The Development of a Sydney VKT Regression Model

Grace Corpuz, Michelle McCabe, Kamila Ryszawa
Transport and Population Data Centre (TPDC)
New South Wales Department of Planning, Australia

1 Introduction

The New South Wales Government's Metropolitan Strategy document *City of Cities, A Plan for Sydney's Future* (2005) provides for the integration of land use and transport planning to achieve one of its key objectives, to "influence travel choices to encourage more sustainable travel". Integral to this aim is the promotion of urban design and land use solutions that encourage walking, cycling and use of public transport in order to reduce car dependency.

As provided for in the Strategy, support will be provided to local councils in their preparation of Local Environmental Plans (LEPs) to ensure that planned development aligns with the Strategy's sustainability and growth targets. This support will take the form of an online planning system that will be used to monitor and assess council plans (NSW Dept of Planning 2005). The scheme that is being developed to enable this quantitative assessment is called METRIX.

To establish a framework for METRIX, the Sustainability Unit of the New South Wales Department of Planning commissioned the Transport and Population Data Centre (TPDC) to develop a Sydney VKT Regression Model. The aim was firstly to identify the factors that impact on car usage, and secondly, to develop a quantitative model to predict the vehicle kilometres travelled (VKT) likely to be generated given a set of socio-economic, locational and urban form characteristics. The Model will be used to inform land use planning by predicting VKT resulting from proposed developments enabling their assessment against the Metro Strategy objective of reducing growth in VKT. In addition, the Model can be applied to gauge the impacts of various development scenarios at a broader sub-regional level.[1]

The main source of the dataset used in the development of the model is TPDC's Household Travel Survey (HTS). The model used seven years of survey data collected from June 1997 to June 2004 on the travel and socio-demographic characteristics of over 16,000 households in the Sydney Statistical Division. (For details about the HTS and the data used in the model development including scope and coverage, please see Appendix 1.) The dataset also included information on dwelling and employment densities derived using statistics from the Australian Bureau of Statistics' 2001 Census and the Local Environment Plan (LEP) spatial data on land use zones. Data on accessibility to public transport used the latest available information on Sydney's public transport network and services.

This paper describes the development of the Sydney VKT Regression Model. It begins with a brief review of literature for results and procedures of related studies. The following section presents the model and explains what it means. Section 4 describes the development process. Subsequent sections discuss the limitations, the validation procedure and scope for further work.

2 Related studies and findings

There is significant discussion in the literature about the relationship between the urban environment and travel behaviour (Boarnet and Crane 2001, Brunton and Brindle 1999, Burke and Brown 2005, Soltani and Primerano 2005, Soltani and Somenahalli 2005).

[1] For further details about the policy application of the model, please refer to a related paper to be presented at this conference (29th ATRF) by Holden, *"I've got a car": the relationship between land use and car dependence and its application for land use planning policy in Sydney*.

However, the conclusions are mixed, even those that use empirical evidence. Boarnet and Crane (2001) state that many studies indicate that higher densities, mixed land-use, and compact neighbourhoods are correlated with low car usage but other studies show the impacts to be either nil or insignificant. In addition, the results about the extent of the effect differ by location suggesting that these are not necessarily transferable from one city to another (Chandra 2005). Thus, the modellers have adopted lessons from the literature in their choice of the influencing variables but remained open to the unique conclusions that may arise from the modelling of Sydney data.

The predictor variables, distance to the CBD, access to public transport, employment density, housing density, land use mix, dwelling type, level of services available locally; number of vehicles, persons, licence-holders and persons of driving age in the household; and income, were tested in the model based on discussion and evidence widely found in literature. Some neighbourhood design characteristics such as road configuration, presence of cycling and recreational paths were identified in the literature as influencing travel patterns and mode usage (Boarnet and Crane 2001, Brunton and Brindle 1999, Burke and Brown 2005, IBI Group 2000, Newman 2005, Soltani and Primerano 2005, Soltani and Somenahalli 2005) but were not included because of the absence of readily available data.

The methodology used in the project was based on a Canadian model which had similar aims and approach (IBI Group 2000). The developers found this procedure particularly attractive because of the relative ease that it can be applied considering the nature of the Sydney data and the short timeframe of the project. The authors noted the criticism by Burke and Brown (2005) that regression-based approaches 'ignore most of the contributions of location'. In the Sydney VKT Regression Model, as in the Canadian model, however, locational characteristics such as 'distance to the Central Business District (CBD)' and accessibility to public transport featured prominently.

3 The Model

3.1 Components of the model

3.1.1 Dependent variable

The number of vehicle kilometres travelled (VKT) is obtained from the HTS as the total road distance travelled by each household for all trips where the mode is *vehicle driver*.

When the dependent variable VKT was fitted, one of the assumptions of linear regression (the requirement for the errors to have constant variation) was violated. A simple square root transformation fixed this problem. Therefore, the recommended model has the square root of VKT as the dependent variable. To predict VKT, we merely square the output from the model.

3.1.2 Predictor variables

The initial set of predictor variables that were tested were jointly chosen by TPDC and the Sustainability Unit. These variables measure three main characteristics: location, socio-demographics and urban form / neighbourhood design.

For the complete list of the variables in the model including definitions and data sources, please refer to Appendix 2.

3.2 The recommended model

The recommended model is represented by the following linear equation relating the square root of VKT to a number of significant predictor variables[2].

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

(square root of the household VKT) = 3.9270 + (2.4510 * number of vehicles if the household) + (0.0124 * closest distance to major centre or CBD) + (-1.8057 * land use mix) + (-0.0021 * local employment) + (-0.0099 * housing density) + (0.0084 * distance to nearest train, ferry, light rail or high frequency bus)

The model has the following characteristics:

- It has a good Rsquare of 0.731, meaning that the model explains nearly three quarters of the variation of VKT.
- The predictor variables were chosen to minimise the effect of multicollinearity[2].
- The model satisfied the assumptions of linear regression analysis[3]

3.3 What does the recommended model mean?

The coefficients ($b_1 \ldots b_n$) estimate the magnitude of the impact on the *square root* of household VKT for every unit change in the corresponding predictor variable. In addition, the signs of the coefficients denote the direction of the impact, that is, whether the square root of VKT is increased or decreased by every unit change in the predictor variable. For example, an increment of one vehicle in the household will increase the square root of VKT by 2.454.

4 Development of the model

The model was developed using the regression analysis facility of the statistical software SPSS version 13. The following were the main stages in the development:

- Preliminary data analysis and preparation
- First stage regression analysis – individual household level
- Second stage regression analysis – data aggregated at the collection district (CD) level and travel zone (TZ) level
- Third stage regression analysis – improvement of the TZ model
- Fourth stage regression analysis – further trials using weighted least squares

The process is described in the following sub-sections. [4]

[2] Collinearity (or multicollinearity) refers to the situation where the predictor variables are correlated with each other and explains similar variability in the dependent variable, square root of household VKT. It is a serious situation which impacts on the reliability of the coefficients and also affects the statistical tests of significance (Neter et al 1996, pp 285-291).

The variables included in the model were chosen to minimise the effect of collinearity. This means that there were in fact other predictor variables which on their own are significantly related to the dependent variable but which have not been included in the model. These excluded variables are highly correlated with and therefore effectively 'captured' by one or more variables that are in the model. (This is discussed further in Section 4.5.)

[3] The final model satisfied the following assumptions for validity (Neter et al 1996):

- The errors (e_i) have a normal distribution with a mean of zero
- The errors (e_i) have constant variation across cases and independent of the variables in the model. Non-constant variance is called to be *heteroscedasticity*
- The errors (e_i) are independent of each other.

[4] The discussion may be more summarised than what some readers may prefer, particularly for the stages leading to the recommended model. This is intentional to assist in the understanding of what proved to be a highly involved process. For further details about each stage, please contact the authors.

4.1 Preliminary data analysis and preparation

This preliminary step involved the examination of descriptive statistics, scatterplots, and histograms. These diagnostics were used to identify, examine and remove outliers and suspect data. The correlations between household VKT and the explanatory variables were also analysed for early indications of which characteristics can be expected to be significant in the model.

4.2 First stage regression analysis – individual household level

In this early stage of the regression analysis, the *stepwise*[5] method was applied. All the variables were entered and the resulting model had an Rsquare of 0.34 and the most influential variables were: the number of household vehicles, the distance to CBD, number of licence holders, household income, and access to non-road transport.

With only a third of the variation explained by the variables in the model, the researchers turned their attention to the distribution of VKT. Since it showed high variability, the modellers trialled various modifications of VKT, including restricting VKT to those within (a) three standard deviations (b) excluding households that made no car driver trips and (c) taking the natural log of VKT. The best Rsquare value in this group of iterations was 0.339 with similar variables as before appearing as significant in the model.

Further runs using the following modified datasets were undertaken:

- Small subset of the data with travel dates between 1 May and 30 June 2004[6]
- Dataset split into weekday and weekend day travel by a variety of methods
- Household VKT transformed into 3, 4, 5 and 8 categories
- Dataset divided into 3 regions based on (a) distance to CBD and (b) groups of Statistical Local Areas.

Various combinations of these modifications were tested. For this group of trials, the best Rsquare was 0.439 for the model where the household VKT was grouped into eight categories. The variables identified as significant in the model were consistent with those previously selected.

However, in addition to the relatively poor fit as shown by the low Rquare values above, the second assumption of constant variance was also not met in all these trials. Statistical literature suggested a square root transformation of VKT to obtain constant variance (Neter et al 1996, p130). When this was undertaken, the computed model had an Rsquare of 0.46 with the following variables being the most influential: the number of household vehicles, the distance to CBD, number of licence holders, household income, number of residents, and access to non-road transport.

4.3 Second stage regression analysis – data aggregated at the collector's district (CD) level and travel zone (TZ) level

In response to the low Rsquare values in the preceding stage, the researchers decided to aggregate each variable at two geographical levels: Census collection district (CD) and travel zone (TZ). *Aggregation,* as the term is used here, refers to *averaging* each variable at the CD and TZ level. For ordinal or categorical variables, such as dwelling structure and household type, the median[7] value was used as the average.

[5] At each step, the independent variable not already in the equation which has the smallest probability of F is entered, if that probability is sufficiently small. Variables already in the regression equation are removed if their probability of F becomes sufficiently large. The method terminates when no more variables are eligible for inclusion or removal.

[6] This was tested to reduce the variability due to the HTS data being collected in a continuous manner over a number of years. In comparison, the Canadian Model was based on data from a one-off survey. The Rsquare in this instance was also low at 0.378.

[7] The researchers noted the disadvantage of using the median and considered the mode as a better measure of central tendency for the ordinal variables. However, to minimise complexity at this exploratory stage, the median which was significantly easier to derive in SPSS was used. If the median value was an average of two types, the next value up was chosen.

Averaging contains the variability in the data by reducing the *noise* which characterises travel behaviour at an individual household level. VKT can vary widely due to factors that have not been or can not be measured. For example, a change in weather or personal preference at a particular time might mean someone is picked up by car from the station rather than walk home. Or, someone can make a detour to a shopping centre on the way home from work in response to a need or opportunity.[8]

Another argument for aggregating the data in this situation is that a number of variables are already based at CD level (access to non-road transport, access to bus services, level of local services, land use mix, housing density, employment, and distance to CBD or major regional centre)[9]. Aggregation also made practical sense even though we lose some of the geographical detail, because the distance measure used in the household VKT variable is calculated based on distances between the origin and destination travel zone centroids.[10]

As with any summary of data, diversity is necessarily lost in the quest to reduce variability. However, with care in interpretation, aggregation can identify underlying trends and relationships which remain hidden by noise at the individual household level. The modellers carried out iterations of the regression model at both the CD and TZ levels.

It should also be noted that the Canadian Model used data averaged at a level similar to the travel zone, that is, at *traffic* zone. It stated that this level of aggregation was chosen as the basis of its analysis since "this provides a convenient means for summarising travel data, and this is also compatible with the need to make comparisons at the neighbourhood level" (IBI Group 2000).

4.3.1 Data aggregated by CD

At this level of aggregation the number of observations was reduced from over 16,000 households to 3,332 CDs.

All variables were entered into the model and the resulting Rsquare was 0.508, an improvement on that of the best models at an individual household level. As the model still exhibited non-constant variance in the errors the CD-aggregated household VKT variable was transformed by taking the square root. The resulting model had an improved Rsquare of 0.579. Since the Rsquare though improved seemed low especially when compared with the Canadian Model which achieved an Rquare of 0.84 (IBI Group 2000), the researchers decided to proceed to aggregation by TZ.

4.3.2 Data aggregated by TZ

At the TZ level of aggregation the number of observations was reduced to 872 TZs, which is in line with the Canadian model which was based on 795 traffic zones (IBI Group 2000).

A preliminary examination of the plots using data aggregated at TZ level showed more apparent linear relationships than at the individual household level or the CD aggregation level. When all the variables were included and processed through the stepwise method, the regression results showed an Rsquare of 0.698.

---

[8] Horowitz (1995) mentioned this difficulty in his work on modelling residential location and mode of travel to work. He explained that the complexity of modelling travel behaviour comes from the large number of variables that influence choices and the fact that some of them are not measurable directly, eg mode choice can depend on fare and travel time considerations but also on highly personal factors such as whether other activities are planned prior or after work or whether the car will be required by another household member. He advised that a model should not attempt to capture all the complexities and contributing factors that impact on choice, if ever this was possible, as this can increase development costs unreasonably and impede its usability. He recommends to aim for a balance between practicality and complexity and develop a model that can be of practical use and also able to predict realistically.

[9] These variables were computed based on the centroid of the CD where the household is located.

[10] Estimates of trip distance or kilometres travelled in the HTS are based on the shortest *road* distance between the centroids of the origin and destination travel zones of the trip. For intra-zonal trips (i.e. trips that start and end within the same travel zone) distance travelled is estimated according to the size of the zone. This methodology has limited accuracy in the estimation of short trips. TPDC is currently enhancing this process using XY coordinates to improve the quality of distance estimates.

As the model above still exhibited non-constant variance in the errors, the modellers transformed the aggregated VKT by taking the square root once again before re-applying the stepwise regression method. The Rsquare was an improvement, 0.766, and the variables selected were: number of household vehicles, distance to CBD, number of licence holders, access to non-road transport, housing density, land use mix, whether three-storey flat.

This model required more work as many of the predictor variables were correlated with each other, which can be an issue for the reliability of the model's predictions, as previously explained. But since the model had the best fit and satisfied the conditions of regression analysis including constant variance of the errors, the researchers decided to settle on this model and focus on improving it.

4.5 Third stage regression analysis – improvement of the TZ model

This stage concentrated on two processes, refining the explanatory variables and dealing with the problem of multicollinearity.

Further work was undertaken to improve the quality of the housing density, employment, land use mix and 'distance to the nearest regional centre' data. The researchers also worked on refining the ordinal variables and improving on their treatment in the model.

Attention was then directed to the problem of multicollinearity. The researchers examined the correlations between the dependent variable and each of the explanatory variables, and between the explanatory variables themselves. Some of the non-significant variables were clearly (practically and statistically) "covered" by other variables. That is, two or more variables explain the same variation in VKT because they are highly correlated. For example,

- Level of local employment with dwelling structure, housing density, level of local services, number of household vehicles in the household
- Distance to major regional centre with distance to CBD and access to non-road transport
- Access to bus with distance to CBD, distance to major and access to non-road transport
- Number of household vehicles with residents aged 16 years and over, residents aged 18 years and over, number of licence holders, number of residents and household income

In light of the strong correlations, statistical literature cautioned that the stepwise method could potentially choose the "wrong" variables. As a result, the modellers manually chose variables that had a stronger influence on VKT and which were 'practically important', and applied the *entry*[11] method in SPSS to specify the variables to include in the model (in contrast to the stepwise method which automatically chooses the variables). The choices were based on a detailed investigation of the correlations, the collinearity statistics and the statistical significance of each variable in the model output.

The modellers removed the variable 'number of licence holders' because of its high correlation with the number of household vehicles.[12] All dummy variables created for dwelling structure and household type were also excluded as these impacted on the collinearity and all except one were insignificant.

Following discussions with the Sustainability Unit and other internal stakeholders, two new variables were also created in line with practical and policy considerations. The first combines two variables which measure access to bus (which was insignificant in the model) and access to non-road transport separately into a single variable, access to all public

[11] All variables in a block are entered in a single step.
[12] In practical terms, information about the number of licence holders is not easily available anyway, whereas we can estimate the number of vehicles per household using census data.

transport (*including* high frequency bus). The second combines 'distance to the CBD' and 'distance to the nearest major centre' into a single variable choosing the shorter of the two distances[13]. The latter was borderline significant at the 0.097 but was included because of its practical importance (Table 1).

Table 1 SPSS output for the final model

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	3.9270	0.2602		15.0910	0.0000		
number of household vehicles	2.4510	0.1086	0.5338	22.5644	0.0000	0.5565	1.7968
distance to CBD or major centre	0.0124	0.0075	0.0431	1.6605	0.0972	0.4633	2.1586
land use mix	-1.8057	0.3458	-0.1128	-5.2224	0.0000	0.6676	1.4978
employment	-0.0021	0.0007	-0.1077	-3.0896	0.0021	0.2565	3.8990
housing density	-0.0099	0.0027	-0.1273	-3.6492	0.0003	0.2560	3.9057
access to public transport	0.0084	0.0011	0.1719	7.3706	0.0000	0.5723	1.7473

Dependent Variable: Square root of household VKT

In the final model (Section 3.2), the Rsquare is 0.731 and the variables chosen are: number of vehicles in the household, access to train, ferry, light rail or high frequency bus, land use mix, housing density, employment density, distance to the CBD or nearest major centre.

4.6 Fourth stage regression analysis – further trials using weighted least squares

During consultation with users and stakeholders, it was suggested that using a square root transformation of household VKT in the model introduces some problems of interpretation. Users expressed preference for an untransformed household VKT as the dependent variable in the model if this were possible. Thus, the researchers decided to trial another treatment for the problem of non-constant variance that would not require a transformation, the weighted least squares procedure.[14]

The procedure was applied on both non-aggregated data at the household level and aggregated data at the TZ level. After a number of trials, the researchers decided not to adopt the weighted least squares approach and the following were the main reasons:

- Statistical literature recommends finding a suitable transformation as the preferred way of satisfying the assumption of non-constant variances.
- As the method transforms the dependent variable (by the weight, w_i) the issue of interpretation of the household VKT predictions remains
- As the errors are unknown and must be estimated to derive the weight w_i, the fit of the model depends on the success of the error estimation procedure. In the trials, the regression equation estimating the errors had a poor fit, about 0.20 at best.
- There were only 3 variables which were significant in the final model: number of vehicles, access to public transport (including bus), land use mix, meaning that a number of practically important variables were not included.
- The methodology and calculations were significantly more complex and resource-intensive.

[13] Both 'distance to the CBD' and 'distance to the nearest major centre' were significant in the model but the latter had an unexpected negative coefficient despite having a statistically positive correlation with the dependent variable. This result reinforced the decision to combine the said two variables.

[14] Statistical literature describes two methods for situations of non-constant variance (Neter et al 1996):
- firstly, find a reasonable transformation to stabilise the variances of the errors, which does not introduce problems of interpretation or upset the functional relationship of VKT with the independent variables, or
- secondly, if a suitable transformation cannot be found, investigate the possibility of weighted least squares. This is achieved by finding an appropriate function to weight the regression variables. If a convincing function is found (one that has substantial R^2 and/or one that makes substantive sense, such as when the error variance is proportional to some measure of the size of the unit) then weight each observation by the inverse of the error of that observation and re-apply regression.

In addition, the procedure opened up a new set of issues that required substantial time and resources to investigate and resolve. At this stage of development, the time constraints did not allow further work on this approach to proceed.

5 Limitations

The regression model makes predictions of VKT for given values of the explanatory variables in the model. These predictions are subject to errors because of the imperfect fit of the model which can predict three quarters but not all of the variability of (the square root of) household VKT. Users should be mindful of this limitation when using the predictions on their own for say, inputs for another process, e.g. estimating resultant emissions. But for the purpose of comparing predictions, such as between locations which is main purpose of the model, the errors of the predictions become less of an issue. The comparisons remain valid because of the use of a single model as the same basis, and especially since the errors have constant variance and do not vary systematically between predictions.

As previously indicated, the measure of household VKT is based on road distances between travel zone origin and destination centroids. This methodology produces estimates but not precise measures of VKT. Nevertheless, since the measurement is applied consistently, the impact of this imprecision on the model development and the predictions is minimised. It should be noted that there was no alternative measure of distances in the HTS at the time of model development. XY coordinates of trip origins and destinations which can produce more accurate distance measurements were not yet available at the time.

Two other characteristics of the final model must be considered: the use of data aggregated at the travel zone level and the use of the square root transformation of the household VKT. Although some diversity is lost due to the use of aggregated data, the procedure exposed the underlying relationships and allowed the development of a practical model. The square root transformation, on the other hand, proved to be a reasonably easy function to work with when making predictions or assessing impacts.

6 Model Validation

Using the recommended model, the *predicted* household VKT was computed for each travel zone and mapped (Figure 1). The results indicate that the model is making sensible predictions. The map shows outlying areas with lower housing and employment densities to be high generators of VKT. In comparison, areas nearer the CBD and those with higher densities generate less. In addition, areas near train lines produce less VKT compared to those located farther away. Even in fringe areas, there is a clear difference between areas located near and far from public transport nodes. For example, areas along the train line in the Blue Mountains, Penrith, Blacktown, Liverpool and Campbelltown have comparatively less VKT than areas farther from these stations.

In comparison to *actual* VKT (Figure 2), the predicted VKT shows similar patterns overall. This suggests that the model has a relatively good fit. In fact, the predicted VKT presents a more coherent pattern than actual VKT which has some areas showing rather unexpected results. This is because as previously indicated, actual VKT has a relatively high variation even at a TZ level. For example, the northern and southern areas in the Blue Mountains have less VKT than those near the train line based on actual VKT. On the other hand, the predicted VKT shows the expected pattern with those areas close to the Blue Mountains rail line generating less VKT.

As a further check, the difference between the predicted and actual VKT were also plotted (Figure 3). This map highlights areas where the predictions are more closely aligned to the actual VKT. On the whole, most areas have a small difference. The exceptions are in

outlying areas where the smaller sample sizes are impacting on the variability of actual VKT and in turn are affecting the difference to the predicted values.

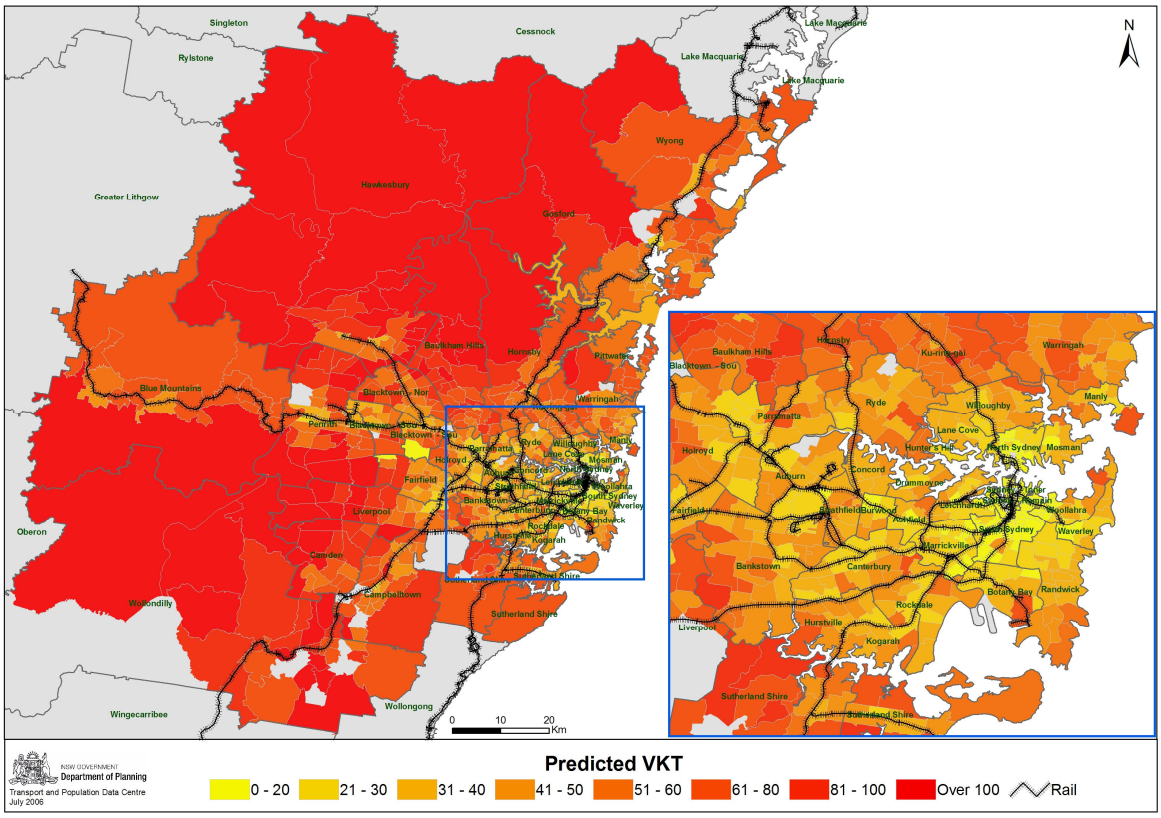

Figure 1 *Predicted* household VKT by travel zone

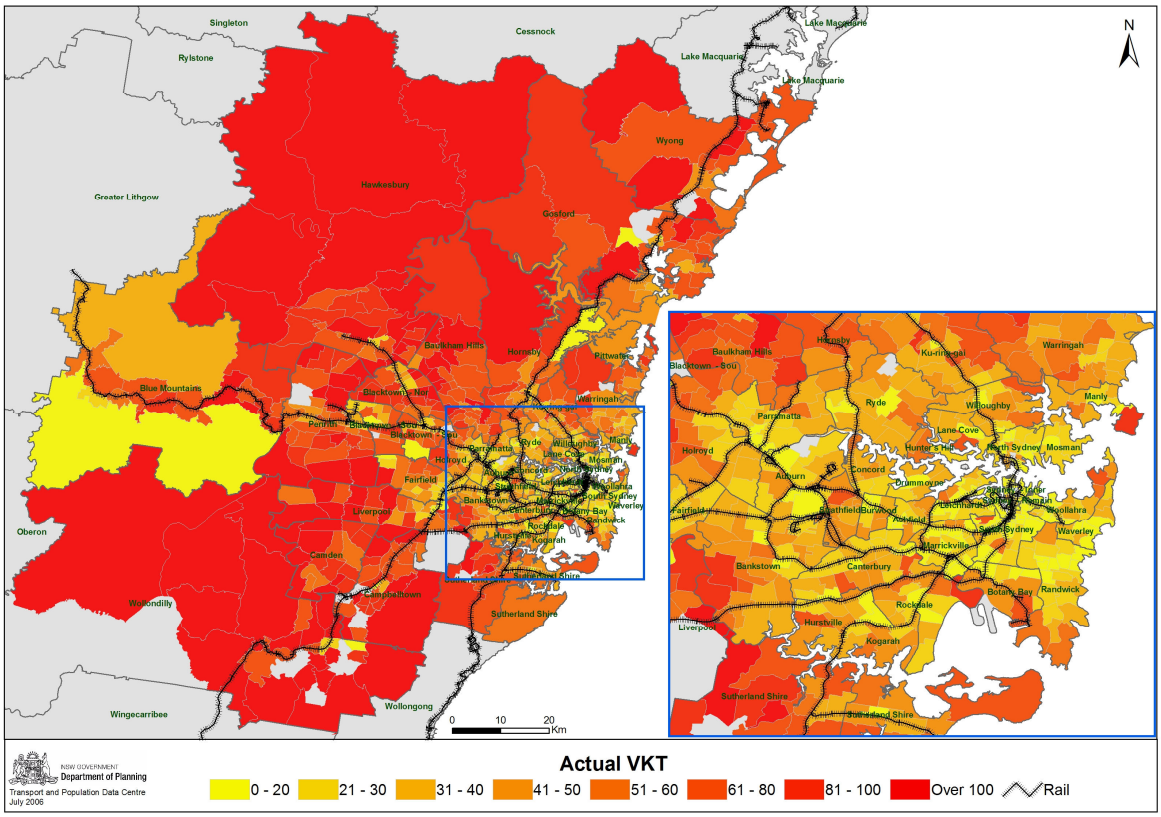

Figure 2 *Actual* household VKT by travel zone

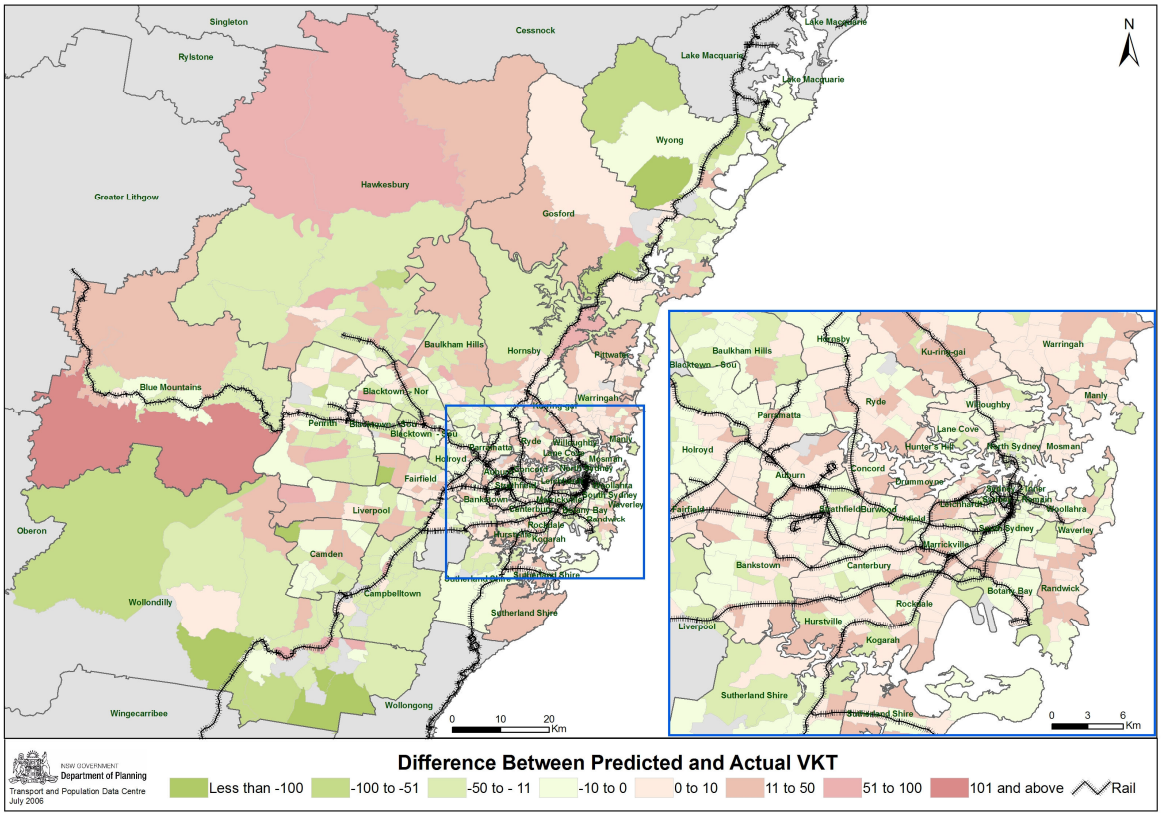

Figure 3 *Difference* between actual and predicted household VKT by travel zone

7 Further work

Replication of the approach to other cities would be beneficial to help assess the transferability of the findings. If demonstrated to be broadly transferable, then the results may be extrapolated for application in similarly configured areas where resources for such a study may be unavailable.

As with all models, further refinement can always be made. For the time and resources that were accorded to this project, the researchers believe that the recommended model was the best that can be achieved. Further work can be done on the sourcing and inclusion of data on other neighbourhood design characteristics such as road configuration and the availability of parking. This project may also benefit from more analysis on the impacts of ordinal variables, especially dwelling type which did not feature in the final model, and the use of distances based on XY coordinates.

8 Summary and recommendations

This paper describes the development of the Sydney VKT Regression Model, designed to predict the number of vehicle kilometres travelled, given a set of socio-economic, locational and urban form characteristics.

Numerous trials were undertaken before settling on the final model which is based on data aggregated to the travel zone level and where the dependent variable has been transformed to the square root of household VKT to deal with a number of statistical issues. This leads to some limitations in the model but none that may be considered as significant to the intended application. The final model related the dependent variable, the square root of household VKT to the number of vehicles if the household, closest distance to major centre or CBD,

land use mix, local employment, housing density and distance to nearest train, ferry, light rail or high frequency bus.

Graphs comparing actual and predicted VKT by location in Sydney show that the model is making sensible predictions, with areas nearer the CBD and those with higher densities generating lower VKT per household. In addition, there is a clear decrease in VKT generation in areas located close to public transport nodes.

The model will be used as part of the quantitative assessment scheme known as METRIX to inform land use planning by predicting VKT from proposed developments enabling their evaluation against the Metro Strategy objective of reducing growth in VKT. In addition, the model can be applied to gauge the impacts of various development scenarios at a broader sub-regional level.

To those who will undertake similar work, the authors have the following lessons to share:

- A significant part of the work is the collation of the data required in the model, ensuring its quality, understanding the relationships and correctly specifying it.
- Many trials should be undertaken using the disaggregated data before deciding to use aggregated data in the model.
- The decision to use aggregated or disaggregated data, and variable transformations should be taken in consideration of the final application of the model.
- Identifying the variables that are significantly related to household VKT is only a part of the model specification work. The other part is choosing which variables to include considering the relevant statistical issues and the practical importance of each variable in the model application.
- The procedure itself would be reasonably easy to replicate once the required data is available.
- A vehicle ownership model will be a useful addition to this exercise.

Appendix 1 About the Household Travel Survey

The Household Travel Survey

Up to 1991, large one-off household travel surveys were conducted in Sydney in ten-year intervals. The last of this was the 1991 Home Interview Survey (HIS) which had a sample of over 12,000 households. Beginning in 1997, a new data collection strategy was implemented that would provide personal travel data on a continuous basis in order to meet the need of transport data users for more timely data. This continuous survey was called the Household Travel Survey (HTS). The HTS sampling methodology was developed for TPDC by the Australian Bureau of Statistics (ABS) with an annual sample of about 3500 households.

The HTS uses the face-to-face interview method which is carried out every day of the survey period. A travel diary is used by each householder to record the details of all travel undertaken for their nominated 24 hour period. For each trip, the interviewer records the mode of travel, trip purpose, start and end location, and time of departure and arrival. Vehicle occupancy, toll roads used and parking is recorded for car trips and fare type and cost for public transport trips. Detailed socio-demographic information is also collected on the household, including dwelling type, household structure and vehicle details, as well as age, gender, employment status, occupation and income of individual household members.

Geographical coverage

The HTS is conducted over an area which includes the Sydney and Illawarra Statistical Divisions (SD) and the Newcastle Sub-Statistical Subdivision as shown in Figure 4. This area extends from Port Stephens in the north to Shoalhaven in the south and the Blue Mountains in the west. For this project, only data from surveyed households located in the Sydney Statistical Division were included.

Figure 4
Geographical Scope of the HTS

Survey sampling design

The HTS uses a stratified, three-stage cluster sampling method. The stratification is by Statistical Local Area. Temporal allocation of the sample is also undertaken. This refers to the allocation of the sample to days of the week and weeks of the year as evenly as possible over the survey period. This design ensures the geographic and temporal representativeness of the dataset.

The survey data

The data used in the model came from the first seven waves of the HTS which correspond to survey data collected on travel for *all days* of the week from June 1997 to June 2004. Only data for fully responding households[15] located within the Sydney SD were included.

[15] Part responding households provide an incomplete enumeration of household VKT and were therefore excluded in the analysis.

Appendix 2 Variables in the model

Variable	Definition	Notes	Transformations[16]	Source of Data
Dependent Variable				
VKT generated by the household			Natural log, Square Root	HTS
Predictor Variables				
Locational variables				
Distance to the Central Business District (CBD)	Distance (km) from the household to designated centre of Sydney CBD			HTS
Distance to Regional Centres	Distance (km) from the household to the nearest Regional Centre			HTS
Accessibility to nearest non-road public transport (train, ferry or light rail)	Accessibility score = 'wait time' + 'walk time'	Wait time = ½ the frequency of the service e.g. wait time is 7.5 minutes if frequency is every 15 minutes. Walk time = in minutes based on distance in kms x 15 minutes	Accessibility scores as 5-6 point ordinal following an approach described in Primerano (2004)	HTS, Sydney Public Transport Information
Accessibility to nearest bus service	As above	As above	As above	As above
Level of local employment	Number of jobs within 5 km radius from the centroid of the CD[17]			ABS 2001 Census
Socio-demographic variables				
Number of vehicles in the household	The number of vehicles usually parked at the household's dwelling			HTS
Number of persons of driving age			Number of persons aged 16 years and above Number of persons aged 18 years and above	HTS
Number of licence-holders				HTS
Household type			A nine-category and five-category variable were tested.	HTS
Number of persons in the household				HTS
Income			Total household income ($000s) Average income per person in the household ($000s)	HTS
Urban form / Neighbourhood design variables				
Dwelling structure	For example: Separate house, Semi-detached, Flat or unit, etc.	Not included in the Canadian Model (IBI Group 2000) but was tested in a South Australian model (Soltani and Somenahalli 2005)	An eleven-category and four-category variable were tested.	HTS
Housing density	No. dwellings in CD centre	Included as shown to be relevant in a South Australian model (Soltani and Somenahalli 2005)		ABS 2001 Census
Level of social and commercial services available locally	Total area with business zoning, to represent shops and services outlets, within a kilometre radius from the centroid of the CD		Total area in hectares as well as square metres were tested.	Local Environment Plan (LEP) data
Land-use mix	This measure is: $- \sum = (pi)^* \text{Ln}(pi) / \text{Ln}(s)$ where pi = area of each land-use (i.e. separate houses; multi unit, commercial & mixed-use) Ln is the natural log S = the number of land-uses			Local Environment Plan (LEP) data

[16] Some other versions and/or transformations of the variables were tested in addition to the standard definitions.
[17] Census Collection District

References

Boarnet M and Crane R (2001) The influence of land use on travel behaviour: specification and estimation strategies. *Transportation Research Part A 35*. pp 823-845

Brunton P and Brindle R (1999) *The relationship between urban form and travel behaviour* Vermont South: ARRB Transport Research Ltd.

Burke M and Brown L (2005) Rating the Transport Sustainability of New Urban Developments: a starting point and ways forward. *Papers from the 28th Australasian Transport Research Forum.* Sydney: ATRF

Chandra L (2005) Putting the transit into Transit Oriented Development. *Papers from Transit Oriented Development - Making it Happen.* Fremantle: Planning and Transport Research Centre (PATREC)
http://www.patrec.org/conferences/TODJuly2005/papers/Chandra.L.pdf
[accessed March 2006]

Horowitz, J (1995) Example: Modelling Choices of Residential Location and Mode of travel to work *The Geography of Urban Transportation Second Edition*. ed Hanson S. New York: The Guildford Press. pp 219-239

IBI Group (2000) *Greenhouse Gas Emissions from Urban Travel: Tool for Evaluating Neighbourhood Sustainability* Canada Mortgage and Housing Corporation

New South Wales Department of Planning (2005) *City of Cities, A Plan for Sydney's Future*

Neter J, Kutner M, Nachtsheim C and Wasserman W (1996) *Applied Linear Statistical Methods*. USA: Times Mirror Higher Education Group Inc

Primerano F (2004) *Development of Accessibility Measures for Transport and Urban Planning*. a doctoral thesis submitted to the University of South Australia. pp 130-133

Soltani A and Primerano F (2005) The Travel Effects of Community Design. *Papers from the 28th Australasian Transport Research Forum.* Sydney: ATRF

Soltani A and Somenahalli S (2005) Household Vehicle Ownership: Does Urban Structure Matter? *Papers from the 25th Australasian Transport Research Forum.* Sydney: ATRF