-values difficult. Often there are economic justifications for including a particular variable in the expense model, such as those outlined in the framework in Section 5.2. Where such theoretical underpinnings exist, then an insignificant p-value may not necessarily preclude us from incorporating it into a preferred model.

This is not to say that p-values are not relevant at all. Rather, p-values may simply not be the most relevant criterion for variable/model selection in the present context. A high R^2 (also having regard to adjusted R^2), low information criteria and a significant overall F-statistic are likely to be preferable in selecting an optimal model.

Substantial academic evidence is accumulating in support of not placing excessive reliance on p-values.[26] In 2016, the American Statistical Association (ASA) released a statement on statistical significance and p-values.[27] It provided a set of principles that were intended to address misconceptions about p-values and thereby improve the interpretation of quantitative analysis. In essence, the ASA concluded that scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. Furthermore, by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. This lends support to the modelling framework that we have adopted in this report, which balances economic theory with statistical indicators.

[24] The formula for adjusted R^2 is: Adj. $R^2 = 1 – [SSE/(N-K) / SST/(N-1)]$. This compares to the conventional R^2 formula: $R^2 = 1 – (SSE/SST)$
Where: SSE = sum of squared errors (unexplained variation in dependent variable)
SST = total sum of squares (total variation in dependent variable)
N = number of observations
K = number of coefficients in the equation
Holding all else constant, the Adjusted R^2 is decreasing in K. This means that, unless the unexplained variation in the dependent variable falls sufficiently, including extra variables may cause the Adjusted R^2 to fall. Adjusted R^2 does not specifically identify which variables are failing to materially improve the explanatory power of the model.

[25] The formulae for the AIC and BIC are as follows:
AIC = ln(SSE/N) + 2K/N
BIC = ln(SSE/N) + K*ln(K)/N

[26] See, for example: Hubbard, R and Lindsay, R.M. (2008) *Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing*, Theory & Psychology, 18:1, 69-88.

[27] American Statistical Association (2016). American Statistical Association releases statement on statistical significance and p-values, March 7.

Predictive capabilities

Finally, because the ultimate objective of this analysis is to inform funding shares, it is important to evaluate how the various models perform in predicting observed data. In the rest of this section, we demonstrate the predictive capacity of the models visually using plots of the actual versus fitted expense values, as well as using residual boxplots by State.

5.4 Preferred model

Based on the theoretical framework set out in the introduction, the preferred model form should have variables from the set or proxy variables for volume and with a set of area specific variables to be consistent with the theoretical framework underpinning the flowing equation:

$$E_i = F(D_i, S_i, C_i) \quad \text{(eq. 5.1)}$$

Where E_i is expenditure (as net per capita expenditure[28]), D_i are demand variables (essentially proxies for volume). S_i are supply or network related variables, some of which are proxies for volume and some of which capture cost of provision and factor price effects and C_i are city specific variables that capture differences between or SUA specific variables that influence differences in cost of provision across SUAs.

Also discussed in the introduction, the expected functional form is an open question to some extent as the presence and direction of economies of scale is unclear. In order to establish the preferred model presented below, we have tested numerous models within the framework set out above, covering many permutations of explanatory variables and functional forms. The models were compared against the statistical model selection criteria described above.

The best model is presented below (Model 1b in Appendix E) and the set of candidate-preferred models that were assessed in detail in Appendix E. Relative to the other models, it performs well on the model selection criteria, with high R^2 and adjusted R^2 values, and low information criteria values. Furthermore, the model adheres to the established theoretical framework, with variables from the demography, topography and proxy variable categories.[29] With regards to the predictive capabilities of the model, a plot of the predicted expense values against the actual values reveals a suitably close fit. Likewise, an inspection of the residuals indicates that the estimates are unbiased overall and by State.

The preferred model uses density ($dense_i$) to depict demand, distance to work ($dist_i$) to represent network complexity, passengers by public transport mode ($pax_{i,mode}$) to represent availability and congestion, and mean land slope ($slope_i$) to account for topography. Formally the model can be specified as:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 \ln(pax_{i,train}) + \beta_5 \ln(pax_{i,bus}) + \varepsilon_i \quad \text{(eq. 5.2)}$$

[28] We also tested models using (unscaled) net expenses and expenses per person as dependent variable. As suspected in Section 3.4, the smaller variation in the population-scaled sample produces substantially more robust models. As per the stage 1 report, we consider expenses per person the preferable dependent variable. Furthermore, this way, for SUAs crossing State borders, models can be estimated using data for the entire SUA and the results of the modelling can be directly applied to the populations on each side of the border to calculate the respective part's contribution to the total expense of the associated State.

[29] The preferred model does not include any socio-economic status or transport infrastructure/vehicle variables. In the case of the former, the inclusion of SEIFA was found not to improve the fit of the model. In the case of the latter, incomplete data limited the observations that were available.

Table 5.4: Model 1b: Test statistics

	Preferred model
Observations	70
F statistic	48.44
Prob > F	0.0000
R^2	0.79
Adjusted R^2	0.77
Akaike information criterion	798
Bayesian information criterion	811
Root MSE	69

Source: Synergies modelling

Table 5.5: Preferred model: Coefficient estimates

	Coefficient estimate	Standard error	95% confidence interval	
Intercept	-154.5637**	46.8811	-248.2194	-60.90792
$dense_i$	0.0715307***	0.0200746	0.0314271	0.1116343
$dist_i$	3.411582*	1.647887	0.1195494	6.703616
$slope_i$	6.963933	4.882911	-2.790803	16.71867
$\ln(pax_{i,train})$	18.07401***	4.036532	10.01011	26.13791
$\ln(pax_{i,bus})$	6.719857	6.659917	-6.584856	20.02457

*** p > |t| <=0.1%
** p > |t| <=1%
* p > |t| <=5%
^ p > |t| <=10%

A sensitivity analysis in which the model is re-estimated with a higher expense level for Melbourne as suggested in Section 2.3 showed only a marginal change in the coefficient estimates and hence the predicted benchmark values.

Coefficient estimates follow intuition as the model suggest that net expenses per person

- increase with urban density (representing demand);
- increase with the distance to work (representing network complexity);
- increase with mean land slope (depicting topographical complexity); and
- increase with train and bus passengers.

The model incorporates passenger mode numbers in a linear-log functional form (the name arises because the independent variables have been transformed by a logarithm, while the dependent variable has not). The linear-log relationship implies that per capita expenses increase as the network becomes more complex but the rate at which this occurs decreases as passenger volumes increase. This holds for buses and rail. Specifically, the linear-log relationship implies that for every 1% increase in passenger mode numbers, per capita expenses increase by a dollar amount equal to the respective estimated coefficient divided by 100.

Consider, under Model 1b, the effect of increasing bus passenger volume by 10% in Darwin versus increasing bus passenger volume by 10% in Kalgoorlie. Holding all other factors constant, this bus passenger increase will increase the per capita expenses in both cities by $0.67. However, because the passenger base in Kalgoorlie (594 passengers on Census night) is only about one tenth of that in Darwin (5100 passengers on Census night),

the expense increase per additional 100 passengers in Kalgoorlie is $0.11 and that in Darwin only $0.01. The same holds for rail where per capita expenses increase by $1.81 with every 10% increase in passengers.

This means the linear-log form of the model can be interpreted as indicative of scale effects in the wider sense as it suggests that growth from additional passengers becomes less substantial as total volume increases.

6. Conclusions

The objective of this report was to develop a model that comprehensively captures the primary drivers of urban transport supply and demand that influence State government expenditure on a per capita basis. Building on the theoretical framework developed in Section 1.1, the preferred model form is:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 \ln(pax_{i,train}) + \beta_5 \ln(pax_{i,bus}) + \varepsilon_i \quad \text{(eq. 6.1)}$$

Where;

- Density ($dense_i$) is the demand variable, which acts as a proxy for traffic volume;
- The bus and train passenger counts ($pax_{i,train}$ and $pax_{i,bus}$) are supply or network related variables, which also capture cost of provision and factor price effects as proxies for congestion and volume; and
- Distance ($dist_i$) and mean slope ($slope_i$) are SUA-specific variables that capture differences between the SUAs and that influence the cost of provision across SUAs.

Based on these principles, we have estimated a model that is consistent with the theoretical framework and that also performs well in statistical tests. As the model captures all key relevant (theoretical) drivers of public transport expenditure, its forecasts can be considered a relevant benchmark for appropriate expenses under each SUA's specific attributes. Hence, we can apply this model to derive a policy neutral benchmark per capita expense level for all SUAs. Figure 6.1 below presents the actual values and those predicted by the preferred model.

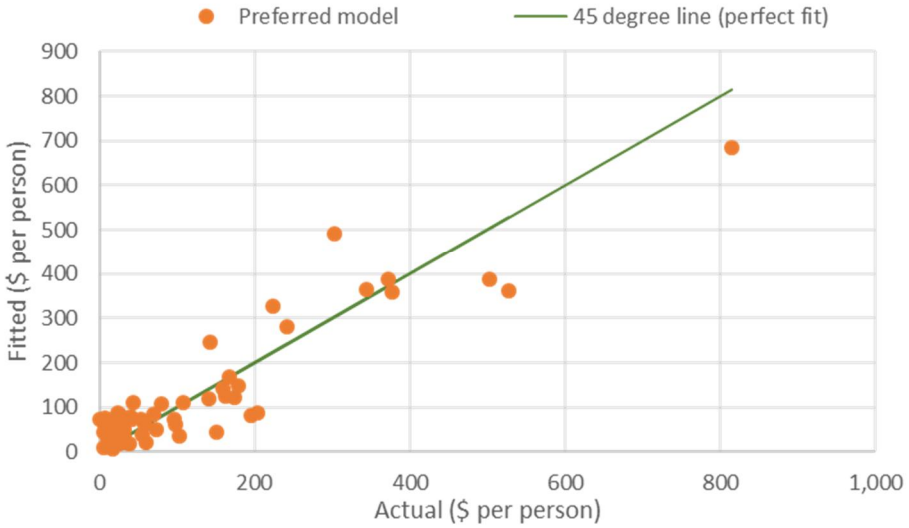


Figure 6.1 Preferred model: Actual vs. fitted values

Data source: Synergies modelling

On the plot above, the green (45 degree) line represents instances in which the actual and the predicted value are identical. Dots above the line represent SUAs in which the predicted value is larger than the actual value indicating expense levels below the benchmark. Dots below the line represent SUAs in which the predicted value is smaller than the actual value indicating expense levels above the benchmark. In other words, the line on the graph could be interpreted as the boundary between SUAs with public transport networks with particularly low per capita cost under their specific characteristics (those above the line) and SUAs with relatively high per capita cost under their specific characteristics (those below the line).

The preferred model described in this report is based on a larger dataset than the model previously identified in the 2015 Review. It also has increased flexibility and a stronger theoretical foundation achieved through the

inclusion of additional variables. Hence, compared to the 2015 Review model, it is likely to produce more accurate estimates and ultimately more representative funding shares.

7. References

Australian Bureau of Statistics 2017, Census 2016, *Employment by SUA*, TableBuilder. Findings based on use of ABS TableBuilder data.

Australian Bureau of Statistics 2017, Census 2016, *Population by SUA*, TableBuilder. Findings based on use of ABS TableBuilder data.

Australian Bureau of Statistics 2017, Census 2016, *SEIFA by SUA*, TableBuilder. Findings based on use of ABS TableBuilder data.

Australian Bureau of Statistics 2017, Census 2016, *Income by SUA*, TableBuilder. Findings based on use of ABS TableBuilder data.

Australian Bureau of Statistics 2017, Census 2016, *Place of work vs. place of usual residence by SA2 and State*, TableBuilder. Findings based on use of ABS TableBuilder data.

Australian Bureau of Statistics 2017, Census 2016, *Distance of journey to work by SUA*, TableBuilder. Findings based on use of ABS TableBuilder data.

Australian Bureau of Statistics, 2017, *Australian Statistical Geography Standard (ASGS): Volume 4 - Significant Urban Areas, Urban Centres and Localities, Section of State,* cat. no. 1270.0.55.004, viewed 12 September 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.004>

Australian Bureau of Statistics, 2011, *Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011,* cat. no. 2033.0.55.001, viewed 12 September 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2033.0.55.001main+features100042011>

Bierman, S. and Martinus, K. (2017) *Boundary Objects as Tools for Integrated Land Use Planning*. In Bierman, S. Olaru, D and P. Valeria, editors. Planning Boomtown and Beyond. Perth: UWA Press

Bierman, S. and Martinus, K. (2018) *Strategic Planning for Employment Self-Containment in Metropolitan Sub-Regions*, Urban Policy and Research, 36:1, 35-47.

Bureau of Infrastructure, Transport and Regional Economics (BITRE), *Australian Infrastructure Statistics Yearbook 2016*; 2016

Bureau of Infrastructure, Transport and Regional Economics (BITRE), *Information sheet 74: Traffic and congestion cost trends for Australian Capital Cities*; 2015

Commonwealth Grants Commission (2015), Report on GST Revenue Sharing Relativities 2015 Review – Volume 2 – Assessment of State Fiscal Capacities.

Jacobs (2017), *Modelling of urban transport recurrent and infrastructure expenditure requirements: Stage 1 report to the Commonwealth Grants Commission*

Total Cost Review of Regular Bus Services Operated in Sydney's Four Largest Regions. Prepared for The Independent Pricing and Regulatory Tribunal NSW (IPART). Indec. Sydney, 2009

Western Australian Planning Commission and Department of Planning, Lands and Heritage (2018), *Perth and Peel at 3.5 million*.

Appendix A. Detailed assessment of demand variables

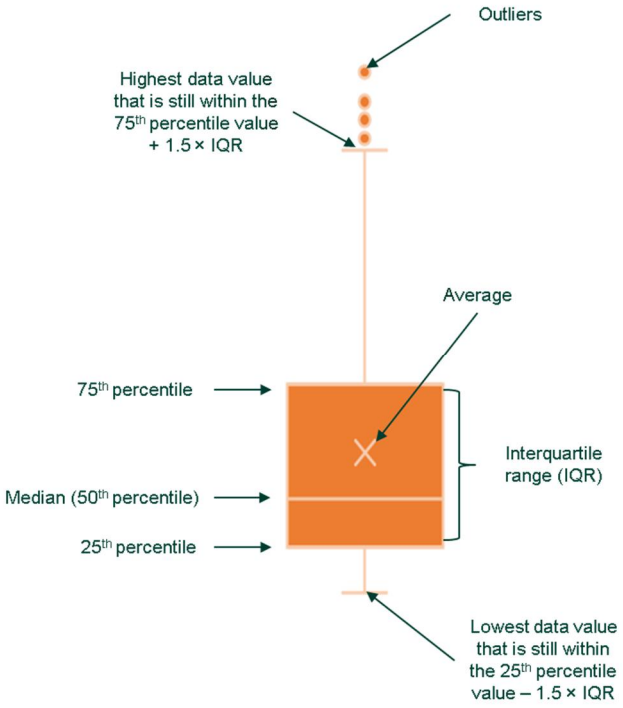
This Appendix presents a range of demand variables that account for various demographic factors that are likely to affect transport usage. In essence, demand variables capture consumer behaviour in relation to transit decisions.

Throughout this report, boxplots are used to illustrate data. The box below, Table A.1, summarises what such plots show and how they can be interpreted.

Table A.1: Box plots

A box plot graphically depicts groups of numerical data through their quartiles. Box plots also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles. Box plots display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacing between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data.

In the adjacent chart, the thick bar represents the median, the average is represented by an "X" and the box (interquartile range or IQR) contains approximately 70% of observations. The tips of the whiskers represent the maximum and minimum values excluding the outliers, which are represented as dots outside the whiskers. These are illustrated in the figure on the right.



A.1 Key Stage 1 report findings

The Stage 1 report observes that the urban transport task is highly correlated with population. However, it points out that the current model in which population is the sole driver could be improved by including other influences that explain the variation in travel demand between cities. Such influences could include travel distance to work, education and other social and cultural activities. Specifically, it identifies the following demand related factors:

- **Population serviced by an urban transport network**
  The urban transport task is a derived demand, as travel occurs because people want to undertake specific activities at different locations in an area. Thus, the transport activity only occurs because of some other underlying location dependent demand. The underlying demand is driven by the desire of a city's population for commuting, education, business, leisure, retail and recreational activities.
- **Commuting journeys**
  Public transport trip generation by commuters (employed persons and students) tends to be highly concentrated and occur during a morning and an afternoon peak. Since the system needs to be designed to accommodate these peaks, a larger number of commuters is likely to cause higher expenditure (all else equal) as the peaks are likely to be more pronounced than in a system with lower commuter numbers.

Taking the above factors into account candidate demand variables include: population, the number of persons employed and student enrolments. Density is also a candidate demand variable, but for the purpose of the variable assessment we have considered it under the cost variable category.

A.2 Variables

In addition to the population focused variables demand related factors set out in the Stage 1 Report and the previous section, we consider socio-economic factors to be potentially relevant as well. Therefore, this section also introduces Socio-Economics Indexes for Areas (SEIFA) scores and average incomes as potential candidate variables.

Employment

Employment is one of the candidate variables reflecting population composition. Travel to employment is one of the main reasons for travel during peak times. Consequently, the level of employment will affect recurrent expenditure levels. The different options for measuring employment are shown in Table A.2.

Table A.2: Employment

Variable description	For the purpose of employment, we consider both full-time and part-time workers. We also consider the unemployed looking for either full-time or part-time work, as these individuals are also likely to be utilising public transport when searching for employment.
Reason for inclusion	Employment-related travel induces a particular kind of demand, characterised by many commuters travelling in the same direction at the same time. This means that recurrent expenditure must be sufficient to deal with these peak loads. Jacobs commented in the Stage 1 report that “Commute trips drive the AM and PM peak travel demand. The average operating cost per passenger km may be lower during these peak periods as occupancy is higher in peak hours compared to the off-peak.”
Expectation	There are two options for the employment variable: employment by place of work (POW), and employment by place of usual residence (PUR). We consider that employment by POW will be the more useful indicator, as it considers where people actually work. Holding all else constant, it is anticipated that higher levels of employment will require higher recurrent expenditure.
Statistical level	SUA
Data modifications required	Employment is calculated as the sum of all part-time and full-time workers, plus those currently looking for work.
Maximum available observations	101 (70 in expense sample)
Average value	90,649 persons (125,100 persons in sample)
Median value	11,974 persons (15,840 persons in sample)
Maximum value	2,092,602 persons (2,092,602 persons in sample)
Minimum value	3,075 persons (3,346 persons in sample)
Standard variation	316,128 persons (373,299 persons in sample)
Robust variable for analysis	Yes (Yes)

Source: Synergies analysis of Census 2016 data

In logarithmic form, Figure A.1 shows a relatively balanced distribution compared to the untransformed variable. However, even after the logarithmic transformation, the dominance of the capital cities can still be observed. The majority of SUAs have a labour force of less than 100,000 workers, with almost half of the labour force concentrated in Sydney, Melbourne and Brisbane.

The boxplot for the sample of 70 (left panel) resembles that for all 101 closely. This indicates that the sample of the 70 SUAs is very likely to be representative for all 101 SUAs.

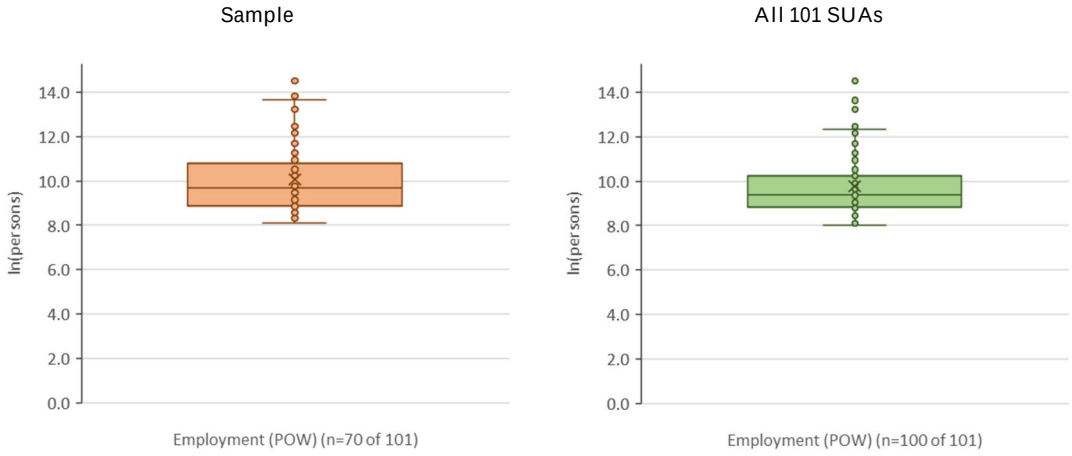


Figure A.1: Distribution of number of persons in the labour force (logarithmic form)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

School enrolment

The number of school enrolments is relevant for ascertaining transport needs, as students are more reliant on public transport due to their age. As such, this variable can be considered a population composition factor. Details regarding the enrolment data from ACARA are shown in Table A.3.

Table A.3: School enrolment

Variable description	ACARA (Australian Curriculum, Assessment and Reporting Authority) collects data on school enrolments of primary and secondary school students.
Reason for inclusion	Unless they can rely on their parents or other family members for car transport, school students are heavily dependent on a reliable public transport network. As identified in the Stage 1 report, educational trips are also likely to be heavily subsidised, which poses implications for expenditure levels.
Expectation	Our expectation is that a high proportion of students in an SUA (especially secondary students) will lead to an increased reliance on public transport, thereby demanding higher expenditure.
Statistical level	SUA
Data modifications required	Primary and secondary school enrolments will be aggregated into total school enrolments to preserve degrees of freedom. However, the distinction is still valuable, as secondary school students are likely to be less dependent on their parents for school transport requirements.
Maximum available observations	101 (70 in expense sample)
Average value	32,499 persons (34,675 persons in sample)
Median value	5,714 persons (6,802 persons in sample)
Maximum value	693,445 persons (693,445 persons in sample)
Minimum value	908 persons (908 persons in sample)
Standard variation	104,627 persons (100,057 persons in sample)
Robust variable for analysis	Yes (Yes)

Source: ACARA provided by the CGC

Figure A.2 shows the range of school enrolments by SUA. The demand for education-related transport will be most concentrated in capital cities and other significant regional centres, such as Newcastle-Maitland, Central Coast and Sunshine Coast. 57 SUAs have fewer than 10,000 enrolments.

In the ACARA data, the number of enrolments in Melbourne SUA appear relatively small. The Melton SUA in contrast, which is just south of Melbourne, has a very large number of enrolments that does not seem to align with the number of schools in the SUA. In the related Census data this anomaly does not seem to occur.

The boxplot for the sample of 70 (left panel) resembles that for all 101 closely. This indicates that the sample of the 70 SUAs is very likely to be representative for all 101 SUAs.

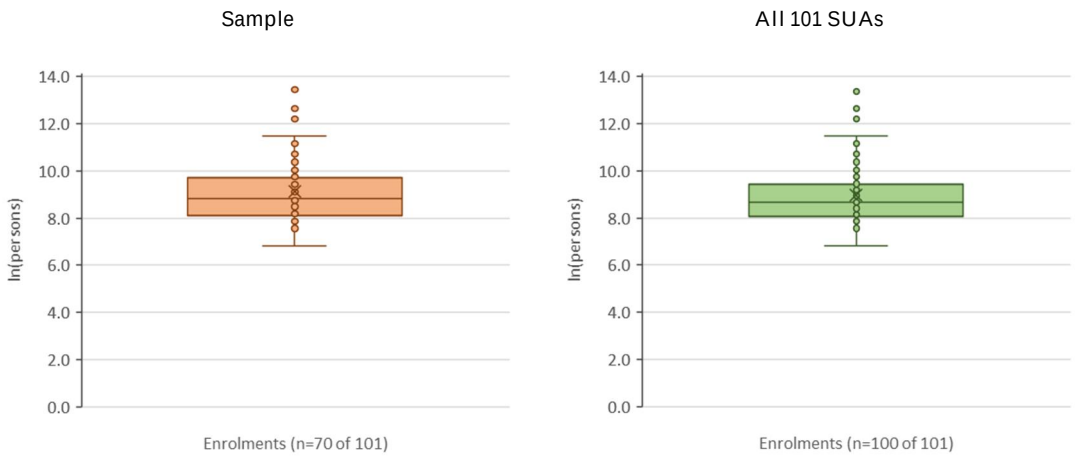


Figure A.2: Distribution of school enrolments (logarithmic form)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of ACARA data provided by the CGC

SEIFA (Socio-Economics Indexes for Areas)

Socio-economic status is likely to be a key determinant of public transport use. As detailed in Table A.4, SEIFA is likely to provide a robust proxy for socio-economic status in the regression specifications. SEIFA represents a measure for economic circumstances.

The ABS derives a suite of four SEIFA indexes:

- Index of Relative Socio-economic Disadvantage (IRSD)
- Index of Relative Socio-economic Advantage and Disadvantage (IRSAD)
- Index of Economic Resources (IER)
- Index of Education and Occupation (IEO)

For the purpose of the analysis in this report, we have used IRSAD, because it is the most comprehensive of the four measures. According to the ABS, IRSAD is preferred in situations where a general measure of advantage and disadvantage is sought.[30] The alternative indexes on specific aspects of disadvantage, such as education or wealth.

Table A.4: SEIFA (Socio-Economic Indexes for Areas)

Variable description	SEIFA ranks areas in Australia according to relative socio-economic advantage and disadvantage. It is based on census information. For each SA2 region, the data identifies how many households fall under each decile.
Reason for inclusion	The ABS identifies one of the common uses for SEIFA as: “determining areas that require funding and services”. Applied to the context of transport, it would seem feasible that demand for transport services would be related to socio-economic status.
Expectation	In the context of transport, we would expect those areas with greater socio-economic disadvantage to be more reliant on public transport, holding all other factors constant.
Statistical level	Originally SA2, aggregated to SUA.
Data modifications required	In order to generate a single indicator for each SA2 region, we have weighted the number of households in each decile to establish the average decile for each region

[30] See ABS catalogue 2033.0.55.001 - Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011

Maximum available observations	101 (70 in expense sample)
Average value	3.5 index value (3.7 index value in sample)
Median value	3.4 index value (3.8 index value in sample)
Maximum value	8.4 index value (8.4 index value in sample)
Minimum value	1.1 index value (1.1 index value in sample)
Standard variation	1.7 index value (1.8 index value in sample)
Robust variable for analysis	Yes (Yes)

Source: Synergies analysis of Census 2016 data

The distribution of average deciles for the SUAs is shown in Figure A.3. With the exception of three outliers (Canberra-Queanbeyan, Gisborne–Macedon and Karratha), average deciles range between 1.11 and 8.40 (with a higher value corresponding to higher socioeconomic status). The interquartile range lies between 2.15 and 4.69.

The boxplot for the sample of 70 (left panel) resembles that for all 101 closely, although the SUAs with SEIFA values above 7 are deemed to be outliers when using all SUAs. This being said, the sample of the 70 SUAs is still very likely to be representative for all 101 SUAs.

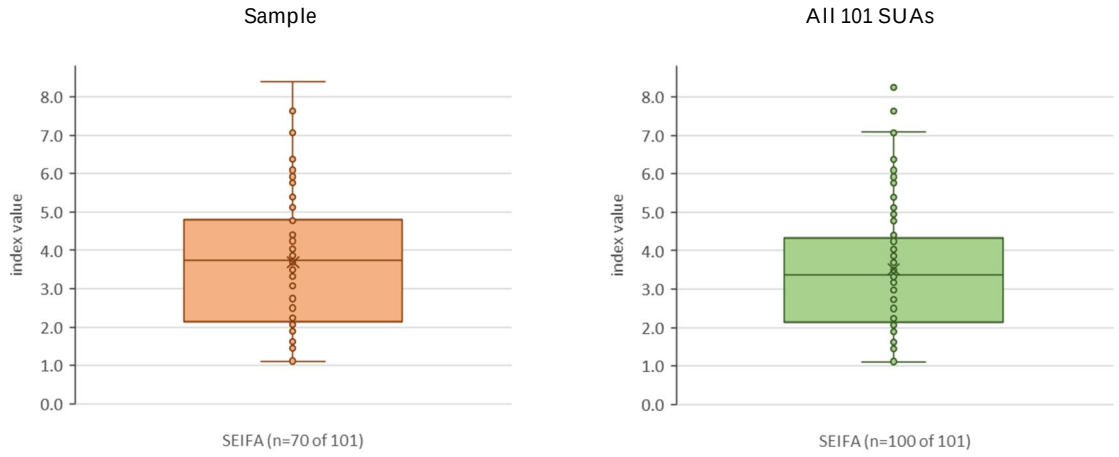


Figure A.3: Distribution of SEIFA values

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

Income

Similar to SEIFA, income represents socio-economic status and is hence a measure of economic circumstances. Income is likely to determine the propensity to utilise public transport, but the possibility of high correlation between income and SEIFA may result in only one of these variables being included in a regression specification.

Table A.5: Income

Variable description	Income is based on median total personal weekly income for each SUA. We favour personal income over household income but will investigate both as a robustness check.

Reason for inclusion	Income is likely to be related to reliance on public transport, because income influences the availability of private transport alternatives.
Expectation	It is expected that higher incomes will be associated with a decreased reliance on public transport, as car ownership is likely to be higher. However, lower incomes may suggest that there are other sectors competing for funding, which could have the adverse effect of decreasing transport funding. We will also need to verify that income is not overly correlated with SEIFA.
Statistical level	SUA
Data modifications required	None
Maximum available observations	101 (70 in expense sample)
Average value	635 $/week (650 $/week in sample)
Median value	605 $/week (618 $/week in sample)
Maximum value	1,361 $/week (1,361 $/week in sample)
Minimum value	471 $/week (472 $/week in sample)
Standard variation	159 $/week (173 $/week in sample)
Robust variable for analysis	Yes (Yes)

Source: Synergies analysis of Census 2016 data

Except for 8 outliers (predominantly SUAs associated with mining activities), Figure A.4 shows that average weekly personal income ranges between $472 and $730. The interquartile range is $542 to $668.

The boxplot for the sample of 70 (left panel) resembles that for all 101 closely. More outliers are observed in the sample with all SUAs, but these are still bounded by the outliers included in the 70-SUA sample. This indicates that the sample of the 70 SUAs is very likely to be representative for all 101 SUAs.

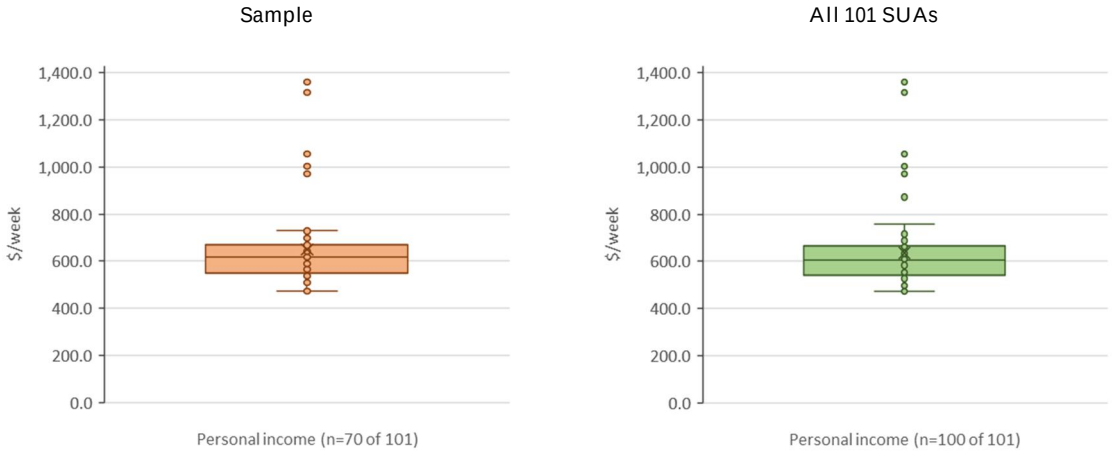


Figure A.4: Distribution of average weekly incomes

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

Demand variable summary

Demand variables encompass consumer and demographic patterns that are likely to determine urban transport needs. The demand variables available for this analysis consist of employment, enrolment, income and SEIFA.

Employment is expected to be a strong determinant of per capita expenses, given that it is an important driver of transport demand during peak hours. School enrolment is another potential indicator, but its inclusion will rest on whether it adds explanatory power on top of that already offered by employment, especially if it exhibits high correlation with employment.

When specifying potential expense models, the primary challenge is that some of the demand variables will explain the same variation in per capita expenses. For example, socio-economic status could be proxied by either income or SEIFA, but there will be no need to include both in the same specification. This would potentially lead to the statistical problem of multicollinearity, which we address in Section 5.

With regards to the statistical robustness of the variables, all indicators have sufficient observations for regression analysis. Furthermore, boxplots indicate that the 70-SUA sample is representative of all 101 SUAs.

Appendix B. Detailed assessment of supply variables

Candidate variables in this category typically relate to service levels. These are deemed to be supply-oriented, as the level of transport infrastructure and vehicles determines the availability of public transport for prospective users.

B.1 Key Stage 1 report findings

Different public transport modes are associated with different cost structures. The Stage1 Report points out that on a dollar per vehicle kilometre basis, the Australia wide average cost of operating trains is approximately six times more expensive than operating buses and about 1.5 times more expensive than light rail. However, on a dollar per passenger kilometre basis average operating costs are similar for the three modes.[31]

Therefore, the presence/absence of modes should be controlled for in developing urban transport recurrent and infrastructure expenditure models. Among eight Australian capital cities, five major cities have train services and four have light rail services. Statistically, a dummy variable, taking the value of either 0 or 1, is an accepted way to control for the existence of a particular mode. For example, the dummy variable for train takes the value 1 in Sydney with the service and 0 in Hobart without the service. This will allow the control of existence if train services in SUAs in the models.

Beyond the existence of a mode, there are a range of supply variables that capture the relationship between public transport provision of specific modes and costs. For example, rail track length is expected to correlate with costs of provision. These variables are discussed more fully in the following section.

B.2 Variables

As the provision of urban public transport is a State responsibility, the variables presented here tend to be policy-related. While using dummy variables as suggested in the Stage 1 report is a viable way of including supply in the econometric model, the Team considers it important to consider testing the influence of more detailed variables to avoid introducing omitted variable bias and to be able to potentially control for (some of) their impacts (see Section 2.1).

The limited data availability observed in this section is likely resulting from State governments or their entities not owning transport assets and the contracted private operators not necessarily reporting their asset bases to the States. For this reason, we will revisit the proxy set of supply variables developed in Section 2.4.

Track length, lane length, and number of wharves

The quantity of existing track length, bus lane length and ferry wharves are likely to drive recurrent expenditure on transport.

Table B.1: Track/lane length / number of wharves

Variable description	These variables include heavy rail track km, light rail track km, busway lane km and the number of ferry wharves
Reason for inclusion	Track and lane length serve as a useful representation for the scale and complexity of a given transport network
Expectation	Higher quantity of existing infrastructure will be associated with higher recurrent expenditure due to maintenance costs.
Statistical level	SUA
Data modifications required	Some concerns have been raised with data reporting, given the number of zero values.

[31] See Stage 1 Report Table 2.3 on p. 15: Public transport operating cost in Australian capital cities

	Heavy Rail Track Km	Light Rail Track Km	Busway Lane Km	No. of Ferry Wharves
Maximum available observations	10 (10 in expense sample)	5 (5 in expense sample)	2 (2 in expense sample)	2 (2 in expense sample)
Average value	422 track kms (422 track kms in sample)	120 track kms (120 track kms in sample)	13 lane kms (13 lane kms in sample)	20 (20 in sample)
Median value	220 track kms (220 track kms in sample)	28 track kms (28 track kms in sample)	13 lane kms (13 lane kms in sample)	20 (20 in sample)
Maximum value	1,492 track kms (1,492 track kms in sample)	529 track kms (529 track kms in sample)	24 lane kms (24 lane kms in sample)	39 (39 in sample)
Minimum value	10 track kms (10 track kms in sample)	0 track kms (0 track kms in sample)	1 lane kms (1 lane kms in sample)	0 (0 in sample)
Standard variation	517 track kms (517 track kms in sample)	229 track kms (229 track kms in sample)	16 lane kms (16 lane kms in sample)	28 (28 in sample)
Robust variable for analysis	No (No)	No (No)	No (No)	No (No)

Source: Synergies analysis of State data collected by the CGC

Upon inspection, there are a number of data-related issues that need to be highlighted.

To use Brisbane as an example, the asset quantity data indicate that Brisbane has no busways or ferry wharves, which is clearly not the case. This may be because these forms of infrastructure are managed by Brisbane City Council, rather than the State Government.

In the case of light rail, there are only four SUAs with this form of infrastructure. This makes it difficult to use this as a variable in our regression. Instead, it may be more feasible to include the presence of light rail infrastructure as a dummy variable.

Heavy rail track, as shown in Figure B.1, is concentrated in only 10 SUAs (all capital cities except for Wollongong, Gold Coast-Tweed Heads, Central Coast, Sunshine Coast and Newcastle-Maitland).

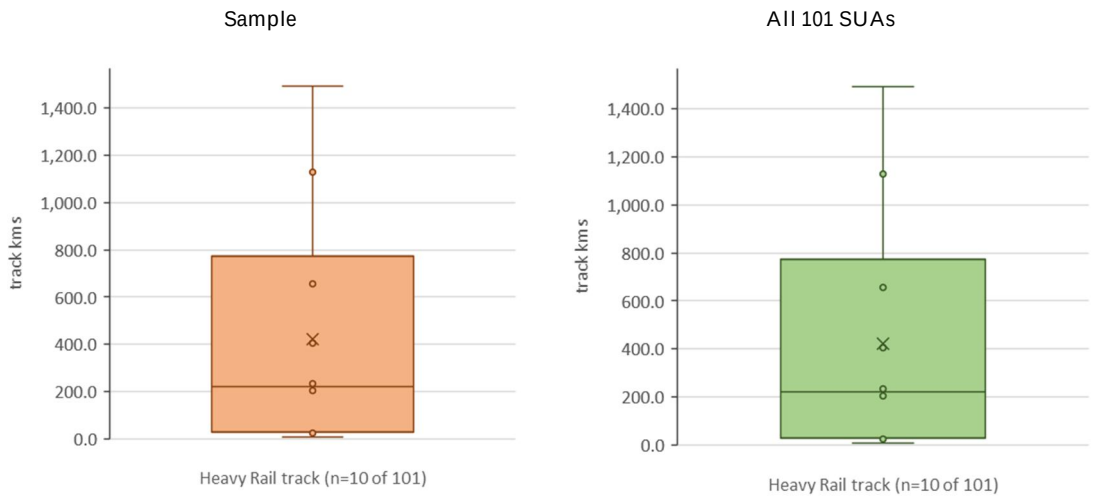


Figure B.1: Distribution of rail track km

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

The boxplot for light rail track km is presented in Figure B.2.

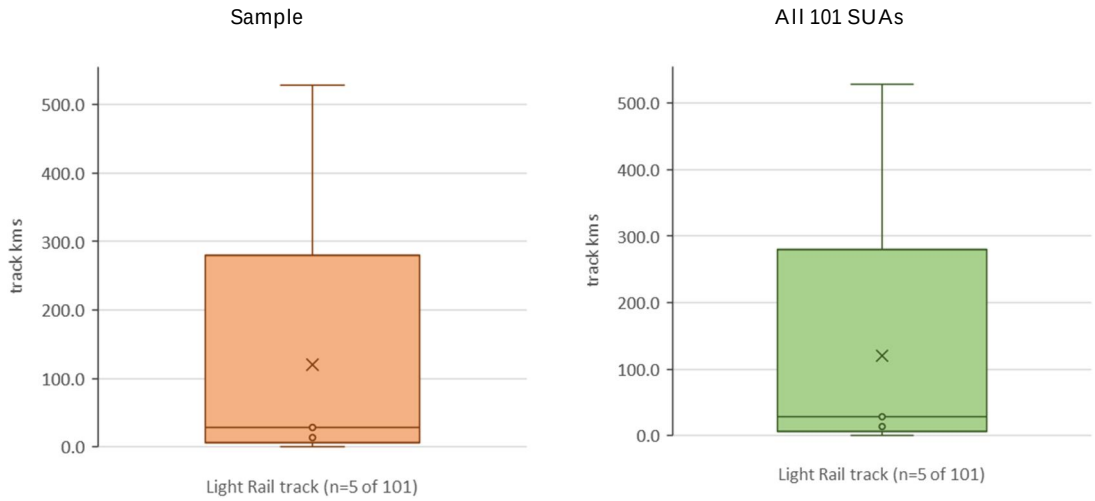


Figure B.2: Distribution of light rail track km

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Figure B.3 presents the data on busways. Only Perth and Adelaide report non-zero busway lane km.

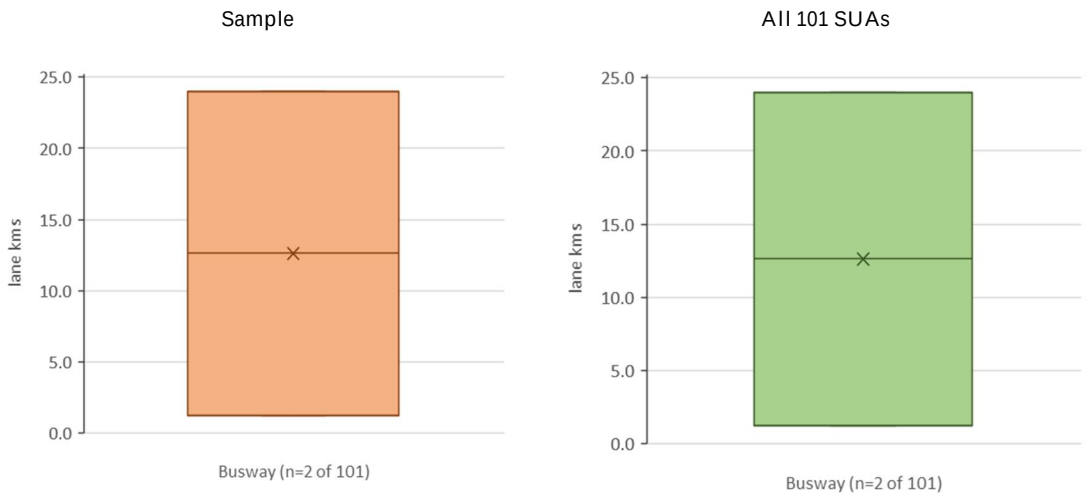


Figure B.3: Distribution of busway km

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Figure B.4 shows the number of ferry wharves. Only Sydney is reported to have ferry wharves (39). However, as documented below, Perth is reported as having ferries, but no ferry wharves.

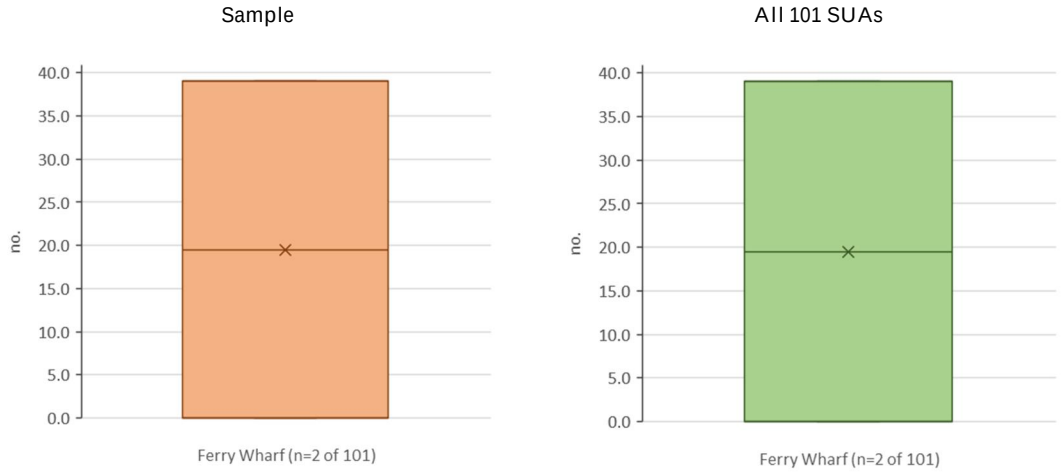


Figure B.4: Distribution of count of ferry wharves

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

It is evident that data availability is affecting the shape of the boxplots for track length, lane length and number of wharves. In general, it seems that the quantity of some assets is either understated or missing for certain SUAs. For instance, Brisbane has an extensive ferry network, which has not been captured in this data. Where asset quantity data is questionable, it may be more appropriate to model these factors as dummy variables, which simply indicate the presence or absence of a particular mode of transport for each SUA.

Transport vehicles

The number of transport vehicles also falls under the category of service levels. The number of trains or buses or ferries in a fleet is expected to affect the level of expenditure on transport. There is, however, a risk that the transport vehicle variables could be highly correlated with the track and lane length variables in the previous subsection. We explore this possibility in the correlation matrix in Appendix D.

Table B.2: Quantity of transport vehicles (heavy rail cars, light rail cars, buses and vessels)

Variable description	These variables list the number of heavy rail cars, light rail cars, buses and vessels within a given statistical area.			
Reason for inclusion	Recurrent transport expenditure is likely to be a function of the existing transport fleet, due to factors such as maintenance, staffing and replacement costs over time.			
Expectation	Our expectation is that expenditure will increase in line with the quantity of different mode of transport vehicles. As discussed in Section 5.2, it will be important to verify that there is no excessive collinearity between the number of transport vehicles and the quantity of line length that these vehicles service (as shown in Section 5.2).			
Statistical level	SUA			
Data modifications required	The number of available observations limits the degrees of freedom for the regression analysis. To make aggregation possible, it may be possible to compare these different transport vehicles on the basis of passenger capacity			
	Heavy rail cars	Light rail cars	Buses	Vessels
Maximum available observations	10 (10 in expense sample)	5 (5 in expense sample)	23 (22 in expense sample)	2 (2 in expense sample)
Average value	347 no. of cars (347 no. of cars in sample)	113 no. of cars (113 no. of cars in sample)	432 no. (452 no. in sample)	17 no. (17 no. in sample)
Median value	193 no. of cars (193 no. of cars in sample)	14 no. of cars (14 no. of cars in sample)	29 no. (35 no. in sample)	17 no. (17 no. in sample)
Maximum value	1,812 no. of cars (1,812 no. of cars in sample)	516 no. of cars (516 no. of cars in sample)	4,144 no. (4,144 no. in sample)	32 no. (32 no. in sample)
Minimum value	13 no. of cars (13 no. of cars in sample)	0 no. of cars (0 no. of cars in sample)	0 no. (0 no. in sample)	2 no. (2 no. in sample)
Standard variation	545 no. of cars (545 no. of cars in sample)	226 no. of cars (226 no. of cars in sample)	947 no. (964 no. in sample)	21 no. (21 no. in sample)
Robust variable for analysis	No (No)	No (No)	No (No)	No (No)

Source: Synergies analysis of State data collected by the CGC

Boxplots for each transport vehicle type are presented in the following subsections. All of the plots are heavily skewed owing to the dominance of capital cities in certain modes of transport.

The distribution of heavy rail cars for each SUA is displayed in Figure B.5. Consistent with the data on heavy track km, 10 SUAs have heavy rail cars, mainly either capital cities or other significant regional areas. This number is not necessarily unreasonable, given that commuter train transport is unlikely to be viable in most regional areas.

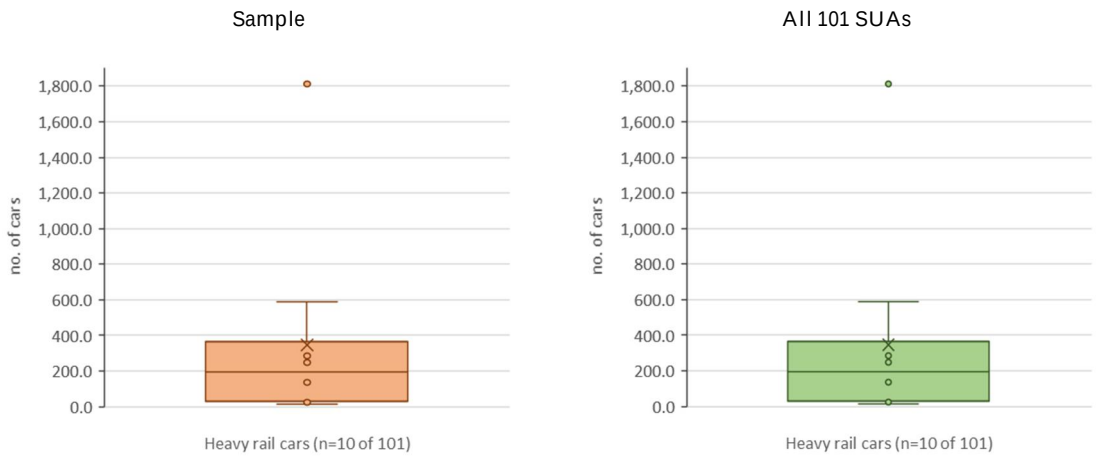


Figure B.5: Distribution of count of number of heavy rail cars

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

As discussed in the previous section, only 4 SUAs have a light rail system, as reflected in the shape of the boxplot in Figure B.6.

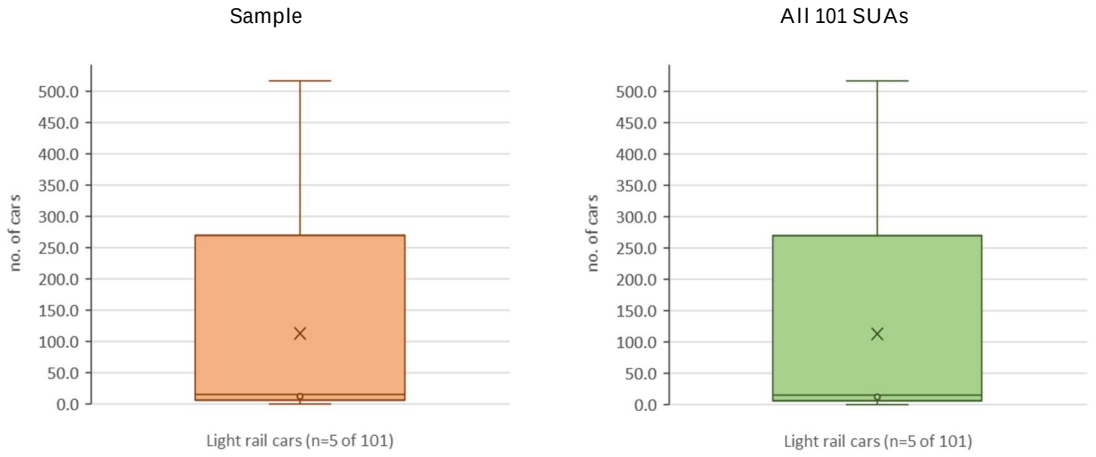


Figure B.6: Distribution of count of number of light rail cars

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Only 21 of the 74 SUAs report having any buses, and of these some of the values are lower than anticipated (e.g. 49 buses in the case of Brisbane).[32] This can possibly be attributed to the exclusion of some bus services (such as private contracts) in the data, the distribution of which is presented in Figure B.7. We also understand that the Northern Territory did not provide asset quantity data.

[32] According to the Brisbane City Council website, the size of its current fleet is more than 1,200 buses.

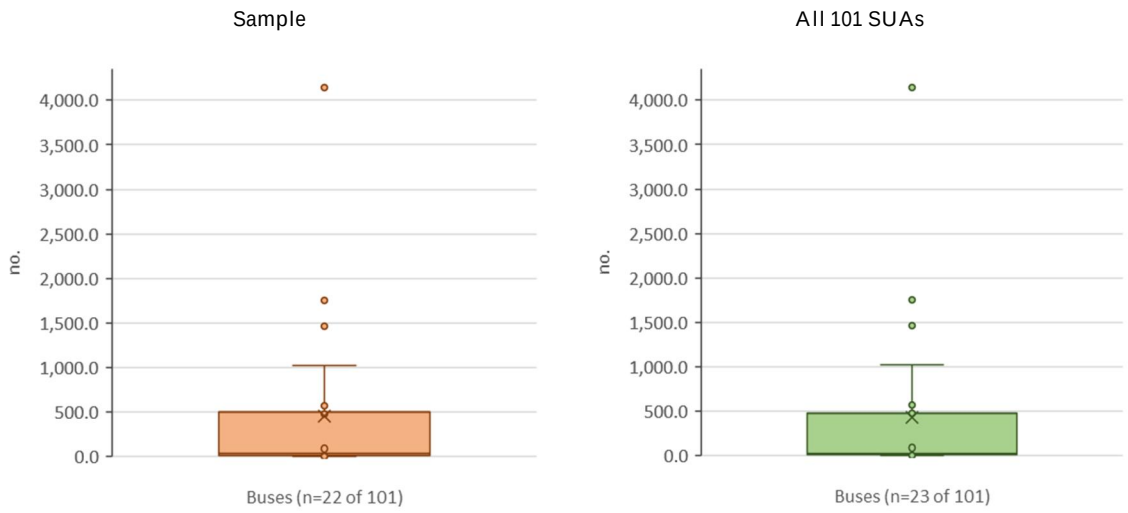


Figure B.7: Distribution of count of number of buses

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

In Figure B.8, Sydney (32 vessels) and Perth (2 vessels) were the only SUAs to report non-zero values, although in the case of Perth, there were no corresponding ferry wharves.

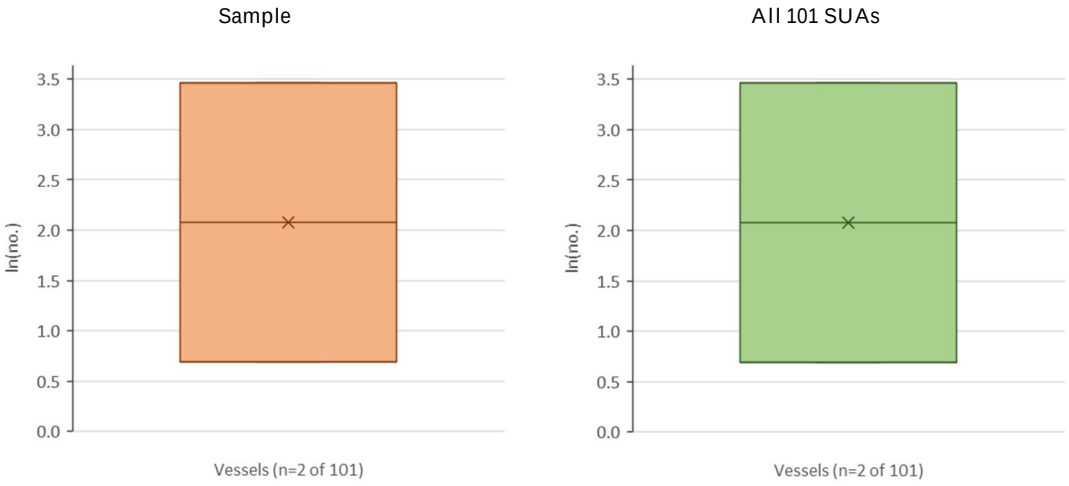


Figure B.8: Distribution of count of ferries

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Similar to the track/lane length and ferry wharf data, the unconventional appearance of the boxplots for transport vehicles reflects the quality of the data that is available. Specifically, it seems that the quantity of some assets is understated for certain SUAs. Given this paucity of data, it may be more appropriate to model these factors as dummy variables, which simply indicate the presence or absence of a particular mode of transport for each SUA.

Consolidated revenue km travelled

Consolidated revenue km travelled represents the distance travelled by public transport vehicles, from which fare revenue could be generated. Consequently, this variable is closely related to the service levels factor identified by the CGC. Our analysis has shown that the inclusion of this variable may present challenges when it comes to disentangling frequency and fleet effects. Further discussion of these issues is presented in Table B.3.

Table B.3: Consolidated revenue km travelled

Variable description	Revenue km travelled represents how far the vehicles in a public transport fleet travel.
Reason for inclusion	Rather than focusing simply on the size of a given fleet, revenue km travelled captures how far the fleet travels, thereby proxying for the level of service. One drawback of this is that it is difficult to differentiate between fleet and frequency; in effect, a single vehicle (such as a bus) could register higher revenue km travelled if it operates at a higher frequency.
Expectation	Our expectation is that higher revenue km travelled will be associated with higher transport expenditure. This is because more kilometres travelled requires higher levels of staffing and also increases depreciation and operating costs for transport vehicles.
Statistical level	SUA
Data modifications required	None, although there is considerable missing data, as highlighted below.
Maximum available observations	36 (28 in expense sample)
Average value	24 million km (31 million km in sample)
Median value	1 million km (2 million km in sample)
Maximum value	392 million km (392 million km in sample)
Minimum value	0 million km (0 million km in sample)
Standard variation	73 million km (81 million km in sample)
Robust variable for analysis	No (No)

Source: Synergies analysis of State data collected by the CGC

Data limitations could be another obstacle to the use of this variable. Only 36 SUAs report a non-zero value for consolidated revenue km travelled. The clear outliers are (in order) Sydney, Melbourne, Brisbane and Perth, all with consolidated revenue km travelled in excess of 80 million km. The next highest is Wollongong at 25 million km.

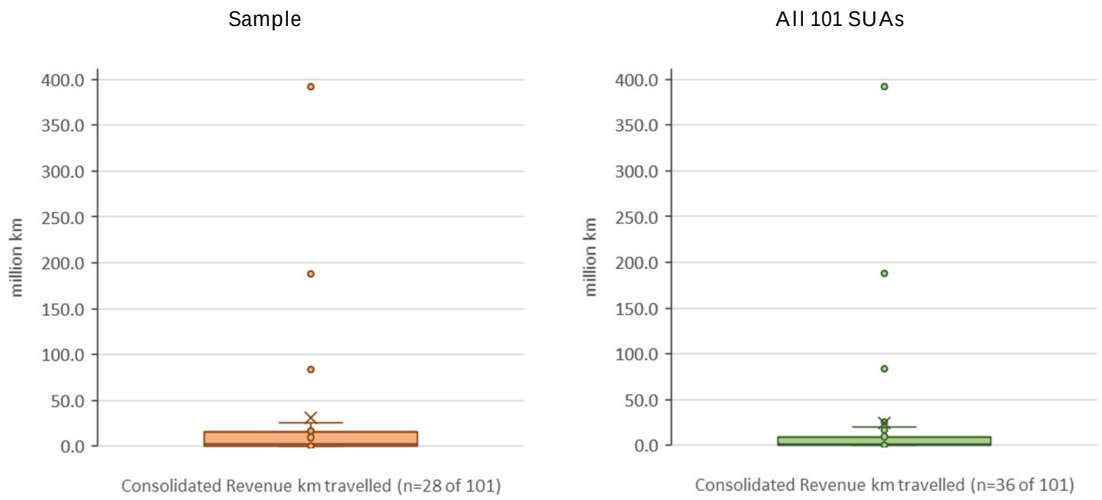


Figure B.9: Distribution of revenue kilometres travelled

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Again, as data is sparse it may be more appropriate to model this variable as dummy variables, which simply indicate the presence or absence of a particular mode of transport for each SUA.

As discussed in Section 2.4 relative demand for public transport mode and/or average distance to work can be used as proxy variables for kilometres travelled. Short assessments of both variables are presented below.

Relative demand by public transport mode

Demand for a certain mode in an SUA can be extracted from Census data on method of travel to work. There is the potential for these variables to be closely correlated with other candidate variables such as population and employment, so mode dummies can also be constructed that indicate the presence or absence of a mode in each SUA.

Table B.4: Relative demand by public transport mode

Reason for inclusion	Proxy variable for revenue kilometres travelled.			
Expectation	As rail and light rail require more complex infrastructure than buses or ferries, higher demand for these modes is likely to drive expenses up.			
Statistical level	SUA			
Data modifications required	None			
	Train (mode use level)	Bus (mode use level)	Ferry (mode use level)	Light rail (mode use level)
Maximum available observations	100 (70 in expense sample)	100 (70 in expense sample)	100 (70 in expense sample)	100 (70 in expense sample)
Average value	7,659 no. (10,877 no. in sample)	3,532 no. (4,992 no. in sample)	159 no. (218 no. in sample)	657 no. (937 no. in sample)

Median value	16 no. (19 no. in sample)	81 no. (100 no. in sample)	5 no. (7 no. in sample)	3 no. (3 no. in sample)
Maximum value	358,661 no. (358,661 no. in sample)	137,186 no. (137,186 no. in sample)	8,901 no. (8,901 no. in sample)	55,333 no. (55,333 no. in sample)
Minimum value	0 no. (0 no. in sample)	5 no. (5 no. in sample)	0 no. (0 no. in sample)	0 no. (0 no. in sample)
Standard variation	43,537 no. (51,813 no. in sample)	15,843 no. (18,786 no. in sample)	964 no. (1,148 no. in sample)	5,551 no. (6,629 no. in sample)
Robust variable for analysis	Yes (Yes)	Yes (Yes)	Yes (Yes)	Yes (Yes)

Source: Synergies analysis of Census 2016 data

Shown in logarithmic form, the boxplots for the samples of 70 (left panel) resembles that for all 101 closely for all four modes. This indicates that the sample of the 70 SUAs is very likely to be representative for all 101 SUAs.

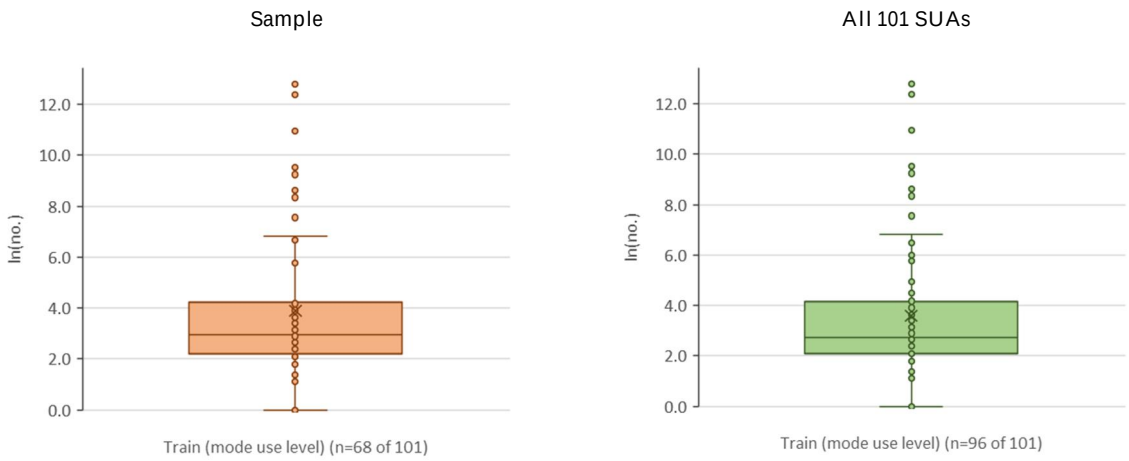


Figure B.10: Distribution of train demand levels (logarithmic form)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

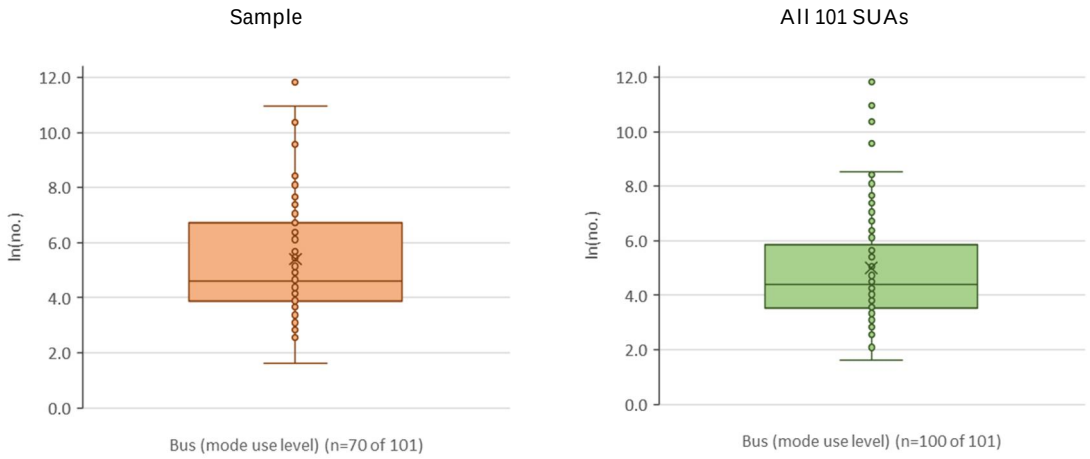


Figure B.11: Distribution of bus demand levels (logarithmic form)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

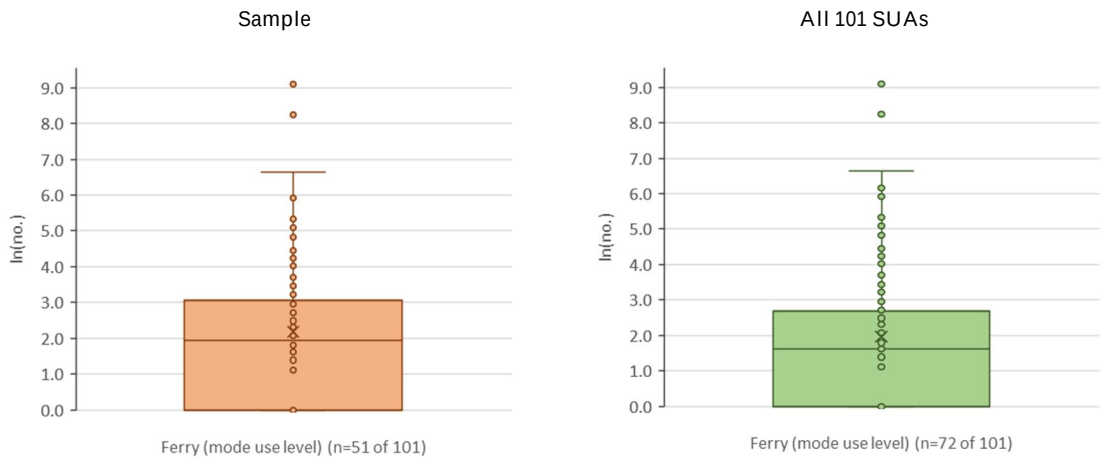


Figure B.12: Distribution of ferry demand levels (logarithmic form)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

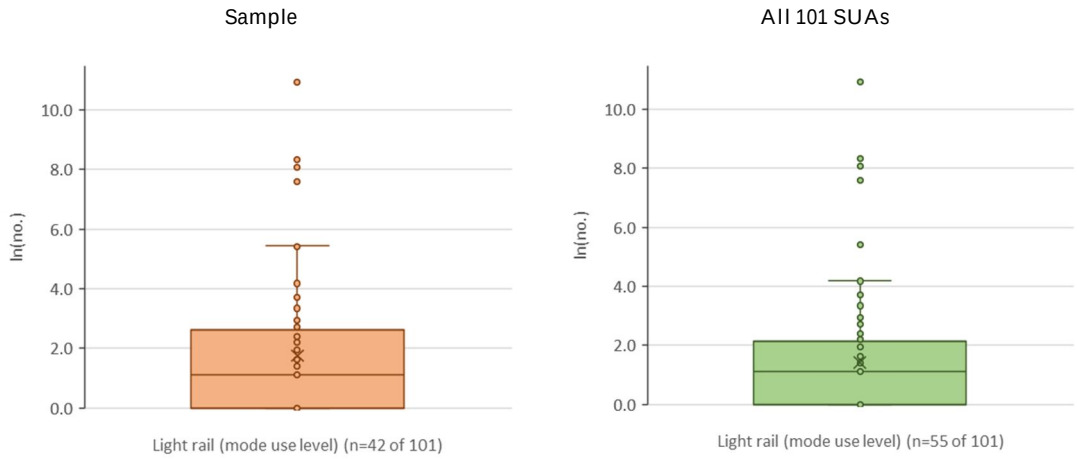


Figure B.13: Distribution of light rail demand levels (logarithmic form)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

Journey to work

Travel demand during morning and evening peak times are driven by commuters journeying to work. Table B.5 outlines the available data for distance travelled.

Table B.5: Journey to work

Variable description	Distance: The 2016 census collected data on the distance travelled to work for each SUA.
Reason for inclusion	Demand for public transport will be a function of the distance required to travel to work. This will have an impact not only on the volume of public transport required, but also on the appropriate modes of transport. It is a proxy variable for revenue kilometres travelled.

Expectation	The longer the distance required to journey to work, the more recurrent expenditure on transport would be expected. The relationship may not be this clear-cut though. Shorter travel distances may encourage greater public transport use, and longer distances may make use of a private vehicle more feasible.
Statistical level	SUA
Data modifications required	Responses are currently grouped into 12 intervals (e.g. nil distance, 0-1km, 1-2.5km etc.)
Maximum available observations	100 (70 in expense sample)
Average value	21 km (19 km in sample)
Median value	19 km (18 km in sample)
Maximum value	49 km (34 km in sample)
Minimum value	11 km (11 km in sample)
Standard variation	7 km (6 km in sample)
Robust variable for analysis	Yes (Yes)

Source: Synergies analysis of Census 2016 data

The distribution of distances to work are shown in Figure B.14. The highest estimated journey to work is for Morisset – Cooranbong (located between Central Coast and Newcastle), at 37.58km. The highest capital city journey to work is in Perth (25km).

The boxplot for the sample of 70 (left panel) resembles that for all 101 closely. There are more outliers when all SUAs are analysed, mostly relating to regional areas or SUAs on the periphery of metropolitan areas. On the whole, the sample of the 70 SUAs is very likely to be representative for all 101 SUAs.

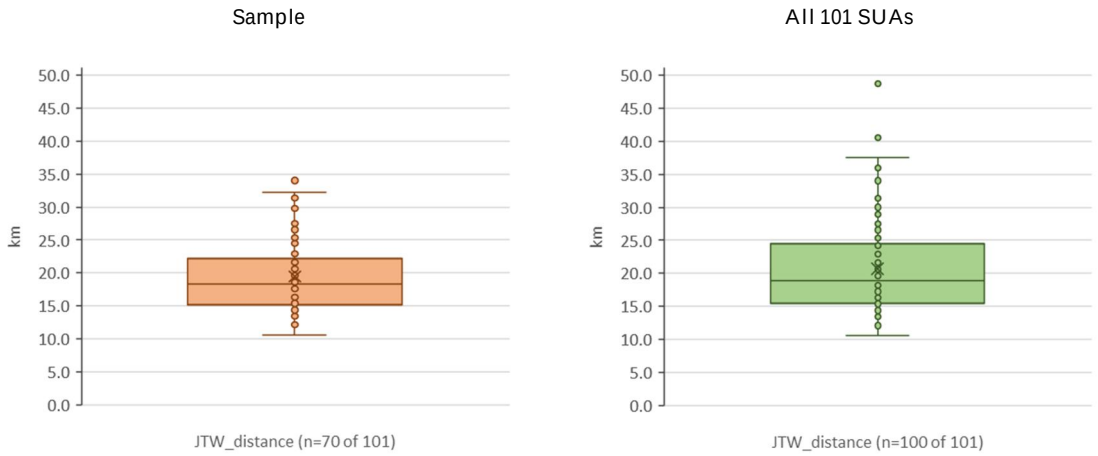


Figure B.14: Distribution of distances to work

Supply variable summary

Supply variables contribute to the modelling framework by providing an indication of service levels and how different modes of transport may result in different cost structures. Variables available for analysis include information on infrastructure by mode (heavy and light rail track, busway lane length and ferry wharves), transport vehicles by mode (trains, trams, buses and ferry vessels), and consolidated revenue km.

Data limitations are an important consideration for this category of variables. Asset quantities (such as track length, busway lane length and ferry wharves) are available for no more than 10 SUAs, most of which are capital cities. Data availability for corresponding transport vehicles is similarly constrained. As a result of this, data on public transport mode usage may offer the most viable way forward for proxying supply factors. This data is available for all SUAs, and it can also be modelled as a simple usage dummy (taking a value of 0 or 1 depending on whether a particular mode is used in an SUA).

Data on consolidated revenue km travelled has also been collected by the States, but there are insufficient observations for this to meet the criteria of a robust variable.

Appendix C. Detailed assessment of cost variables

Cost variables capture any factors pertaining to the urban form and landscape in an SUA, which may result in higher or lower transport-related expenditure compared to otherwise similar areas.

C.1 Key Stage 1 report findings

The Stage 1 report identifies a range of variables that can influence the costs associated with the provision of public transport that cannot be controlled by the public transport provider.

- **Urban congestion**
  The level of urban congestion has a significant impact on bus operating costs, particularly where bus routes share road space with general traffic for part or their entire route. Direct impacts of congestion include increased wear and tear on vehicles and tyres, from repeated stopping and starting in traffic, and reduced fuel efficiency. Indirectly, urban congestion contributes to persistent late-running services and unreliable travel times. Buses not only experience intersection and mid-block delays faced by general traffic, but also delays entering and exiting bus stops. The result is 'bus bunching', where regularly scheduled services instead run near-simultaneously. This can lead to poor utilisation of the existing services while incurring the same operational costs.
- **Urban density**
  When residential density is concentrated around train stations or other transport infrastructure, this increased accessibility to public transport is expected to attract higher public transport patronage. In many cases though, infrastructure projects in high density areas mean more property acquisition and high land value. Higher density can also trigger the development of higher cost options such as tunnel or elevated rail track options.
- **Urban terrain**
  Urban terrain is likely to cause variation in bus operating costs. The report presents a model showing that as road slope increases, bus operating costs also increase.

All three of these are well accepted impacts on costs. For example, in a review of bus costs in areas of Sydney for IPART it was concluded that: "Operating conditions in the four largest contract regions are characterised by high levels of traffic congestion, a high passenger density and a winding geographical topography that the efficient benchmark operator is not subject to. It was concluded that "these characteristics result in additional efficient hourly, kilometre and overhead costs in these regions".[33]

C.2 Variables

Population density and journey to work describe the urban form. Land slope variables describe variability in terrain, while road and railway bridge variables may capture the influence of waterways and other geographic characteristics that may shape transport networks.

Population density

Population density may provide a richer explanation of expenditure patterns than population as it can also be a proxy measure for the complexity of the required infrastructure. From a statistical point of view, population density might be favourable over population as it might cover a narrower range of values. Specifically, we have calculated population-weighted density, which is a widely-recognised measure of density that more closely captures true urban density.[34] One weakness of the conventional density measure is that it may understate the actual density for populated areas of an SUA if the SUA also contains significant portions of unpopulated land. Table C.1 provides a summary of our expectations regarding this variable.

[33] *Total Cost Review of Regular Bus Services Operated in Sydney's Four Largest Regions. Prepared for The Independent Pricing and Regulatory Tribunal NSW (IPART).* Indec. Sydney, 2009.

[34] Morton A.B. (2014) Population-Weighted Density, Density-Weighted Population, Granularity, Paradoxes: A Recapitulation, arXiv:1412.4332v2.

Table C.1: Population-weighted density

Variable description	Population-weighted density for each SUA is calculated as the sum of SA1 parcel densities weighted by the SA1 population share of the SUA.
Reason for inclusion	The concentration of people in a given area is as important as a basic measure of population, due to the potential impacts upon transport utilisation. The Stage 1 report noted that population density can play a role in decisions to provide underground or above ground infrastructure, which is also relevant for expenditure levels.
Expectation	Urban areas are expected to require higher expenditure, holding all else constant, as the added complexity of metropolitan transport networks may outweigh any scale benefits. As documented in the Stage 1 Report, higher residential density is associated with higher public transport patronage, which necessitates higher expenditure. Infrastructure construction in higher density areas typically requires more property reclamations at relatively high land values.
Statistical level	SUA
Data modifications required	The generalised formula for population-weighted density is as follows: Summation over SA1 regions of [(SA1 estimated resident population / SA1 area) * (SA1 population / SUA population)]
Maximum available observations	101 (70 in expense sample)
Average value	1,459 persons/sqkm (1,578 persons/sqkm in sample)
Median value	1,387 persons/sqkm (1,461 persons/sqkm in sample)
Maximum value	6,206 persons/sqkm (6,206 persons/sqkm in sample)
Minimum value	563 persons/sqkm (643 persons/sqkm in sample)
Standard variation	709 persons/sqkm (792 persons/sqkm in sample)
Robust variable for analysis	Yes (Yes)

Source: Synergies analysis of Census 2016 data

The distribution of population-weighted densities across all SUAs is presented in the right panel of Figure C.1. There are two notable outliers (Sydney and Melbourne), but otherwise population densities range between approximately 600 and 3,000 people/sq.km. The boxplot illustrates that the distribution is significantly skewed, with approximately two-thirds of SUAs having a population density of 1,500 people/sq.km or less.

The boxplot for the sample of 70 (left panel) resembles the boxplot for all 101 closely. This indicates that the sample of the 70 SUAs is very likely to be representative for all 101 SUAs.

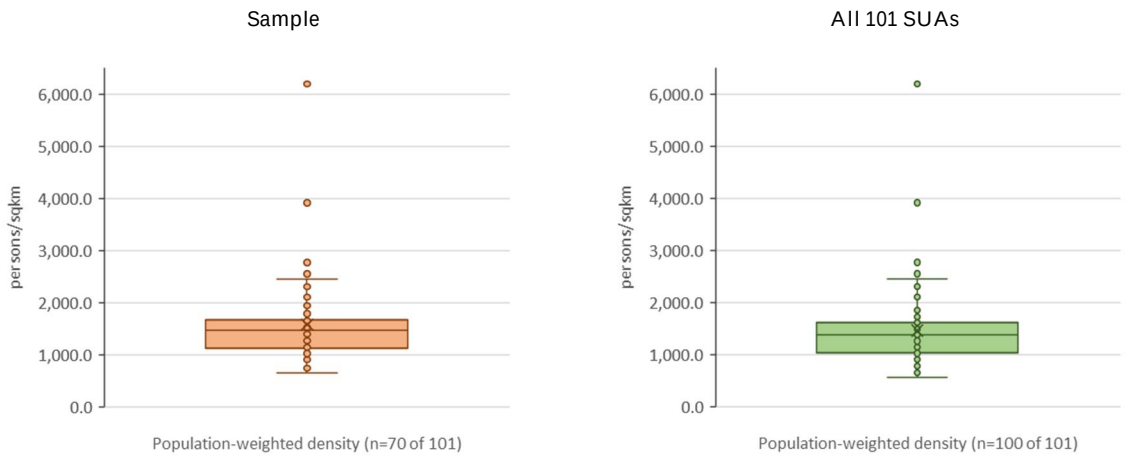


Figure C.1: Distribution of population-weighted density

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of Census 2016 data

Land slope indicators

Land slope indicators provide another perspective on recurrent expenditure, by incorporating geographical factors that may complicate transport networks.

Table C.2: Land slope indicators

Variable description	The land slope indicators consist of the following: Zero slope land area: Area of land within the GCCSA with a slope percent value of 0-2% Land slope mean: Mean of the slope segment values in degrees Land slope SD: Standard deviation of the slope segment values in degrees		
Reason for inclusion	Geographical indicators are likely to have an effect on the required complexity of infrastructure construction and hence drive investment volumes as well as recurring maintenance/replacement expenses. While it is unlikely that all three variables will appear in the final specifications, we will explore the relative impacts of the different indicators.		
Expectation	Our expectation is that a higher zero slope land area will be associated with lower recurrent expenditure. On the other hand, a higher mean slope segment is expected to add to recurrent expenditure. In addition, a higher standard deviation in land slope could be relevant to the extent that it implies more diverse terrain, which means that transport options need to be more flexible to accommodate this.		
Statistical level	SUA		
Data modifications required	None		
	Zero Slope Land Area	Land Slope Mean	Land Slope SD
Maximum available observations	101 (70 in expense sample)	101 (70 in expense sample)	101 (70 in expense sample)
Average value	68,246,682 square metres (87,964,039 square metres in sample)	3 degrees (3 degrees in sample)	3 standard deviation (3 standard deviation in sample)

Median value	23,403,667 square metres (30,530,316 square metres in sample)	2 degrees (2 degrees in sample)	2 standard deviation (2 standard deviation in sample)
Maximum value	1,074,412,589 square metres (1,074,412,589 square metres in sample)	11 degrees (11 degrees in sample)	9 standard deviation (9 standard deviation in sample)
Minimum value	1,272,720 square metres (1,272,720 square metres in sample)	0 degrees (0 degrees in sample)	0 standard deviation (0 standard deviation in sample)
Standard variation	151,509,737 square metres (177,329,123 square metres in sample)	2 degrees (2 degrees in sample)	2 standard deviation (2 standard deviation in sample)
Robust variable for analysis	Yes (Yes)	Yes (Yes)	Yes (Yes)

Source: Synergies analysis of Geoscience Australia data

The following figures present boxplots for the various land slope indicators. Figure C.2 presents a boxplot for zero slope land area. Shown in logarithmic form, the area of flat land seems to be evenly distributed. Furthermore, there appears to be sufficient variation in this variable to warrant consideration for regression analysis.

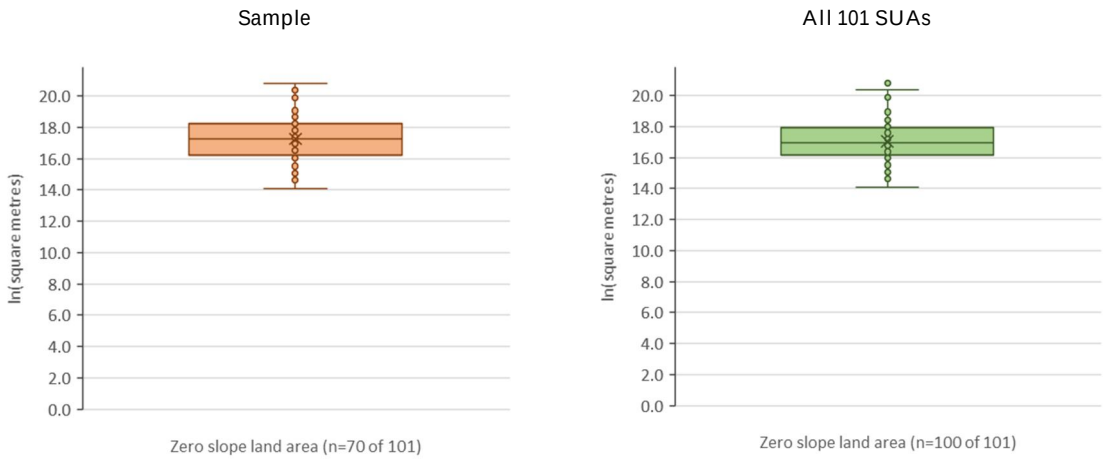


Figure C.2: Distribution of zero slope land area (logarithmic transformation)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Figure C.3 displays the mean slope segment for each SUA measured in degrees. Once again, the data in logarithmic form shows a relatively even distribution.

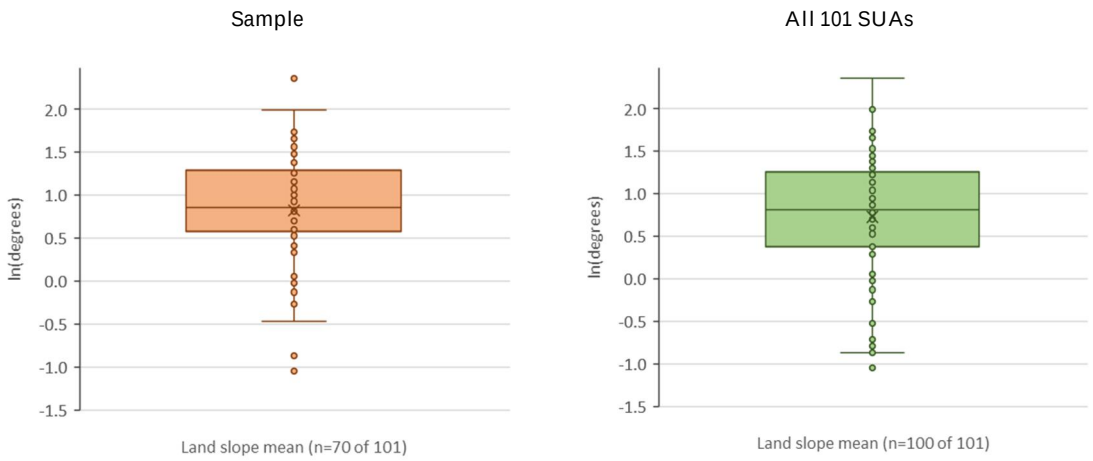


Figure C.3: Distribution of land slope mean (logarithmic transformation)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

As discussed in Figure C.4, land slope standard deviation will possibly serve as a useful proxy for landscape diversity.

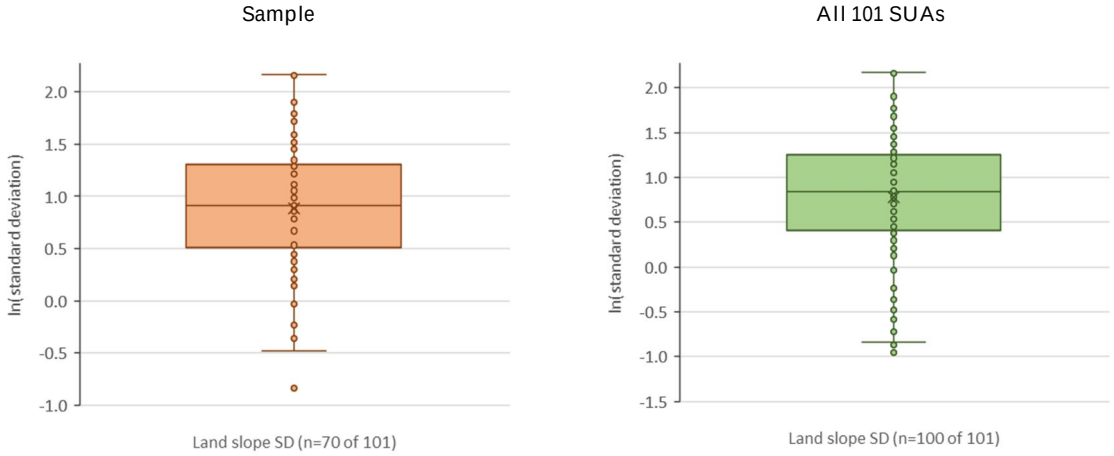


Figure C.4: Distribution of land slope standard deviation (logarithmic transformation)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Road and railway bridge indicators

Whereas land slope indicators consider the topography of each area more generally, the road and railway bridge indicators describe the status of infrastructure already in place.

Table C.3: Road and railway bridge indicators

Variable description	Indicators consist of the following: Road (Rail) Bridge Line (count): The number of intersections of bridges within the SUA (if the entire bridge in the same SUA - counts as 1 line for the bridge) Road (Rail) Bridge Point (count): The number of intersections of bridges within the SUA (if the entire bridge intersects 2 SUAs - counts as 1 point for the bridge in that SUA) Road (Rail) Bridge Line Dimension: Length of the bridge in metres					
Reason for inclusion	Infrastructure complexity will in part determine recurrent transport expenditure					
Expectation	Recurrent expenditure will be positively associated with bridge length and the number of bridge intersections					
Statistical level	SUA					
Data modifications required	None					
	Road Bridge Line (count)	Road Bridge Point (count)	Road Bridge line dimension	Rail Bridge Line (count)	Rail Bridge Point (count)	Rail Bridge line dimension
Maximum available observations	56 (43 in expense sample)	56 (43 in expense sample)	56 (43 in expense sample)	56 (43 in expense sample)	56 (43 in expense sample)	56 (43 in expense sample)
Average value	3 no. (4 no. in sample)	2 no. (3 no. in sample)	1,107 metres (1,343 metres in sample)	1 no. (1 no. in sample)	1 no. (1 no. in sample)	270 metres (322 metres in sample)
Median value	1 no. (1 no. in sample)	1 no. (1 no. in sample)	404 metres (475 metres in sample)	0 no. (0 no. in sample)	0 no. (0 no. in sample)	0 metres (0 metres in sample)
Maximum value	34 no. (34 no. in sample)	19 no. (19 no. in sample)	13,117 metres (13,117 metres in sample)	9 no. (9 no. in sample)	14 no. (14 no. in sample)	3,953 metres (3,953 metres in sample)
Minimum value	0 no. (0 no. in sample)	0 no. (0 no. in sample)	0 metres (0 metres in sample)	0 no. (0 no. in sample)	0 no. (0 no. in sample)	0 metres (0 metres in sample)
Standard variation	6 no. (6 no. in sample)	4 no. (4 no. in sample)	2,343 metres (2,626 metres in sample)	2 no. (2 no. in sample)	2 no. (2 no. in sample)	651 metres (732 metres in sample)
Robust variable for analysis	Yes (Yes)	Yes (Yes)	Yes (Yes)	Yes (Yes)	Yes (Yes)	Yes (Yes)

Source: Synergies analysis of Geoscience Australia data

There is no road or rail bridge data for 27 of the 74 SUAs in the expense sample.

The road bridge line data has an interquartile range of only 0 to 2 bridges. Only 5 (Sydney, Melbourne, Brisbane, Sunshine Coast, Central Coast) of the 48 SUAs with available data have more than 5 bridges in this category. 11 SUAs report a value of zero, and a further 15 report a value of 1.

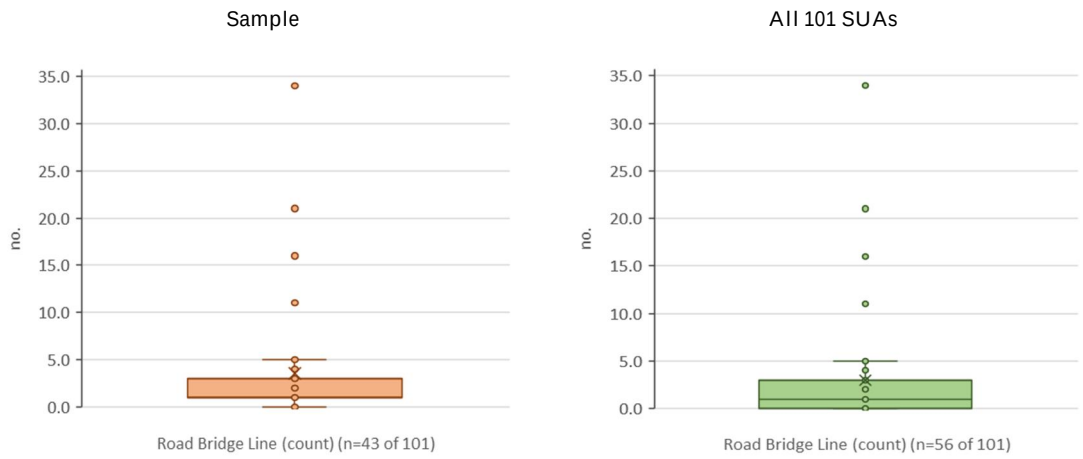


Figure C.5: Distribution of road bridge line (count)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

With regards to road bridge points, the interquartile range is only 0 to 1 bridges. Using this measure, Sydney, Melbourne, Brisbane, Adelaide and Cairns are the only SUAs with more than 4 bridges. 21 of the 48 SUAs report no bridges, and a further 10 report having only 1 bridge.

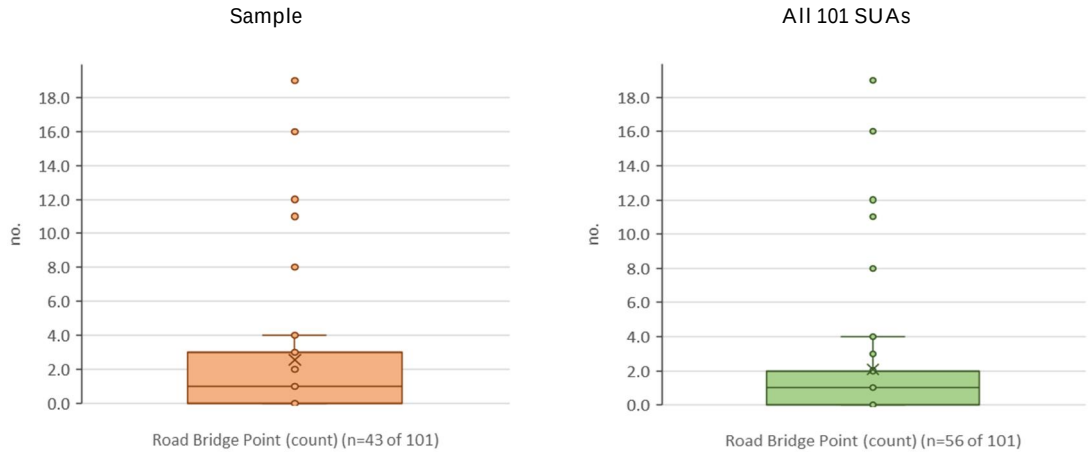


Figure C.6: Distribution of road bridge point (count)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Similar to the other indicators, high values for road bridge line dimension are concentrated in capital cities.

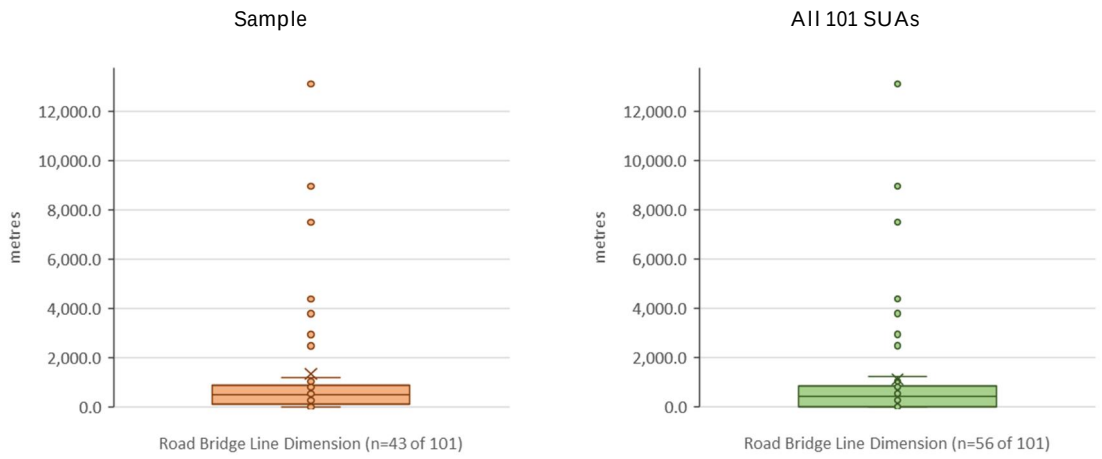


Figure C.7: Distribution of road bridge line dimension

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

One risk with econometric analysis based on these variables is that there may be insufficient variation in it to explain recurrent expenditure patterns. The road bridge line dimension variable could be more appropriate as it measures precise bridge lengths and takes on more diverse values.

Rail data follows a similar pattern, but with even less diversity in values among SUAs.

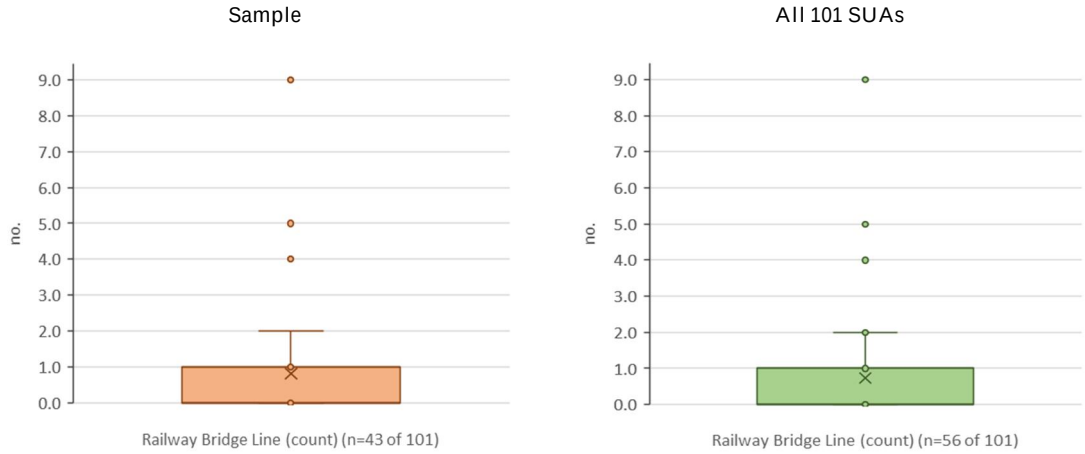


Figure C.8: Distribution of railway bridge line (count)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

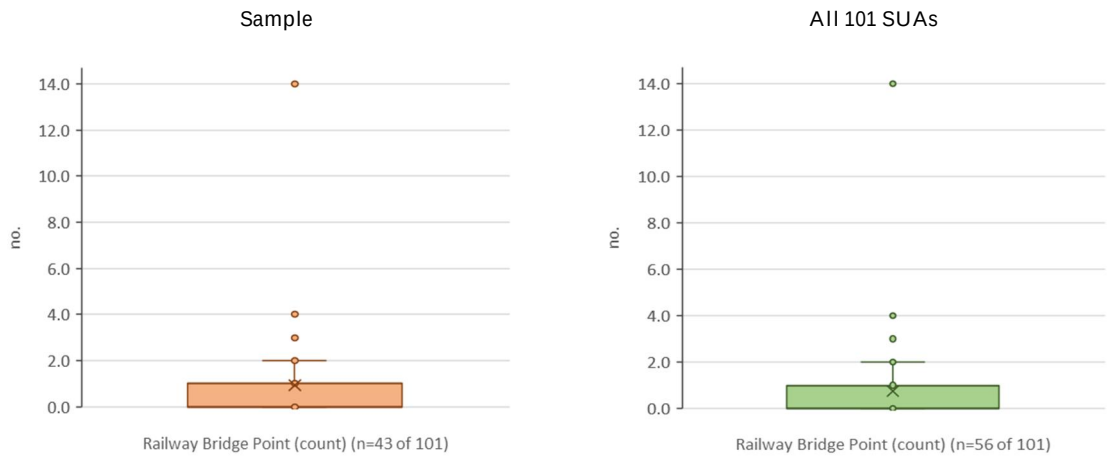


Figure C.9: Distribution of railway bridge point (count)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

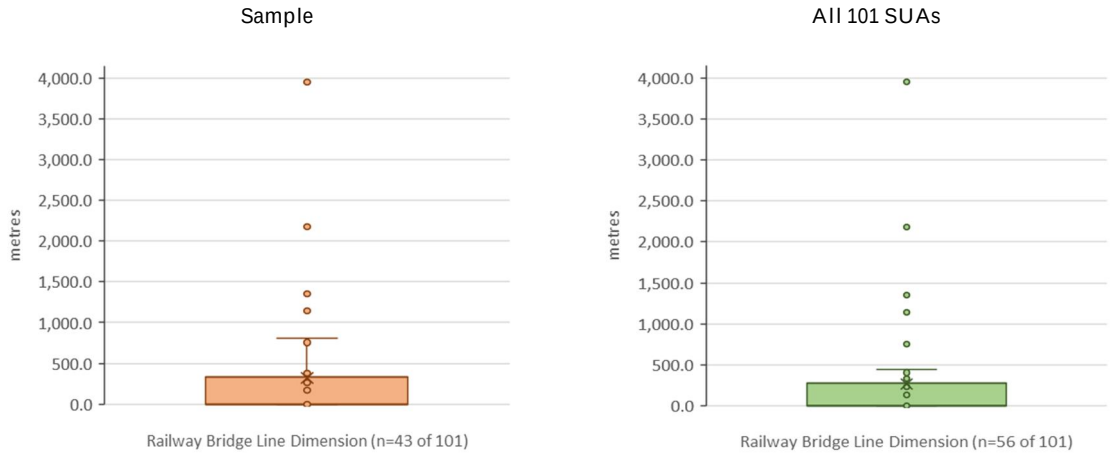


Figure C.10: Distribution of railway bridge line dimension

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

Railway segment indicators

Railway segment indicators are an alternative to road and railway bridge data. These variables focus more on the magnitude of elevation, as well as variability in elevation.

Table C.4: Railway segment indicators

Variable description	Indicators consist of the following: Railway segment slope degree: Percent slope of line segment Railway segment rise positive: Absolute value of elevation change across line segment in metres Railway segment rise length: Length of line segment in metres		
Reason for inclusion	If the slope of railway segments are either very steep and/or highly variable, this is likely to have an impact on transport expenditure.		
Expectation	Recurrent expenditure will be positively associated with degree of, and change in, elevation across railway segments.		
Statistical level	SUA		
Data modifications required	None		
	Railway segment slope degree	Railway segment rise positive	Railway segment rise length
Maximum available observations	81 (56 in expense sample)	81 (56 in expense sample)	81 (56 in expense sample)
Average value	1 percent slope (1 percent slope in sample)	1 metres (2 metres in sample)	588 metres (777 metres in sample)
Median value	1 percent slope (1 percent slope in sample)	1 metres (1 metres in sample)	189 metres (226 metres in sample)
Maximum value	3 percent slope (3 percent slope in sample)	5 metres (5 metres in sample)	8,474 metres (8,474 metres in sample)
Minimum value	0 percent slope (0 percent slope in sample)	0 metres (0 metres in sample)	5 metres (5 metres in sample)
Standard variation	0 percent slope (0 percent slope in sample)	1 metres (1 metres in sample)	1,355 metres (1,595 metres in sample)
Robust variable for analysis	Yes (Yes)	Yes (Yes)	Yes (Yes)

Source: Synergies analysis of Geoscience Australia data

Compared to the bridge indicators, the railway segment data is more widely reported, with only 14 of the expense sample SUAs missing observations.

In Figure C.11, Figure C.12 and Figure C.13, railway segment data exhibits greater variation, partly because it can take on non-integer values.

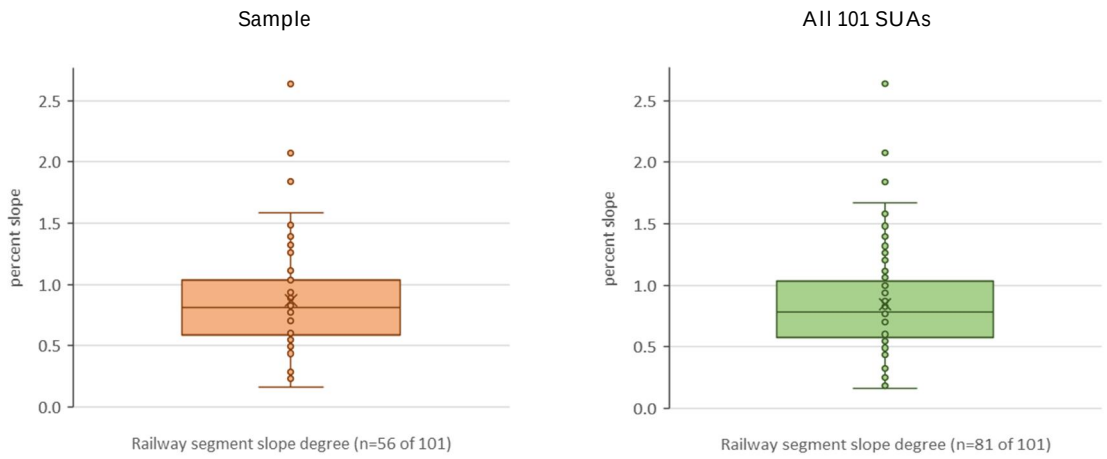


Figure C.11: Distribution of railway segment slope degree

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

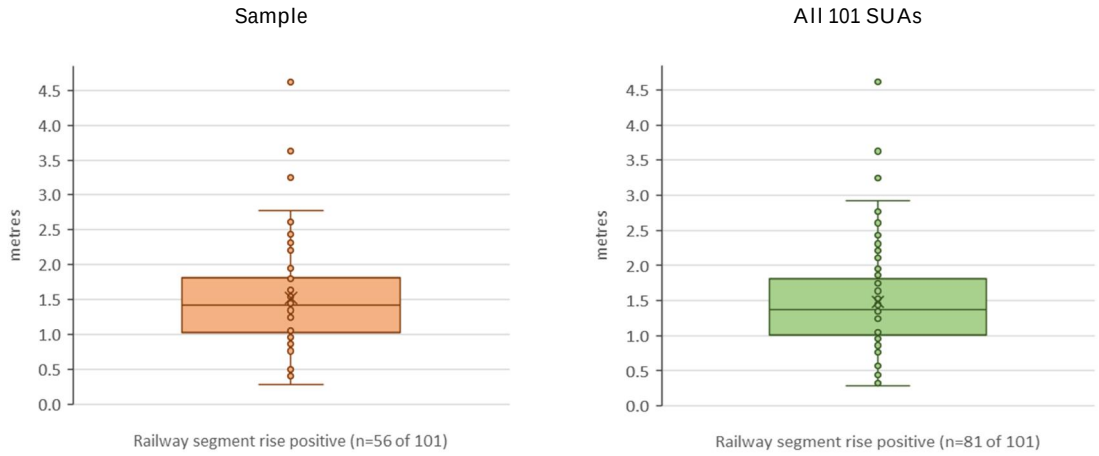


Figure C.12: Distribution of railway segment rise positive

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

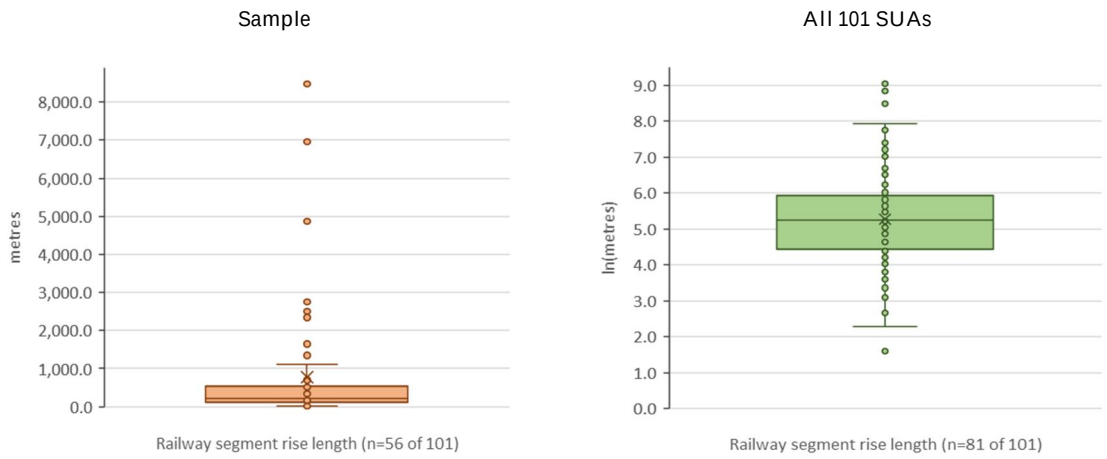


Figure C.13: Distribution of railway segment rise length (logarithmic transformation)

Note: Perth and Yanchep have been aggregated so the maximum number of SUA captured is 100

Data source: Synergies analysis of State data collected by the CGC

The number of missing observations for the road and rail bridge indicators is one of the reasons why those plots are quite heavily skewed. On the other hand, the railway segment data still exhibits some asymmetry, but the data is less incomplete, allowing a more balanced distribution to be derived. With regards to potential regression specifications, the key question will be whether the railway segment data is sufficiently differentiated from the standard land slope data, to warrant including both in a regression. We explore this in the following section.

Cost variable summary

The objective of using cost variables is to identify factors relating to urban form or landscape that may cause transport-related expenditure to be higher or lower relative to otherwise similar areas.

The Stage 1 report identified congestion as a determinant of operating costs. We do not have sufficient data on congestion – the available data is for capital cities only, which offers too few observations for a regression. However, we have illustrated its co-movement with proxy variables, such as employment and public transport passengers.

Population-weighted density is anticipated to provide a more comprehensive indicator of transport requirements than either a standard measure of density or raw population. It is calculated based on a population-weighted average of the density of all SA1 parcels in a given SUA. This variable will feature prominently in the econometrics analysis because of its dual purpose as a demand driver.

There are three candidates for land slope variables. The data underpinning these variables is complete for all but 4 SUAs in the expense sample, and they have a strong theoretical basis. At the same time though, the correlation analysis in Section 5 will demonstrate that only one variable from this set will be required, so as to avoid multicollinearity.

The cost variables also include an array of road and railway bridge indicators. The challenge with these variables is that there are many missing observations, including for SUAs in the expense sample. In some cases, there are fewer than 50 observations. This may be sufficient for econometric modelling, but it does mean that several SUAs are dropped from the analysis and none of their information whatsoever contributes to the regression.

Appendix D. Correlation between variables

The extent of any multi-collinearity is typically evaluated using correlation coefficients between variables. Commonly, a threshold of 0.8 is chosen, with correlation coefficients above this level symptomatic of multicollinearity. For the purpose of Figure D.1, a threshold of 0.7 has been specified. This lower threshold has been chosen because a smaller sample will amplify the consequences of correlation between variables. With a smaller sample, the precision of the estimates (as measured by their standard errors) will already be weaker, holding all else constant.

The colour-coding of the correlation matrix reflects the severity of the correlation. Independent variables with low correlation coefficients (defined as between -0.35 and 0.35) have been shaded in green. Including these pairs of variables in the same regression should be possible without unnecessarily inflating standard errors. Correlation coefficients between 0.35 and 0.70 (or between -0.35 and -0.70) have been assigned yellow shading, indicating that caution should be taken when these pairs of variables are included in the same regression. Red shading indicates correlation above 0.70 (or below -0.70). As discussed above, correlation above this level is likely to significantly hamper the precision of the estimates.

Several important insights can be drawn from the correlation matrix. In column 1, SEIFA appears to be highly correlated with income, which is to be expected. In column 2, population is perfectly correlated with employment, and it is also strongly correlated with consolidated revenue km; heavy rail track; rail and bus asset quantities; mode use variables (when measured in levels); zero slope land area; various road and rail bridge indicators; avoidable congestion cost; and metropolitan public transport task. In the third column, population density is strongly correlated with population, employment, consolidated revenue km, heavy rail track, heavy rail cars and buses, as well as the train, bus and ferry mode use variables. It is also correlated with a range of road and railway indicators.

In column 4, employment exhibits correlation with school enrolments and heavy rail track length, as well as heavy rail cars and buses, and multiple rail and road bridge variables. Employment has been found to be concentrated in capital cities, as are school enrolments and the various transport indicators. In turn, column 7 shows that enrolments are correlated with many of these same factors.

High correlation is also observed between the quantity of transport infrastructure and the quantity of transport vehicles. The correlation between heavy rail track length and the quantity of heavy rail cars is 0.89, which suggests that only one of these variables in necessary in a regression. Heavy track length is also correlated with the number of buses and ferry vessels, as well as many of the bridge indicators. Meanwhile, tram cars and light rail track are perfectly correlated with each other; this means it is not statistically possible to include both of these variables in the same regression.

The mode use variables in levels are strongly correlated with population, employment and enrolments, as well as consolidated revenue km and several transport vehicle and infrastructure quantity variables. Tram mode use in levels is perfectly correlated with light rail track and the quantity of tram cars, which is an intuitive result.

Land slope mean, and land slope deviation are almost perfectly correlated with each other (correlation coefficient = 0.92). Thus, there is likely to be no need to have regard to both of these variables in any one regression specification. Likewise, the red and yellow shaded cells in columns 28-33 show that several of the bridge indicators are correlated with each other, which suggests that they all capture similar variation in bridge infrastructure.

Acknowledging any data limitations, consolidated revenue km in its current form is highly correlated with employment, enrolments, heavy rail track and rollingstock, ferry wharves and vessel, buses, and various road and railway bridge indicators.

It is also important to point out which variables are not overly correlated with each other. Although income is correlated with SEIFA, and to a lesser extent with population density and zero slope land area, it is relatively uncorrelated with most other variables. Similarly, journey to work distance is relatively uncorrelated with the rest of the variables in the data.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1 SEIFA	1.00																																					
2 Population	0.41	1.00																																				
3 Poulation density	0.53	0.81	1.00																																			
4 Employment	0.41	1.00	0.81	1.00																																		
5 Income	0.71	0.10	0.33	0.11	1.00																																	
6 JTW_distance	0.34	-0.05	-0.06	-0.06	0.14	1.00																																
7 Enrolments	0.28	0.61	0.66	0.61	0.08	0.07	1.00																															
8 Consolidated Revenue km travelled	0.35	0.92	0.81	0.93	0.09	-0.05	0.66	1.00																														
9 Heavy Rail track	0.36	0.98	0.80	0.98	0.08	-0.05	0.63	0.96	1.00																													
10 Light Rail track	0.20	0.66	0.39	0.65	0.03	-0.05	0.02	0.45	0.59	1.00																												
11 Busway	0.09	0.17	0.13	0.16	0.00	-0.05	0.18	-0.01	0.11	0.05	1.00																											
12 Ferry Wharf	0.21	0.63	0.66	0.64	0.06	-0.06	0.64	0.87	0.73	0.04	-0.01	1.00																										
13 Rolling Stocks	0.31	0.82	0.77	0.82	0.08	-0.04	0.74	0.95	0.89	0.18	0.06	0.92	1.00																									
14 Trams cars	0.19	0.64	0.37	0.63	0.03	-0.05	0.00	0.42	0.57	1.00	0.03	0.01	0.15	1.00																								
15 Buses	0.36	0.88	0.80	0.88	0.10	-0.06	0.66	0.95	0.91	0.40	0.21	0.85	0.90	0.37	1.00																							
16 Vessels	0.22	0.64	0.67	0.65	0.06	-0.05	0.65	0.87	0.74	0.04	-0.01	1.00	0.93	0.01	0.86	1.00																						
17 Train (mode use level)	0.32	0.94	0.79	0.94	0.08	-0.06	0.60	0.98	0.97	0.58	0.02	0.81	0.89	0.56	0.93	0.82	1.00																					
18 Bus (mode use level)	0.38	0.87	0.79	0.88	0.12	-0.06	0.76	0.93	0.91	0.24	0.20	0.85	0.97	0.21	0.91	0.86	0.88	1.00																				
19 Ferry (mode use level)	0.29	0.76	0.72	0.77	0.08	-0.05	0.72	0.91	0.84	0.11	0.01	0.92	0.98	0.09	0.82	0.92	0.85	0.95	1.00																			
20 Tram (mode use level)	0.20	0.66	0.40	0.66	0.03	-0.05	0.03	0.45	0.59	1.00	0.06	0.05	0.19	1.00	0.41	0.05	0.59	0.25	0.12	1.00																		
21 Train (mode use dummy)	0.53	0.49	0.46	0.47	0.08	0.34	0.49	0.38	0.42	0.23	0.21	0.20	0.36	0.22	0.40	0.21	0.35	0.41	0.30	0.24	1.00																	
22 Bus (mode use dummy)	0.54	0.38	0.52	0.37	0.35	0.15	0.25	0.28	0.31	0.17	0.15	0.15	0.26	0.16	0.30	0.16	0.26	0.32	0.24	0.17	0.41	1.00																
23 Ferry (mode use dummy)	0.33	0.81	0.60	0.81	0.11	-0.02	0.53	0.73	0.79	0.46	0.00	0.44	0.66	0.45	0.66	0.46	0.72	0.72	0.64	0.46	0.34	0.34	1.00															
24 Tram (mode use dummy)	0.30	0.74	0.67	0.74	0.05	-0.05	0.42	0.64	0.69	0.56	0.49	0.49	0.56	0.54	0.69	0.49	0.69	0.62	0.50	0.57	0.41	0.30	0.42	1.00														
25 Zero slope land area	0.44	0.89	0.67	0.88	0.18	0.00	0.46	0.66	0.80	0.70	0.27	0.25	0.52	0.69	0.64	0.27	0.69	0.65	0.45	0.71	0.49	0.47	0.78	0.65	1.00													
26 Land slope mean	0.02	0.07	0.11	0.07	-0.16	0.15	0.05	0.09	0.07	-0.02	0.02	0.09	0.10	-0.02	0.10	0.09	0.07	0.09	0.09	-0.02	0.15	0.02	0.00	0.08	-0.05	1.00												
27 Land slope SD	0.09	0.10	0.21	0.10	-0.06	0.15	0.08	0.11	0.09	0.01	0.07	0.11	0.12	0.00	0.13	0.11	0.09	0.11	0.10	0.01	0.14	0.11	0.01	0.15	0.02	0.89	1.00											
28 Road Bridge Line (count)	0.31	0.75	0.69	0.76	0.04	0.04	0.61	0.81	0.80	0.25	0.05	0.72	0.84	0.23	0.70	0.72	0.77	0.83	0.85	0.25	0.40	0.36	0.59	0.48	0.55	0.11	0.16	1.00										
29 Road Bridge Point (count)	0.32	0.72	0.60	0.73	0.09	-0.07	0.51	0.62	0.72	0.38	0.58	0.49	0.64	0.36	0.64	0.48	0.64	0.73	0.62	0.39	0.36	0.36	0.50	0.69	0.66	0.07	0.15	0.68	1.00									
30 Road Bridge Line Dimension	0.34	0.84	0.74	0.84	0.08	-0.01	0.59	0.86	0.87	0.41	0.01	0.69	0.84	0.39	0.73	0.69	0.84	0.84	0.85	0.42	0.39	0.36	0.67	0.53	0.65	0.12	0.15	0.96	0.67	1.00								
31 Railway Bridge Line (count)	0.25	0.81	0.70	0.82	0.04	-0.06	0.56	0.84	0.85	0.41	0.05	0.70	0.81	0.39	0.74	0.70	0.83	0.82	0.82	0.42	0.32	0.29	0.63	0.55	0.62	0.08	0.15	0.86	0.64	0.89	1.00							
32 Railway Bridge Point (count)	0.26	0.75	0.52	0.74	0.08	-0.03	0.17	0.58	0.69	0.91	0.17	0.23	0.36	0.90	0.55	0.23	0.69	0.42	0.30	0.91	0.25	0.27	0.50	0.64	0.73	0.03	0.07	0.42	0.61	0.54	0.55	1.00						
33 Railway Bridge Line Dimension	0.27	0.81	0.73	0.82	0.05	-0.08	0.55	0.88	0.85	0.45	0.00	0.76	0.82	0.42	0.80	0.76	0.88	0.81	0.81	0.45	0.32	0.26	0.60	0.58	0.57	0.10	0.16	0.82	0.59	0.87	0.94	0.57	1.00					
34 Railway segment slope degree	0.08	0.13	0.14	0.13	-0.06	0.09	0.04	0.14	0.13	0.04	0.01	0.11	0.15	0.03	0.14	0.11	0.12	0.14	0.13	0.04	0.20	0.11	0.06	0.11	0.06	0.54	0.42	0.13	0.07	0.17	0.13	0.03	0.12	1.00				
35 Railway segment rise positive	0.08	0.13	0.14	0.13	-0.06	0.09	0.04	0.14	0.13	0.04	0.01	0.11	0.15	0.03	0.14	0.11	0.12	0.14	0.13	0.04	0.20	0.11	0.06	0.11	0.06	0.54	0.42	0.13	0.07	0.17	0.13	0.03	0.12	1.00	1.00			
36 Railway segment rise length	0.38	0.96	0.82	0.96	0.08	-0.04	0.61	0.91	0.96	0.57	0.17	0.65	0.85	0.55	0.86	0.66	0.91	0.88	0.79	0.58	0.49	0.40	0.76	0.69	0.81	0.15	0.18	0.78	0.74	0.86	0.82	0.68	0.81	0.26	0.26	1.00		
37 Avoidable congestion cost	0.36	0.98	0.79	0.99	0.09	-0.07	0.64	0.96	0.99	0.59	0.13	0.73	0.88	0.57	0.92	0.74	0.97	0.92	0.83	0.60	0.40	0.30	0.81	0.70	0.81	0.05	0.07	0.77	0.71	0.85	0.83	0.70	0.84	0.12	0.12	0.94	1.00	
38 Metropolitan public transport task	0.34	0.96	0.79	0.96	0.09	-0.07	0.60	0.98	0.98	0.61	0.06	0.78	0.88	0.59	0.92	0.78	1.00	0.89	0.84	0.62	0.36	0.27	0.75	0.70	0.74	0.06	0.08	0.77	0.67	0.85	0.84	0.71	0.87	0.12	0.12	0.92	0.99	1.00

Legend

Green: Between -0.35 and 0.35

Yellow: Between 0.35 and 0.70 (or between -0.35 and -0.70)

Red: greater than 0.70 or less than -0.70

Figure D.1: Correlation matrix

Note: Numbers in column headings correspond to those in the row labels

Data source: Synergies analysis

Appendix E. Econometric analysis: Technical details

In order to establish a set of candidate models, we have tested numerous models within the framework set out in Section 1.1, covering many permutations of explanatory variables and functional forms. In this Appendix, we do not present the results for all of these estimations. Rather we present results only for models that we consider promising for the purpose of informing future funding shares.

The following presents each identified model by summarising the theory behind it and which aspects it covers. We then present key test statistics for the model, coefficient estimates and their test statistics and illustrate the model's quality of fit in an actual vs. fitted values plot as well as a plot of the residual overall and by State. The last plot will be of particular interest as it assesses whether a particular model is likely to favour a particular State. Each model assessment concludes by a short interpretation of the key findings.

The model with population as the only explanatory variable is presented first as it will be used as the reference case to assess the improvement in performance and quality of fit achieved by the alternative models. To be considered as a preferred model, an alternative model must be at least as good as the population model.

E.1 Reference model

The reference model uses population as the only explanatory variable. We tested a linear form and one with population in natural logarithms.

Formally the models can be specified as:

$$exp_i = \beta_0 + \beta_1 pop_i + \varepsilon_i \qquad (0a)$$

$$exp_i = \beta_0 + \beta_1 \ln(pop_i) + \varepsilon_i \qquad (0b)$$

Table E.1: Reference Model

	Linear model (0a)	**Linear – log model (0b)**
Observations	70	70
F statistic	80.95	120.36
Prob > F	0.0000	0.0000
R^2	0.54	0.64
Adjusted R^2	0.54	0.63
Akaike information criterion	844	828
Bayesian information criterion	849	832
Root MSE	99	88

Source: Synergies modelling

Table E.2: Reference Model

	Coefficient estimate	**Standard error**	**95% confidence interval**	
Linear model (0a)				
Intercept	68.62919***	12.55391	104.3105	163.7658
pop_i	134.0382***	14.89759	43.57825	93.68014
Linear – log model (0b)				
Intercept	335.0252***	23.44174	288.2479	381.8025
$\ln(pop_i)$	78.49471***	7.154926	64.21728	92.77214

******* p > |t| <=0.1%
****** p > |t| <=1%
***** p > |t| <=5%
^ p > |t| <=10%

Source: Synergies modelling

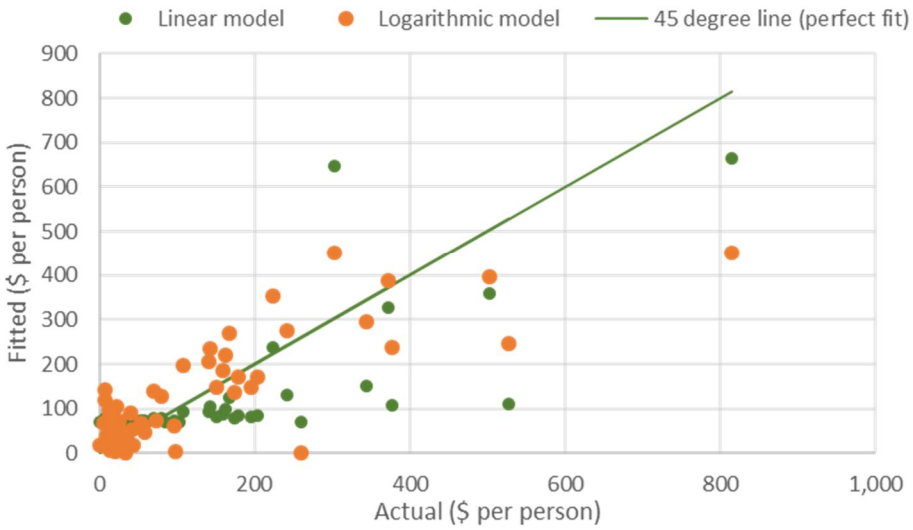


Figure E.1: Actual vs. fitted values for the two models

Data source: Synergies modelling

The actual vs. fitted plot shows that the linear population model reproduces the data relatively poorly as most values are below the 45-degree line. This is confirmed by its R^2 value that is significantly lower than that of the logarithmic form. The logarithmic form tends to underestimate expense per person for SUAs with higher values. Its coefficient estimates are similar to the model specified in the 2010 Review. As discussed throughout this report, since population is on both sides of the equation, the high significance of the coefficient estimates and the reasonably high R^2 of 0.65 were to be expected.

As Figure E.2 shows, its estimates are unbiased on average overall and for all States with more than three observations. However, New South Wales and Victoria show significant downwards but very little upwards deviation. This is further indication that the model does not perform particularly well for cities with high per capita expenses.

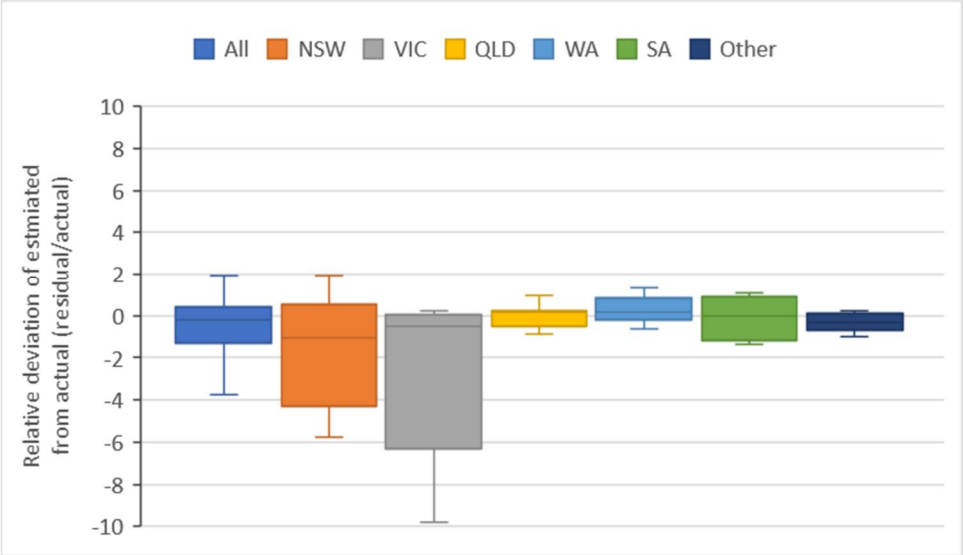


Figure E.2: Plot of relative residuals overall and by State

Data source: Synergies modelling

Despite its reasonably good fit and relatively high R^2 value, the reference model appears to have three key shortcomings:

- It oversimplifies the functional relationship and does not explain the composition of per capita expenses.
- It underestimates expenses per person for SUAs with higher values.
- It performs much poorer for New South Wales and Victoria than it does in the other States.

We will consider an alternative model a candidate for the preferred model if its fit is at least as good as that of the reference model and it improves on at least one of the three shortcomings.

E.2 Model 1

Model 1 uses density ($dense_i$) to depict demand, distance to work ($dist_i$) to represent network complexity, passengers by public transport mode ($pax_{i,mode}$) to represent availability and congestion, and mean land slope ($slope_i$) to account for topography. We tested a linear form and one with the passenger numbers in natural logarithms. Formally the models can be specified as:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 pax_{i,train} + \beta_5 pax_{i,bus} + \varepsilon_i \quad (1a)$$

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 \ln(pax_{i,train}) + \beta_5 \ln(pax_{i,bus}) + \varepsilon_i \quad (1b)$$

Table E.3: Model 1: Test statistics

	Linear model (1a)	**Linear – log model (1b)**
Observations	70	70
F statistic	39.68	48.44
Prob > F	0.0000	0.0000
R^2	0.76	0.79
Adjusted R^2	0.74	0.77
Akaike information criterion	808	798
Bayesian information criterion	822	811
Root MSE	75	69

Source: Synergies modelling

Table E.4: Model 1: Coefficient estimates

	Coefficient estimate	**Standard error**	**95% confidence interval**	
Linear model (1a)				
Intercept	-229.4174***	47.3626	-324.0351	-134.7997
$dense_i$	0.1113166***	0.0222505	0.0668661	0.1557671
$dist_i$	6.457876***	1.641414	3.178775	9.736977
$slope_i$	7.956897	5.237084	-2.505381	18.41918
$pax_{i,train}$	-0.0007752^	0.0004006	-0.0015755	0.0000252
$pax_{i,bus}$	0.0039925***	0.0011005	0.0017941	0.0061909
Linear – log model (1b)				
Intercept	-154.5637**	46.8811	-248.2194	-60.90792
$dense_i$	0.0715307***	0.0200746	0.0314271	0.1116343
$dist_i$	3.411582*	1.647887	0.1195494	6.703616
$slope_i$	6.963933	4.882911	-2.790803	16.71867
$\ln(pax_{i,train})$	18.07401***	4.036532	10.01011	26.13791
$\ln(pax_{i,bus})$	6.719857	6.659917	-6.584856	20.02457

*** p > |t| <=0.1%
** p > |t| <=1%
* p > |t| <=5%
^ p > |t| <=10%

Source: Synergies modelling

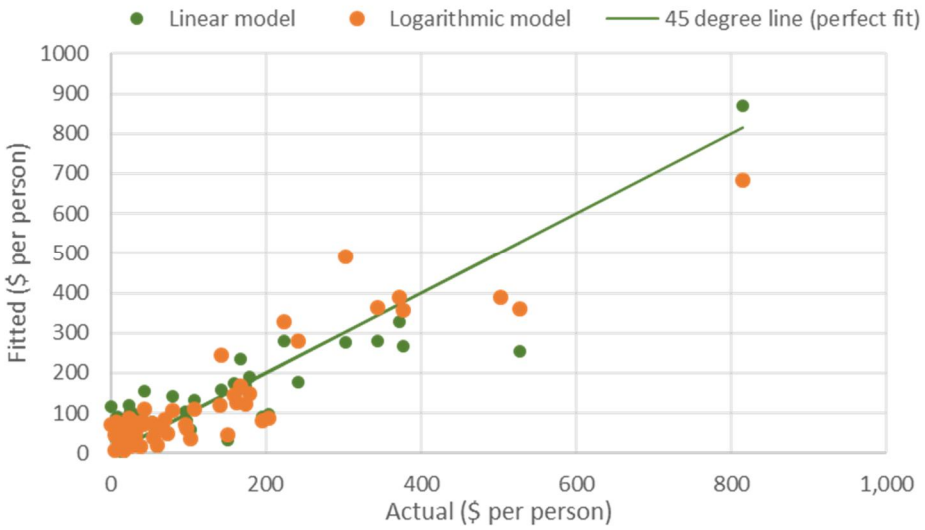
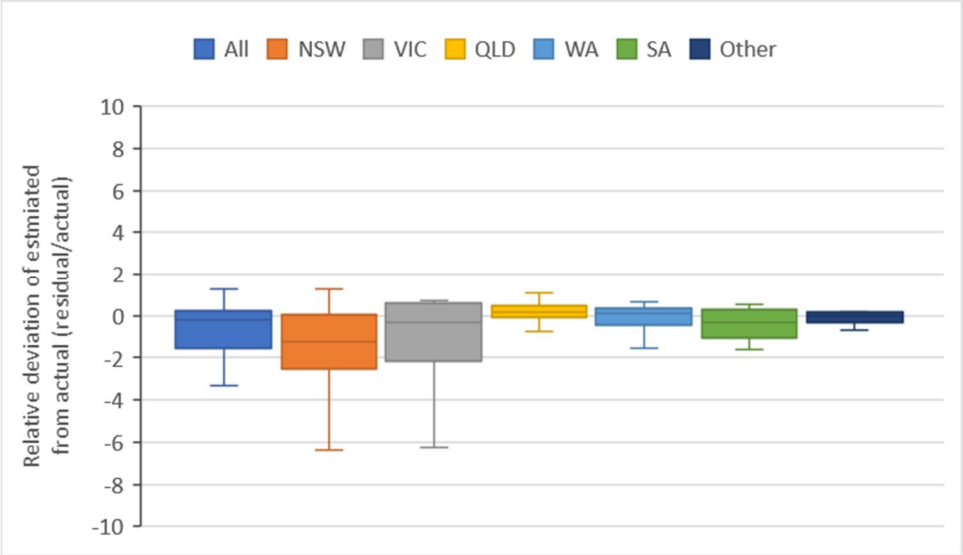


Figure E.3: Actual vs. fitted values for the two models

Data source: Synergies modelling

The actual vs. fitted plot shows a reasonable fit for both the linear and logarithmic form of the model. However, the test statistics prefer the logarithmic model. Therefore, we will consider 1b the better model and assess it further.

As Figure E.4 shows, its estimates are unbiased overall and for all States with more than three observations (grouped as other). The medians for New South Wales and Victoria are weighed down by the SUAs of Albury-Wodonga, Coffs Harbour, Goulburn, St Georges Basin - Sanctuary Point and Warrnambool which all have per capita net expenses below $15 and a combined population accounting for 1.1% of Australia's urban population.



Excluding Albury-Wodonga, Coffs Harbour, Goulburn, St Georges Basin – Sanctuary Point and Warrnambool

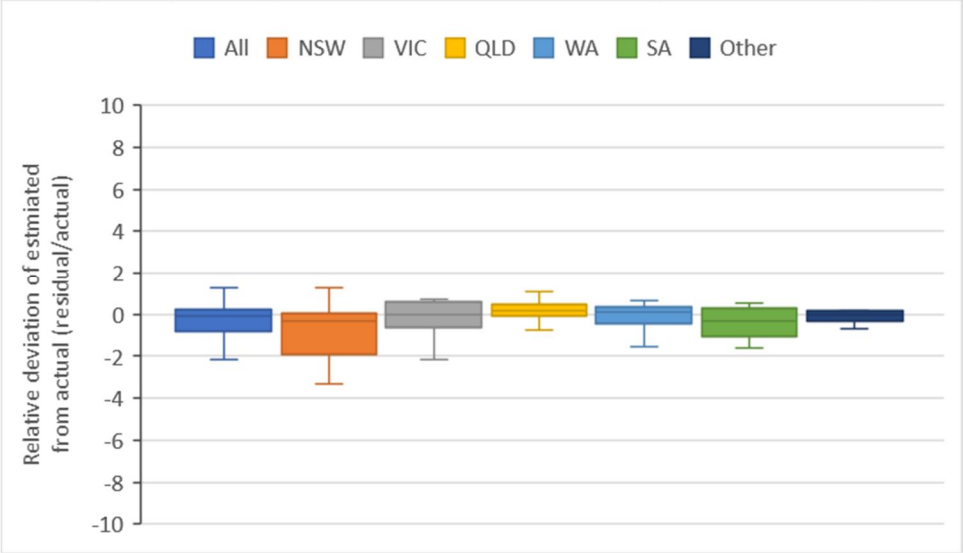


Figure E.4: Plot of relative residuals overall and by State

Data source: Synergies modelling

Coefficient estimates also follow intuition as the model suggest that net expenses per person

- increase with urban density (representing demand);
- increase with the distance to work (representing network complexity);
- increase with mean land slope (depicting topographical complexity); and
- increase with train and bus passengers.

Model 1b incorporates passenger mode numbers in a linear-log functional form (the name arises because the independent variables have been transformed by a logarithm, while the dependent variable has not). The linear-log relationship implies that per capita expenses increase as the network becomes more complex but the rate at which this occurs decreases as passenger volumes increase. This holds for buses and rail. Specifically, the linear-log relationship implies that for every 1% increase in passenger mode numbers, per capita expenses increase by a dollar amount equal to the respective estimated coefficient divided by 100.

Consider, under Model 1b, the effect of increasing bus passenger volume by 10% in Darwin versus increasing bus passenger volume by 10% in Kalgoorlie. Holding all other factors constant, this bus passenger increase will increase the per capita expenses in both cities by $0.67. However, because the passenger base in Kalgoorlie (594 passengers on Census night) is only about one tenth of that in Darwin (5100 passengers on Census night), the expense increase per additional 100 passengers in Kalgoorlie is $0.11 and that in Darwin only $0.01. The same holds for rail where per capita expenses increase by $1.81 with every 10% increase in passengers.

This means the linear-log form of the model can be interpreted as indicative of scale effects in the wider sense as it suggests that growth from additional passengers becomes less substantial as total volume increases.

Including variables that together account for all key drivers of expenses identified in the analytical framework Model 1b has a robust theoretical basis. The statistical and graphical tests indicate a strongly significant and unbiased functional relationship between dependent and explanatory variables that is stronger than that of the reference model. This makes Model 1b a candidate for the preferred model.

E.3 Model 2

Model 2 uses density ($dense_i$) to depict demand, a combination of distance to work ($dist_i$) and employment (emp_i) to represent demand (availability) and network complexity, and mean land slope ($slope_i$) to account for topography. We tested a linear form and one with the employment in natural logarithms. The correlation of population density and employment could lead to multi-collinearity issues in this model. We therefore test a version without population density as Model 3. Formally the models can be specified as:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 emp_i + \varepsilon_i \quad (2a)$$

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 \ln(emp_i) + \varepsilon_i \quad (2b)$$

Table E.5: Model 2: Test statistics

	Linear model (2a)	**Linear – log model (2b)**
Observations	70	70
F statistic	40.83	47.12
Prob > F	0.0000	0.0000
R^2	0.72	0.74
Adjusted R^2	0.70	0.73
Akaike information criterion	817	810
Bayesian information criterion	829	821
Root MSE	80	76

Source: Synergies modelling

Table E.6: Model 2: Coefficient estimates

	Coefficient estimate	**Standard error**	**95% confidence interval**	
Linear model (2a)				
Intercept	-239.3264***	49.71914	-338.6224	-140.0304
$dense_i$	0.118144***	0.0231505	0.0719094	0.1643787
$dist_i$	6.380205***	1.758083	2.869072	9.891338
$slope_i$	9.007778	5.609009	-2.194187	20.20974
emp_i	0.0000739	0.0000491	-0.0000241	0.0001719
Linear – log model (2b)				
Intercept	-511.4805***	85.97888	-683.1922	-339.7688
$dense_i$	0.0965532***	0.0201411	0.0563287	0.1367776
$dist_i$	5.718093***	1.682321	2.358266	9.077921
$slope_i$	7.433261	5.342233	-3.235916	18.10244
$\ln(emp_i)$	33.01957**	10.61787	11.81423	54.22492

*** p > |t| <=0.1%
** p > |t| <=1%
* p > |t| <=5%
^ p > |t| <=10%

Source: Synergies modelling

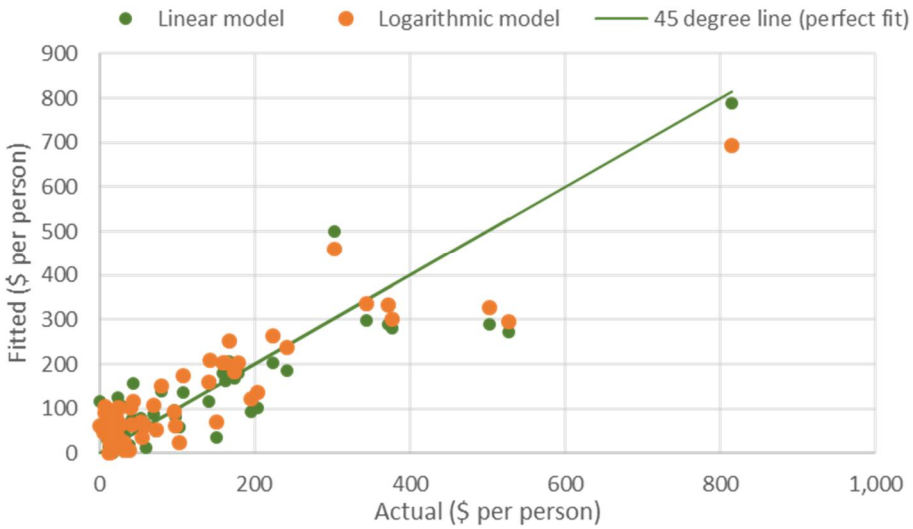


Figure E.5: Actual vs. fitted values for the two models

Data source: Synergies modelling

The actual vs. fitted plot shows a reasonable fit for both the linear and logarithmic form of the model. Since Model 2b has a higher R^2 value it is preferred to Model 2a. As Figure E.6 shows its estimates are unbiased overall and for all States with more than three observations.

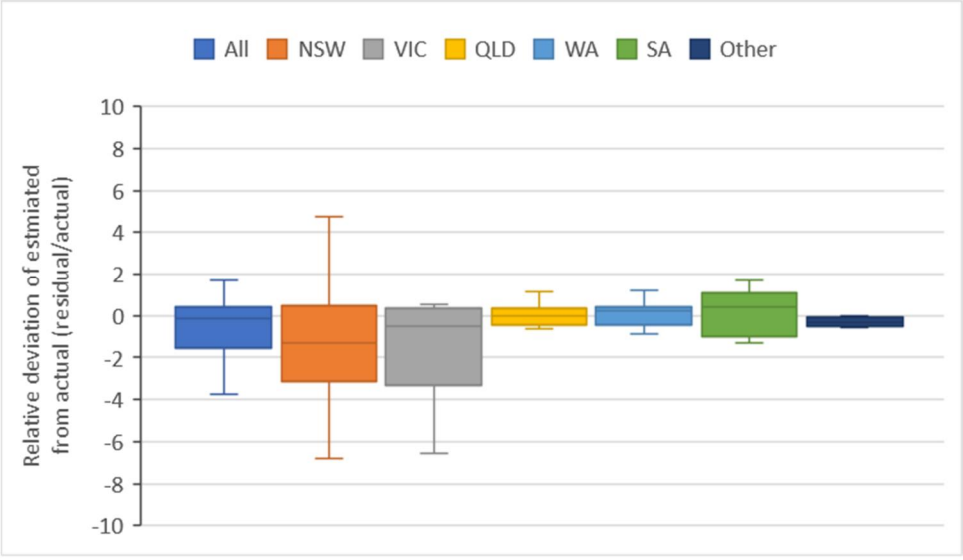


Figure E.6: Plot of relative residuals overall and by State

Data source: Synergies modelling

Coefficient estimates also follow intuition as the model suggest that net expenses per person

- increase with urban density (representing demand);
- increase with the distance to work (representing network complexity);
- increase with mean land slope (depicting topographical complexity);

- increase with the number of persons employed in the SUA.

Including variables that together account for the key aspects of the main drivers of expenses identified in the analytical framework, Model 2b has a robust theoretical basis. The statistical and graphical tests indicate a strongly significant and unbiased functional relationship between dependent and explanatory variables that is stronger than that of the reference model. Subject to the potential multi-collinearity issue mentioned above, this makes Model 2b a candidate for the preferred model.

E.4 Model 3

Model 3 is similar to Model 2 but excludes population density to avoid potential multi-collinearity issues resulting from this variables correlation with employment. Formally the models can be specified as:

$$exp_i = \beta_0 + \beta_1 dist_i + \beta_2 slope_i + \beta_3 emp_i + \varepsilon_i \quad (3a)$$

$$exp_i = \beta_0 + \beta_1 dist_i + \beta_2 slope_i + \beta_3 \ln(emp_i) + \varepsilon_i \quad (3b)$$

Table E.7: Model 3: Test statistics

	Linear model (3a)	**Linear – log model (3b)**
Observations	70	70
F statistic	33.18	41.38
Prob > F	0.0000	0.0000
R^2	0.60	0.65
Adjusted R^2	0.58	0.64
Akaike information criterion	839	829
Bayesian information criterion	848	838
Root MSE	94	88

Source: Synergies modelling

Table E.8: Model 3: Coefficient estimates

	Coefficient estimate	**Standard error**	**95% confidence interval**	
Linear model (3a)				
Intercept	-78.56964^	45.17957	-168.7736	11.63428
$dist_i$	6.154163**	2.064214	2.032828	10.2755
$slope_i$	10.27105	6.581366	-2.869073	23.41117
emp_i	0.0002865***	0.0000304	0.0002258	0.0003473
Linear – log model (3b)				
Intercept	-755.1824***	80.05823	-915.0239	-595.341
$dist_i$	4.668835*	1.925858	0.8237356	8.513935
$slope_i$	6.434232	6.163318	-5.87123	18.73969
$\ln(emp_i)$	74.62882***	7.061022	60.53104	88.72661

*** p > |t| <=0.1%
** p > |t| <=1%
* p > |t| <=5%
^ p > |t| <=10%

Source: Synergies modelling

The specifications above indicate that Model 3 is very similar to Model 2 which means that the suspected multi-collinearity in Model 2 is not a major issue. Model 3 is inferior to Model 2 in all other aspects, and as such, it does not yield further consideration.

E.5 Model 4

Model 4 combines Model 1 and Model 2 by including a dummy variable for the presence of public transport modes other than buses ($D_{i,mode}$). We tested a linear form and one with employment in natural logarithms. Formally the models can be specified as:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 emp_i + \beta_5 D_{i,train} + \beta_6 D_{i,tram} + \beta_7 D_{i,ferry} + \varepsilon_i \quad (4a)$$

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 \ln(emp_i) + \beta_5 D_{i,train} + \beta_6 D_{i,tram} + \beta_7 D_{i,ferry} + \varepsilon_i \quad (4b)$$

In contrast to Model 1, where public transport mode use is captured by passenger numbers extracted from the Census 2016 travel to work data by place of usual residence, Model 4 captures mode use through dummy variables, which simply indicate the presence or absence of a public transport mode. Each mode dummy takes a value of 1 if passengers are reported for a certain mode in an SUA, or zero if not.

The number of public transport passengers is very small in many SUAs. The confidentiality processes in the ABS Census data could be one reason for this. To ensure confidentiality in table cells with very small counts, the ABS introduces random error into the data by applying slight adjustments to cells.[35] This can result in small positive numbers being assigned to values that are in reality zero.

This data issue has very little impact when actual passenger numbers are used (as in Model 1) instead of dummies (as in Model 4). When using dummies, especially without a passenger threshold, there is no numerical distinction between an SUA with 100 passengers and an SUA with 100,000 passengers. In Model 1 however, if the number of passengers for a particular mode is very small, then this will be reflected through a minor impact on per capita net expenses. One possible solution in the usage of dummies is to apply a passenger threshold to the dummy variables, which means that the dummy takes a value of 1 only if the number of passengers is larger than a given value.

To demonstrate the challenges in Model 4 arising from the choice of passenger threshold, we have estimated the model with dummies that have a passenger threshold of 250, and again with dummies that do not have a passenger threshold. Table E9 and Table E10 display the sensitivity of the test statistics and coefficient estimates when the two different approaches are adopted. For ease of comparison, we do not present the 95% confidence intervals for the Model 4 coefficients.

Table E.9: Model 4: Test statistics

	Linear model (4a)		**Linear – log model (4b)**	
	250-passenger threshold	***No passenger threshold***	***250-passenger threshold***	***No passenger threshold***
Observations	70	70	70	70
F statistic	31.80	23.57	31.42	32.00
Prob > F	0.0000	0.0000	0.0000	0.0000
R^2	0.78	0.73	0.78	0.78
Adjusted R^2	0.76	0.70	0.76	0.76
Akaike information criterion	805	820	805	804
Bayesian information criterion	823	838	823	822
Root MSE	72	80	72	72

[35] For more details on ABS confidentiality processes, see: http://www.abs.gov.au/websitedbs/censushome.nsf/home/factsheetsccd?opendocument&navpos=450

Source: Synergies modelling

Table E.10: Model 4: Coefficient estimates

	Coefficient estimate	Standard error	
Linear model (4a)			
Intercept	-194.2908***	48.32675	
$dense_i$	0.1133834***	0.0229213	
$dist_i$	3.880568*	1.68846	
$slope_i$	8.71598^	5.074125	
emp_i	-0.0001156	0.0001158	
$D_{i,train}$	102.682***	27.9556	
$D_{i,tram}$	9.794041	74.67019	
$D_{i,ferry}$	228.6228^	126.1777	
Linear – log model (4b)			
Intercept	-253.8425*	121.5553	
$dense_i$	0.0943145***	0.022707	
$dist_i$	4.142592*	1.704461	
$slope_i$	8.076735	5.091902	
$\ln(emp_i)$	8.411398	13.10598	
$D_{i,train}$	89.42893	35.28419	
$D_{i,tram}$	-43.53836^	52.23771	
$D_{i,ferry}$	105.8958*	53.61403	

*** p > |t| <=0.1%
** p > |t| <=1%
* p > |t| <=5%
^ p > |t| <=10%

Source: Synergies modelling

Although the direction of the coefficients for density, distance to work and mean land slope remain consistent with intuition, several of the dummy coefficients do not align with the expectations formulated in the analytical framework. We therefore consider it inferior to Model 1b which meets all these criteria.

E.6 Model 5

Model 5 uses SEIFA as the variable depicting demand generation. It retains density ($dense_i$) as an additional demand driver, distance to work ($dist_i$) to represent network complexity and mean land slope ($slope_i$) for topography. We tested a linear form and one with land slope in natural logarithms. Formally the models can be specified as:

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 slope_i + \beta_4 SEIFA_i + \varepsilon_i \quad (5a)$$

$$exp_i = \beta_0 + \beta_1 dense_i + \beta_2 dist_i + \beta_3 \ln(slope_i) + \beta_4 SEIFA_i + \varepsilon_i \quad (5b)$$

Table E.11: Model 5: Test statistics

	Linear model (5a)	**Linear – log model (5b)**
Observations	70	70
F statistic	39.58	40.17
Prob > F	0.0000	0.0000
R^2	0.71	0.71
Adjusted R^2	0.69	0.69
Akaike information criterion	819	818
Bayesian information criterion	830	829
Root MSE	81	81

Source: Synergies modelling

Based on these model selection indicators, Model 5 is not superior to either Models 1, 2 or 4, and therefore does not yield further consideration.

Appendix F. Data availability and quality by SUA

The table below shows the SUA's share of Australia's urban population (the part living in the 101 SUAs) to provide an indication of the weight of the excluded SUAs.

Table F.1: Data availability and quality by SUA

State	SUA	Included in analysis	Reason for exclusion	Share of total urban population
ACT	Canberra - Queanbeyan (ACT)	Yes		2.16%
NSW	Albury - Wodonga (NSW)	Yes		0.43%
NSW	Armidale	Yes		0.11%
NSW	Ballina	Yes		0.12%
NSW	Batemans Bay	Yes		0.08%
NSW	Bathurst	Yes		0.17%
NSW	Bowral - Mittagong	Yes		0.18%
NSW	Broken Hill	Yes		0.09%
NSW	Camden Haven	No	Expense data not provided	0.07%
NSW	Central Coast	Yes		1.59%
NSW	Coffs Harbour	Yes		0.33%
NSW	Dubbo	Yes		0.18%
NSW	Forster - Tuncurry	Yes		0.10%
NSW	Goulburn	Yes		0.11%
NSW	Grafton	Yes		0.09%
NSW	Griffith	Yes		0.10%
NSW	Kempsey	Yes		0.06%
NSW	Lismore	Yes		0.14%
NSW	Lithgow	Yes		0.06%
NSW	Morisset - Cooranbong	No	Expense data not provided	0.11%
NSW	Mudgee	Yes		0.06%
NSW	Muswellbrook	Yes		0.06%
NSW	Nelson Bay	No	Expense data not provided	0.13%
NSW	Newcastle - Maitland	Yes		2.33%
NSW	Nowra - Bomaderry	Yes		0.17%
NSW	Orange	Yes		0.19%
NSW	Parkes	Yes		0.05%
NSW	Port Macquarie	Yes		0.23%
NSW	Singleton	No	Expense data not provided	0.07%
NSW	St Georges Basin - Sanctuary Point	Yes		0.09%
NSW	Sydney	Yes		22.23%
NSW	Tamworth	Yes		0.19%
NSW	Taree	Yes		0.12%
NSW	Ulladulla	Yes		0.08%
NSW	Wagga Wagga	Yes		0.27%

State	SUA	Included in analysis	Reason for exclusion	Share of total urban population
NSW	Wollongong	Yes		1.45%
NT	Alice Springs	Yes		0.14%
NT	Darwin	Yes		0.62%
Qld	Brisbane	Yes		10.83%
Qld	Bundaberg	No	Derived using population	0.34%
Qld	Cairns	Yes		0.74%
Qld	Emerald	No	Derived using population	0.07%
Qld	Gladstone - Tannum Sands	No	Derived using population	0.22%
Qld	Gold Coast - Tweed Heads (QLD)	Yes		3.06%
Qld	Gympie	No	Derived using population	0.09%
Qld	Hervey Bay	No	Derived using population	0.26%
Qld	Kingaroy	No	Derived using population	0.05%
Qld	Mackay	Yes		0.41%
Qld	Maryborough	No	Derived using population	0.12%
Qld	Mount Isa	No	Derived using population	0.10%
Qld	Rockhampton	No	Derived using population	0.38%
Qld	Sunshine Coast	Yes		1.39%
Qld	Toowoomba	Yes		0.63%
Qld	Townsville	Yes		0.88%
Qld	Warwick	Yes		0.07%
Qld	Yeppoon	No	Derived using population	0.09%
SA	Adelaide	Yes		6.34%
SA	Mount Gambier	Yes		0.13%
SA	Murray Bridge	Yes		0.09%
SA	Port Augusta	Yes		0.07%
SA	Port Lincoln	Yes		0.08%
SA	Port Pirie	Yes		0.07%
SA	Victor Harbor - Goolwa	Yes		0.12%
SA	Whyalla	Yes		0.11%
Tas	Burnie - Wynyard	Yes		0.13%
Tas	Devonport	No	Derived using population	0.14%
Tas	Hobart	Yes		0.95%
Tas	Launceston	Yes		0.42%
Tas	Ulverstone	No	Derived using population	0.07%
Vic	Bacchus Marsh	No	Derived using population	0.09%
Vic	Bairnsdale	No	Derived using population	0.06%
Vic	Ballarat	Yes		0.47%
Vic	Bendigo	Yes		0.46%
Vic	Colac	No	Derived using population	0.06%
Vic	Echuca - Moama (VIC)	No	Derived using population	0.09%
Vic	Geelong	Yes		1.16%

State	SUA	Included in analysis	Reason for exclusion	Share of total urban population
Vic	Gisborne - Macedon	No	Derived using population	0.08%
Vic	Horsham	No	Derived using population	0.08%
Vic	Melbourne	Yes		21.68%
Vic	Melton	No	Derived using population	0.28%
Vic	Mildura - Wentworth (VIC)	Yes		0.20%
Vic	Moe - Newborough	No	Derived using population	0.08%
Vic	Portland	No	Derived using population	0.05%
Vic	Sale	No	Derived using population	0.07%
Vic	Shepparton - Mooroopna	Yes		0.24%
Vic	Swan Hill	No	Derived using population	0.05%
Vic	Traralgon - Morwell	No	Derived using population	0.20%
Vic	Wangaratta	No	Derived using population	0.09%
Vic	Warragul - Drouin	No	Derived using population	0.14%
Vic	Warrnambool	Yes		0.16%
WA	Albany	Yes		0.16%
WA	Broome	Yes		0.07%
WA	Bunbury	Yes		0.36%
WA	Busselton	Yes		0.15%
WA	Esperance	Yes		0.05%
WA	Geraldton	Yes		0.18%
WA	Kalgoorlie - Boulder	Yes		0.16%
WA	Karratha	Yes		0.09%
WA	Perth	Yes		9.69%
WA	Port Hedland	Yes		0.07%
WA	Yanchep	No	Part of metropolitan Perth	0.04%
Total included		**70**		**96.21%**
Total excluded		**31**		**3.79%**

Source: Synergies analysis of State data collected by the CGC

Appendix G. Self-sufficiency index values by SUA

Table G.1: Self-sufficiency index values by SUA

State	SUA	Capital city index	SUA index	SA2 index
ACT	Canberra - Queanbeyan	0.96	0.05	0.97
NSW	Albury - Wodonga	0.01	0.20	1.00
NSW	Armidale	0.00	0.10	0.10
NSW	Ballina	0.00	0.37	0.48
NSW	Batemans Bay	0.01	0.25	0.54
NSW	Bathurst	0.01	0.17	0.43
NSW	Bowral - Mittagong	0.16	0.30	0.60
NSW	Broken Hill	0.00	0.04	0.04
NSW	Camden Haven	0.01	0.57	0.57
NSW	Central Coast	0.19	0.33	0.81
NSW	Coffs Harbour	0.00	0.11	0.63
NSW	Dubbo	0.00	0.10	0.58
NSW	Echuca - Moama	0.02	0.64	0.64
NSW	Forster - Tuncurry	0.01	0.22	0.47
NSW	Gold Coast - Tweed Heads	0.11	0.21	0.89
NSW	Goulburn	0.01	0.17	0.17
NSW	Grafton	0.00	0.14	0.14
NSW	Griffith	0.00	0.27	0.27
NSW	Kempsey	0.00	0.18	0.18
NSW	Lismore	0.00	0.26	0.55
NSW	Lithgow	0.07	0.33	0.33
NSW	Mildura - Wentworth	0.00	0.14	0.60
NSW	Morisset - Cooranbong	0.08	0.62	0.72
NSW	Mudgee	0.00	0.27	0.27
NSW	Muswellbrook	0.00	0.28	0.28
NSW	Nelson Bay	0.02	0.37	0.48
NSW	Newcastle - Maitland	0.01	0.15	0.80
NSW	Nowra - Bomaderry	0.01	0.20	0.49
NSW	Orange	0.00	0.17	0.54
NSW	Parkes	0.00	0.23	0.23
NSW	Port Macquarie	0.01	0.17	0.52
NSW	Singleton	0.01	0.46	0.46
NSW	St Georges Basin - Sanctuary Point	0.03	0.58	0.71
NSW	Sydney	0.96	0.04	0.86
NSW	Tamworth	0.00	0.10	0.60
NSW	Taree	0.00	0.15	0.30
NSW	Ulladulla	0.02	0.23	0.23
NSW	Wagga Wagga	0.00	0.11	0.69

State	SUA	Capital city index	SUA index	SA2 index
NSW	Wollongong	0.15	0.24	0.83
NT	Alice Springs	0.00	0.11	0.77
NT	Darwin	0.88	0.12	0.88
Qld	Brisbane	0.92	0.08	0.86
Qld	Bundaberg	0.00	0.13	0.79
Qld	Cairns	0.00	0.10	0.81
Qld	Emerald	0.00	0.28	0.28
Qld	Gladstone - Tannum Sands	0.00	0.15	0.75
Qld	Gold Coast - Tweed Heads	0.11	0.21	0.89
Qld	Gympie	0.01	0.22	0.46
Qld	Hervey Bay	0.01	0.22	0.69
Qld	Kingaroy	0.00	0.23	0.23
Qld	Mackay	0.01	0.18	0.84
Qld	Maryborough	0.01	0.22	0.46
Qld	Mount Isa	0.00	0.10	0.10
Qld	Rockhampton	0.01	0.16	0.82
Qld	Sunshine Coast	0.06	0.17	0.74
Qld	Toowoomba	0.01	0.14	0.80
Qld	Townsville	0.01	0.08	0.84
Qld	Warwick	0.01	0.14	0.14
Qld	Yeppoon	0.01	0.50	0.50
SA	Adelaide	0.94	0.06	0.86
SA	Mount Gambier	0.01	0.15	0.53
SA	Murray Bridge	0.09	0.28	0.28
SA	Port Augusta	0.01	0.12	0.12
SA	Port Lincoln	0.01	0.12	0.12
SA	Port Pirie	0.01	0.10	0.10
SA	Victor Harbor - Goolwa	0.14	0.27	0.44
SA	Whyalla	0.01	0.20	0.20
Tas	Burnie - Wynyard	0.00	0.18	0.72
Tas	Devonport	0.01	0.24	0.63
Tas	Hobart	0.94	0.06	0.83
Tas	Launceston	0.01	0.15	0.85
Tas	Ulverstone	0.00	0.53	0.72
Vic	Albury - Wodonga	0.01	0.20	1.00
Vic	Bacchus Marsh	0.44	0.67	0.67
Vic	Bairnsdale	0.01	0.17	0.17
Vic	Ballarat	0.04	0.16	0.76
Vic	Bendigo	0.02	0.16	0.77
Vic	Colac	0.01	0.14	0.14
Vic	Echuca - Moama	0.02	0.64	0.64

State	SUA	Capital city index	SUA index	SA2 index
Vic	Geelong	0.13	0.21	0.78
Vic	Gisborne - Macedon	0.54	0.70	0.76
Vic	Horsham	0.00	0.14	0.14
Vic	Melbourne	0.95	0.05	0.86
Vic	Melton	0.62	0.73	0.86
Vic	Mildura - Wentworth	0.00	0.14	0.60
Vic	Moe - Newborough	0.04	0.58	0.58
Vic	Portland	0.00	0.12	0.12
Vic	Sale	0.01	0.35	0.35
Vic	Shepparton - Mooroopna	0.01	0.18	0.57
Vic	Swan Hill	0.00	0.11	0.11
Vic	Traralgon - Morwell	0.02	0.25	0.46
Vic	Wangaratta	0.01	0.20	0.20
Vic	Warragul - Drouin	0.20	0.42	0.56
Vic	Warrnambool	0.01	0.14	0.57
WA	Albany	0.01	0.17	0.60
WA	Broome	0.00	0.08	0.08
WA	Bunbury	0.02	0.24	0.79
WA	Busselton	0.02	0.25	0.35
WA	Esperance	0.01	0.13	0.13
WA	Geraldton	0.01	0.12	0.66
WA	Kalgoorlie - Boulder	0.01	0.20	0.61
WA	Karratha	0.01	0.35	0.35
WA	Perth	0.91	0.09	0.86
WA	Port Hedland	0.02	0.10	0.49
WA	Yanchep	0.57	0.80	0.84

Note: Index values were calculated as the SA2 level. The values presented in the table are population weighted averages.

Source: Synergies analysis of Census 2016 data