Validating stated preference surveys through the use of hedonic regression models with an application to housing prices around transit oriented developments

B. Smith & D. Olaru
University of Western Australia, Australia

Abstract

A survey of factors affecting residual relocation choices for families who had recently moved onto the Mandurah Railway corridor (Perth, Western Australia) has been undertaken using stated preference methods. While stated preference models are appropriate tools for valuating attributes of choice, there remains the concern of realism in an experimentally controlled environment. In this paper we validate stated preference models using hedonic pricing regressions based on observed real estate prices for three Perth-Mandurah Railway precincts. The results of the regressions are compared to those of the discrete choice models by examining the levels of significance of housing or neighbourhood characteristics in each modelling paradigm, as well as a comparison of the closeness of the valuations for these characteristics.

The findings indicate that factors affecting residential location are consistent whether observed by experimental data or by revealed choices in the market. From a planning perspective, the results indicate that not only housing features, but also neighbourhood characteristics such as proximity to public transport hubs or local schools affect residential property values.
Keywords: discrete choice modelling, hedonic regressions, housing attribute valuation, transit-oriented development.

1 Introduction

This research investigates similarities and differences between hedonic pricing and discrete choice modelling for housing valuation along transit-oriented

WIT Transactions on The Built Environment, Vol 116, © 2011 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT110261

developments (TOD) in Perth. Both modelling approaches are used to investigate the valuation of housing, neighbourhood and transport attributes. However, the models differ in the type of data and multivariate technique. The discrete choice model is estimated based on location choice stated preference data, elicited from households when asked to compare hypothetical properties with various characteristics. The hedonic pricing model is based on observed prices and characteristics of traded properties in the Real Estate Institute of Western Australia (REIWA) traded properties database.

The fundamental proposition in both methods is that a residential property is a composite of multiple characteristics, each of which contributes to its price. The property value embodies commonly recognised tangible housing attributes (size of the block, number of bedrooms, or garage presence for example) or accessibility to transport, as well as the less tangible environmental or landscape elements. The results given here indicate that TOD features play a role in the values of the property, whether sourced from stated preference or market data.

The paper has the following structure: Section 2 surveys the relevant literature and Section 3 briefly presents the data sources used in this research. The results are presented in Section 4. The paper concludes with discussion of results and some suggestions for future research.

2 Two models of housing choice

2.1 Hedonic regression models

The hedonic pricing (HP) model assumes the utility for a household is formed over housing attributes and the price the consumer, with income y, is willing to pay is a combined value of intrinsic characteristics of the dwelling (d), natural environment (n) and the built environment (b). Let X be the *amount* of residential characteristic provided by a housing alternative. The consumer places a value P(X) of the house as a combined value of the sources of utility:

$$P(X) = P(X_d) + P(X_n) + P(X_b), \text{ subject to } y=P(X)+G \quad (1)$$

where G is the expenditure made on non-housing purchases available at residual income y-$P(X)$, i.e., the consumer's calculation of the price she/he is willing to spend on housing to leave sufficient income for other needs and wants. The decomposition of pricing components need not be linear as given in eqn (1).

Rosen's [1] seminal work on implicit pricing indicates that market equilibrium conditions exist such that the consumer's perceived value $P(X)$ of the house is equal to its transacted price, P_i. Furthermore the valuation of any one attribute is recoverable by examining the *marginal rate of substitution* between residual income and the quantity of housing attributes, X. An assumption that goods are freely traded on an open market, without transaction or search cost, is made and the implicit prices are recoverable from the Ordinary Least Squares (OLS) regression:

$$P_i = \sum_d p_d X_{id} + \sum_n p_n X_{in} + \sum_b p_b X_{ib} + \varepsilon_i \quad (2)$$

The implicit prices are regression parameters in eqn (2). A second demand equation using prices as instrumental variables may be undertaken to determine the household's demand for intangible goods. Nelson [2] was first to use the technique for the study on demand for clean air. Empirical work in this area is ongoing (Brasington and Hite [3]; Day *et al.* [4]; Bayer *et al.* [5]). In our study we are validating the *implicit prices* obtained from an experiment by using external data. The implicit price regression, eqn (2), is all that is required.

The distinct advantage of the HP approach is that the transaction price and housing attributes may be sourced from real estate industry records and augmented by geographic information systems (GIS) databases which contain locality attributes such as transport facilities, employment densities, or the area of green space. This is relevant because omitted spatial characteristics lead to less robust models (Yoo and Kyriakidis [6]; Paéz *et al.* [7]).

However, transaction data do not contain information about the purchaser, which limits the modeller's capacity to understand preference heterogeneity. Without information about household attitudes, there is no avenue to capturing relationships between housing attributes and household aspirations, such as comfort, safety or status.

2.2 Discrete choice models (DCM)

Discrete Choice Models (DCM) remains highly influential in the valuations literature and it has been successfully applied in transport for valuing time, air pollution and noise (Ortúzar and Willumsen [8]). Unlike the hedonic pricing model the dependent variable is the choice of house. The model assumes a consumer faced with a finite choice set of properties chooses the one that best suits their need and their budget. This is represented by the modeller as a random utility in which V, the systematic utility, depends not only on the vector of dwelling characteristics (X_d, X_n, X_b), but also on the household's attitudes and lifestyle (Z_a) and its socio-economic characteristics (Z_{se}):

$$\mathrm{V_i} = f\left(X_{\mathrm{i}d}, X_{\mathrm{i}n}, X_{\mathrm{ib}}, y - P_{\mathrm{i}} \;; Z_a, Z_{se},\right) \quad (3)$$

The unobserved part of the utility is modelled as a random variable. If the error is assumed to be independently and identically distributed extreme value type 1, the resulting probability choice system is a multinomial logit, giving the probability π_{i} that the household chooses dwelling i.

$$\pi_{\mathrm{i}} = \frac{exp\left[p_d X_{id} + p_n X_{in} + p_{\mathrm{b}} X_{ib} + \lambda\left(y - P_i\right)\right]}{\sum_{\mathrm{j=1}}^{\mathrm{J}} exp\left[p_d X_{\mathrm{j}d} + p_{\mathrm{n}} X_{\mathrm{j}n} + p_{\mathrm{b}} X_{\mathrm{j}b} + \lambda\left(y - P_i\right)\right]} \quad (4)$$

McFadden [9] provides a probabilistic analogue to Rosen's implicit pricing by expressing the marginal substitution between housing attributes and residual income as a ratio between one of the estimated marginal utilities of housing attributes, p_d, p_n or p_b, and the marginal utility of money, λ. Socio-economic characteristics may enter this model as taste moderators for the marginal prices of housing or environmental attributes. However, attitudinal data accounting for preference heterogeneity by way of latent constructs require more advanced discrete choice models. See Olaru *et al.* [10] for constructing latent classes from attitudinal data or Yañez *et al.* [11] for econometric methods to include latent variables in a choice model.

Ellickson [12] adapted the multinomial logit, eqn (4), to the housing market and his model was modified to represent the competition among buyers by Lerman and Kern [13]. Cropper *et al.* [14] examined the strengths of the hedonic regression model with Box-Cox transformations and the multinomial logit model at recovering known parameters in a simulation experiment. They conclude that the hedonic regression perform better at capturing implicit prices, but the multinomial logit model performed better on welfare calculations for large changes in housing attributes. It should be noted that even at the time of their investigation, the multinomial logit model was superseded by a variety of more flexible models. Since that time the estimation of mixed-logit models (McFadden and Train [15]) have greatly improved the efficiency and explanatory power of discrete choice models.

The discrete choice models used in this research are based on stated-preference data where the respondents make a choice between two hypothetical dwellings. Stated-preference techniques are widely used in transportation and ecological evaluation research (Louviere *et al.* [16]). However, using experimental data will always lack the external validity of observations in the market. The aim of this paper is to use market observations to validate stated preference models and inferences.

While revealed preference data using discrete choice models is favourable, it does require knowledge of the household's choice set. Earnheart [17] interviewed each respondent to obtain a choice set of two houses, one selected and one rejected. However, such information is not readily available in secondary data sources from real estate market institutes. This paper will use hedonic regression to examine the validity of the estimated implicit prices from the discrete choice modelling exercise. In the next section the empirical setting is described and the validation results are presented in section 4.

3 Empirical setting

3.1 Data sources

This research draws on two sources of data: a) primary data obtained in a quasi-longitudinal study conducted in three transit-oriented development (TOD) precincts along the Perth-Mandurah rail corridor, in Western Australia; and b)

secondary data, purchased in May 2010 from REIWA [18], including property sales between 2006 and 2008 in the three precincts.

3.1.1 Survey information

The overall objective of the data collection was to assess the behavioural responses to emerging TOD precincts. Households within Bull Creek, Cockburn Central, and Wellard station precincts (5 min drive from the station) have been surveyed at pre-rail station opening (November-December 2006), and twice after the opening of the railway corridor (July-September 2008 and 2009). A detailed description of the three-waves data is available in Curtis and Olaru [19]. Using computer-assisted surveys, revealed and stated preference data for household and individual characteristics, car ownership, travel behaviour, location preferences, physical activity and mobility restrictions were collected.

The households were asked to select between two houses and locations with different attributes. The attractiveness of one or other alternative, determined by the combination of features, was used to infer the worth of each characteristic for the household. Three categories of attributes, contributing to the housing selection were included: dwelling, facilities and quality of neighbourhood, and travel.

A sample of 539 interviewed households resulted in 4,094 scenario observations for model estimation, as presented in the following section. Table 1 shows the prices of the houses along with their block size and age of the house for the sample of households interviewed in wave 1 2006 of the TOD study.

Table 1: Dwellings of households interviewed in the three precincts: average and (*standard deviation*).

Variable	Bull Creek	Cockburn Central	Wellard
House value($)	720,000 (*295,200*)	455,000 (*134,500*)	329,500 (*104,400*)
Age house (years)	27 *(12)*	14 *(7)*	22 *(16)*
Block size (m^2)	651 *(231)*	722 *(1,071)*	815 *(491)*
N	285	321	342

3.1.2 Real estate data

Information on 6,665 properties, transacted between 2006 and 2008, was obtained from REIWA. The purchased data set included: transaction price and date, size of the block, number of bedrooms, bathrooms, presence of dining area, family, carport, garage, swimming pool, wall type, and year when the house was built.

Filtering and data checking identified several errors or missing data. Variables with significant missing data were not included in the modelling.

Google Earth was then used to determine the Euclidean distance between the house and the railway station and the presence of a swimming pool.

As less than 3% of the properties had repeated sales records, a cross-sectional hedonic model was proposed. In addition, because the transaction dates were different for the houses in our precincts, we used housing indices to estimate the 2006 values of the houses.

4 Results

To validate the results of the discrete choice models with real data, we ran the two types of models separately and then we compared the findings. The reader may turn to [10] for a full description and summary of the discrete choice models. The hedonic regression models are presented with standardised regression weights and all parameters are significantly different from zero (see Tables 3 and 4, further on).

Table 2, shows the willingness to pay for housing attributes. The discrete modelling results indicate that, other things being equal, people prefer bigger houses/blocks, in greener neighbourhoods, closer to all facilities, but had a lower preference for reduced travel time or cost. Within their budget constraints, households trade-off the dwelling and surrounding features, moving closer to schools, shops, transport, medical facilities, to offset the generalised transport costs.

Housing values are higher in Bull Creek compared to Cockburn Central and Wellard and significantly higher valuation of access to the schools and shops in Bull Creek compared to the other two precincts. The reputation of Rossmoyne Senior High School attracts numerous families with children willing to "incorporate" in the housing prices potential school fees. The highest block size valuation in Bull Creek reflects the relative proximity to the city and the position of the suburbs along the Canning River. The lowest valuations are in the Wellard precinct and they are mirrored in the lowest housing costs (about 2/3 of the average metropolitan median price).

Table 2: Housing valuation using SP data.

Attribute (Values measured in thousand AUD)	Bull Creek	Cockburn Central	Wellard
Additional storey in the house	**144.3**	**124.6**	**14.6**
Additional m^2 in the block	**1.28**	**0.50**	**0.34**
Move 1 minute closer to school	11.0	3.51	1.81
Move 1 minute closer to shops	19.6	9.02	2.50
Move 1 minute closer to train station	6.56	3.26	1.81
Improved amenity	116	89.6	41.4
Save 10 min travel time/day	5.28	4.97	0.46

The first run of validation uses a hedonic regression model, eqn (2), based on the available data from the secondary source. The results are given in Table 3. Unfortunately the data set only matches the experimental design for one variable, being block size. The two modelling paradigm show a fair degree of consistency; the valuations being less than 10% different for each of the three precincts. If we take number of bedrooms/ number of storeys to be proxy variables for indoor size we could make tentative comparisons across data sets. The degree of similarity is reflected in the order of magnitude across the precincts for each data set. However, while the experiment tends to moderately overstate the importance of house size for Bull Creek and Cockburn Central, there is a large discrepancy across data sets for Wellard. This is most likely due to the difference in variable definition as people in Wellard tended to want single storey houses.

It is worth mentioning that in Cockburn Central area, more properties have carports than garages compared to Bull Creek and Wellard, which may affect the household preferences for carports and their valuation.

Table 3: Housing valuation using real estate data.

Characteristic	**Bull Creek**		**Cockburn Central**		**Wellard**	
of the dwelling	B	Beta	B	Beta	B	Beta
Number bedrooms	126.7	0.577	89.8	0.757	69.7	0.810
House area (m^2)	1.31	0.257	0.47	0.154	0.319	0.127
Dining area	32.2	0.035	26.4	0.046	15.0	0.039
Carport	-66.4	-0.025	71.9	0.042	-17.8	-0.033
Garage	16.6	0.082	5.2	0.003	58.3	0.031
Swimming pool	153.9	0.114	34.5	0.045	31.8	0.031
R2-adj	80.8%		90.7%		90.04%	
N	2,233		3,230		1,202	

Note: All parameter estimates statistically significant at 0.01 level.

A second validation is undertaken by only estimating the hedonic equations for households in the experimental survey. However, a number of these houses were not traded during the data collection period. In this case a nearest neighbour estimate is used by finding the closest match to missing household records with a property in the data files. This will mean that comparison is somewhat more tentative. The results are given in Table 4. The relationship between precincts and block size valuations is constant across the three models, except the second hedonic regression sample (matching or nearest neighbour dwellings) had lower block size valuations that the other models. The house size valuation (2^{nd} storey) is much larger in the second regression model. However, the most significant finding is valuation for the distance to the railway works in the opposite direction as expected from the experiment for the Bull Creek and Cockburn Central precincts. This is entirely due to fact that the precincts' rail stations are in the centre of the freeway and the value of being away from the freeway outweighs

the desire to be closer to rail station. This shows a clear example of where an experiment omits important contextual variables. This does not invalidate the partial worth estimates from the discrete choice, all else being equal households will pay for proximity to public transport nodes, but in some cases all else is not equal.

The results presented here have prompted us to think about a more structured way to use secondary data to develop an experimental design and to validate the outputs. The following section outlines this proposal.

Table 4: Housing valuation using real estate data (only properties for households interviewed in the three precincts).

Characteristic of the dwelling	Bull Creek		Cockburn Central		Wellard	
	B	Beta	B	Beta	B	Beta
Block area (m^2)	0.719	0.634	0.335	0.279	0.146	0.407
Additional storey	283	0.460	175	0.381	238	0.695
Age house (years)	5.52	-0.204	-7.42	-0.237	-1.96	-0.149
Distance from railway station (km)	14.2*	0.048	10.3	0.056	-12.0	-0.095
Swimming pool	53.3*	0.036	28.8	0.025	0.804*	-0.001
R2-adj	89.8%		96.6%		96.6%	

*Not significant at the 5% level.

5 Conclusions and future directions

The paper presents a validation method for stated preference surveys on preferences for non-traded housing attributes. The method takes a novel approach by recognising the parallel in the micro-economic theoretical underpinnings of two different classes of models. The hedonic regression is applied to market based data where each observation is limited to the purchased dwelling. Discrete choice models, on the other hand, require observations on the attributes of the alternatives rejected by the consumer as well as the purchased dwelling. However, discrete choice models do lend themselves to experimentally controlled stimuli and repeated observation of choice. These models provide a richer insight into household decision processes.

The validation exercise shows some promise in confirming the usefulness of discrete choice estimates for public policy, which requires inferences to a population level. However, the exercise has also highlighted the potential for local contextual variables (i.e. the freeway and railway sharing the same area) to create a gap between experimental results and market based behaviour.

A possible approach to building more robust models in terms of their insight to household decision making and applicableness to policy and population inference is outlined in Figure 1. Efficient experimental designs rely on a

preliminary asymptotic variance-covariance matrix for the proposed model (Bliemer *et al.* [20]), This, in turn, requires prior parameter estimates, which may assumed to be equal to zero. However, the design is improved by having accurate prior estimates. Hedonic regressions done at stage 1, from a secondary data source, offer an alternative to conducting a pilot survey. The design is optimised using genetic algorithm in stage 2. The experimental survey is delivered and models are estimated (Stage 3). A post estimation validation, like the one outlined in this paper is performed at stage 4, using a holdout sample from the secondary data source. The welfare measures made at stage 5 are balanced between the validity to population inferences and richer understanding of household valuation functions.

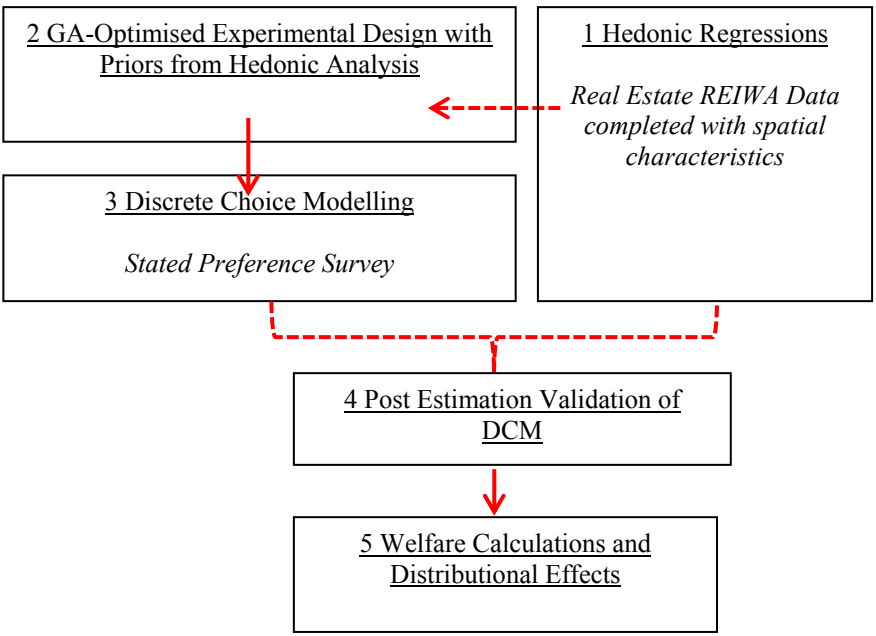

Figure 1: Hybrid hedonic pricing and discrete choice modelling.

Acknowledgements

This research has received funding from the ARC (Linkage Project LP0562422) including 11 partner organisations: Department for Planning and Infrastructure WA, Public Transport Authority of WA, LandCorp (Western Australian Land Authority), The Village at Wellard Joint Venture, Subiaco Redevelopment Authority, City of Melville, East Perth Redevelopment Authority, City of Cockburn, Midland Redevelopment Authority, Town of Kwinana, City of Rockingham.

References

[1] Rosen, S., Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*, **82(1)**, pp. 34-55, 1974.

[2] Nelson, J. P., Residential choice, hedonic prices, and the demand for urban air quality, *Journal of Urban Economics*, **5(3)**,pp. 357-369, 1978.

[3] Brasington, D. M. and Hite, D., Demand for environmental quality: a spatial hedonic analysis. *Regional Science and Urban Economics,* **35(1)**, pp.57-82, 2005.

[4] Day, B., Bateman, I. and Lake, I., Beyond implicit prices: recovering theoretically consistent and transferable values for noise avoidance from a hedonic property price model. *Environmental and Resource Economics* **37(1)**, pp.211-232, 2007.

[5] Bayer, P., Keohane, N. and Timmins, C., Migration and hedonic valuation: The case of air quality. *Journal of Environmental Economics and Management,* **58(1),** pp. 1-14, 2009.

[6] Yoo, E.H. and Kyriakidis, P.C., Area-to-point Kriging in spatial hedonic pricing models, *Journal of Geographical Systems*. **11(4)**, pp. 381-406, 2009.

[7] Paéz, A., Recent research in spatial real estate analysis, *Journal of Geographical Systems*, vol.**11(4),** pp. 311-316, 2009.

[8] Ortúzar, J. de D. and Willumsen, L.G., *Modelling Transport (3rd ed.),* Wiley & Sons: Chichester, UK, 2001.

[9] McFadden, D., Econometric models of Probabilistic Choice (Chapter 5). *Structural Analysis of Discrete Data with Econometric Applications*. ed. D. F. Manski and D. McFadden, MIT Press: Cambridge, pp. 198-272, 1981.

[10] Olaru, D., Smith, B., and Taplin, J.H.E., Residential Location and Transit-Oriented Development in a New Rail Corridor. *Transportation Research A*, 45(3), pp. 219-237, 2011.

[11] Yañez, M.F., Raveau, S., Rojas, M., and Ortúzar, J. de D., Modelling and forecasting with latent variables in discrete choice panel models, *Proc. of ETC*, Leeuwenhorst, The Netherlands, available at: http://etcproceedings.org/paper/modelling-and-forecasting-with-latent-variables-in-discretechoice-panel-model, 2009.

[12] Ellickson, B, An alternative test of the hedonic theory of housing markets. *Journal of Urban Economics,* **9(1),** pp. 56-79, 1981.

[13] Lerman, S. R. and Kern, C. R., Hedonic theory, bid rents, and willingness-to-pay: Some extensions of Ellickson's results. *Journal of Urban Economics*, **13(3), pp.**358-363, 1983.

[14] Cropper, M. L., Leland, D., Kishor, N. and McConnell, K. E., Valuing Product Attributes Using Single Market Data: A Comparison of Hedonic and Discrete Choice Approaches. *The Review of Economics and Statistics* **75(2),** pp.225-232, 1993.

[15] McFadden, D. and Train, K, Mixed MNL models for discrete response. *Journal of Applied Econometrics,* **15(5),** pp. 447-470, 2000.

[16] Louviere J.J., Hensher D.A., and Swait J.D., *Stated Choice Methods: Analysis and Application*, Cambridge University Press: Cambridge, 2000.
[17] Earnhart, D., Combining Revealed and Stated Data to Examine Housing Decisions Using Discrete Choice Analysis. *Journal of Urban Economics,* **51(1),** pp.143-169, 2002.
[18] Real Estate Institute of WA (REIWA) Price growth by suburb Web Site, http://reiwa.com/res/res-pricegrowth-display.cfm.
[19] Curtis, C. and Olaru, D., The impacts of a new railway: Travel Behaviour of Residents in New Station Precincts. *Proc. of 12th WCTR*, Lisbon, Portugal, 2010.
[20] Bliemer, M.C.J., Rose, J.M., and Hensher, D.A., Efficient stated choice experiments for estimating nested logit models. *Transportation Research B*, 43(1), pp. 19-35, 2009.