Correcting for Sample Selection Bias in UK House Price Indices: A Duration Model Approach

Gwilym Pryce[1]
Department of Urban Studies,
University of Glasgow G12 8RS
g.pryce@socsci.gla.ac.uk

ERES, Dublin, 17th June 2005

Working Draft – Please do not quote without permission

ABSTRACT: This paper aims to develop a practical way of correcting for sample selection bias in UK regional house price indices. Such indices are invariably based on transactions data (such as the Land Registry house price index). Unfortunately, properties that transact in any given period are unlikely to be a random sample of all properties, either in terms of location or type. Consequently, such indices may not give a reliable measure of the rate of increase in the value of the housing stock, a variable that is crucial to our understanding of wealth inequality and a range of housing policy decisions. While correction methods have been developed in the US, there is currently no UK equivalent, partly because valuation records are not kept on the whole stock. This paper explores the possibility of developing a reliable correction procedure using data readily available in the UK. A series of duration models are developed for length of stay (based on more than a quarter of a million observations) to establish whether or not the selling decision is characterised by duration dependence. We find very strong evidence that this is indeed the case. We conclude that probit regression (the method used in US studies to control for sample selection bias) may not therefore be the most appropriate method for estimating the inverse Mills ratio.

Introduction

Prices play a pre-eminent role in our understanding of any market. Housing is no exception. Housing, however, is more complex than most goods because each dwelling represents a unique mix of attributes and location. As a result, no one method of measuring house prices can capture all aspects of housing market dynamics. The challenge for housing economists and policy makers is finding the right measure of house price change for the task in hand. There are already many different indices available. A problem common to them all (in the UK at least), however, is that they are based on transactions. Dwellings sold infrequently are underrepresented in such indices

[1] This research stems from research funded by the Office of the Deputy Prime Minister.

and this can cause bias. For some uses of house price information, this bias does not matter. However, I argue that, for certain policy decisions, particularly for those relating to the supply of new housing, transactions bias could result in significant distortions. This is because houses that sell infrequently tend to belong to different submarkets than those that sell frequently and have a different pattern of response to general economic conditions and also to new supply.

Following HM Treasury Barker Review of UK housing Supply (2004), it has become clear that one of the key requirements of a house price measurement is that it must be enable the production of reliable estimates of the impact of new supply. The Barker Review Interim Report (2003, p.58) estimated that the current number of new homes per annum would need to be doubled in order to "achieve the European trend rate", and "more than double to get real price stability". Such estimates are founded on transactions-based indices, however, and we need to ask whether these are the appropriate measures for such analysis. If, for example, newly constructed dwellings tend to be of a type that sell frequently, they will be in the same submarket as properties that repeatedly enter the official indices of house price inflation, and may appear to have the desired dampening effect on those indices. However, these new dwellings may have relatively little impact on the prices of infrequently sold properties, particularly if the latter are of a markedly different size and type. There is a need, therefore, to find a practical way of correcting for sample selection bias in UK house price indices.

The paper is structured as follows: first I offer a discussion of the possible policy implications of using transactions-based house price indices. This discussion of the policy implication is hypothetical because there is little empirical research in the UK on tansactions bias and its implications for policy analysis. This is due to the lack of data in the UK on the housing stock as a whole. The goal of the remainder of the paper, therefore, will be to consider how transactions bias can be measured in the UK given the data limitations. We will first present simple house price inflation estimates for properties with different frequency of sale. We shall then set out a methodology for correcting transactions bias using existing UK data.

Policy Issues

One of the most important implications of the discussion initiated in the Introduction is the choice of house price measure for use as the focus of national and regional supply policy. If the target is to reduce house price inflation, then policy makers have to be aware that current measures carry with them an intrinsic bias towards frequently traded properties. Does this computational bias reflect the desired policy bias? It may be, for example, that there are differential house price inflation rates between different sectors of the housing market according to frequency of sale. If less frequently traded property is experiencing higher inflation than high turnover dwellings, then there may be a case for giving greater policy priority to alleviating demand pressures in the infrequently traded sector.

Policy implications are magnified if length of stay is to some degree a measure of consumer satisfaction with their property type and neighbourhood. Given the emotional and pecuniary upheaval associated with moving house, a family may only consider moving if they anticipate a significant improvement in living standards from doing so (other things being equal). A major reason for moving will therefore be dissatisfaction with current living conditions due to lack of space, unsuitable layout or due to neighbourhood problems. A policy that effects a major increase in the stock of frequently sold dwellings may achieve the superficial policy goal of reducing the rate of change of a transactions-based price index, but at the same time have resulted in an increase in the proportion of the total housing stock with which homeowners are generally dissatisfied.

Another example of the possible unintended consequences of policy founded on transactions-based indices may be to exacerbate the price differential between desirable and undesirable properties, and hence, exacerbate housing wealth inequality. If, as argued in the previous paragraph, standard indices implicitly encourage the construction of dwellings of a type that are frequently traded, then prices will be depressed in this sector relative to the larger, more desirable properties in the infrequently traded sector. As the recent Shelter report indicates, there are important long-term sociological implications of the growing housing gap that should not be ignored in the pursuit of short-term policy goals:

> "… children born this century will be starting life more financially unequal than has been the case since Victorian times… [T]he growing inequality in housing is marginalising a whole section of society... Those whose parents have housing wealth are more likely to be advantaged in childhood and to benefit from financial assistance, for example, in finding their own homes… For the children of the poor there will be large parts of the country to which they cannot consider moving in the future even if they should wish to. When they have problems in their lives, there will not be recourse to family wealth to bail them out, to help with a time when they cannot work or find work, to help pay their way through university…"

Whether or not one shares the political philosophy upon which the Shelter report is predicated, one cannot deny that there are potentially far reaching implications of using house price measures that potentially encourage policy decisions that inadvertently exacerbate housing wealth inequality.

A further possible consequence of founding policy on transactions-based house price indices is the distortion of household formation decisions. Household structure is not independent of the housing market. Whether and how individuals choose to form households is inevitably affected by the price structure of the house types. A housing policy that makes larger properties all the more expensive in relative terms will encourage the formation of small households, exacerbating the atomisation of society, and discriminating against ethnic groups that prefer to form large extended household structures. Construction of larger dwellings occupied by several generations of the same ethnic minority families may not only have a greater impact on price inflation of the total housing stock than the construction of small units, but may also do more to improve social cohesion, yet the construction of such properties may appear to have a relatively small effect on standard measures of house price inflation due to their infrequency of sale.

Finally, there may also be implications for local economic development. In depressed areas, the goal of policy may be to actually raise house prices (or this may at least be perceived to be an indicator that successful regeneration has taken place). Pryce and Gibb (2003) have presented preliminary evidence that new construction adjacent to a deprived area can have a regeneration effect. This positive force for change may be all the more potent if the dwellings being constructed are of high quality and of the kind that encourage the formation of stable, long-stay communities (rather than high-density, high-turnover estates). Using a transactions based measure of house price appreciation may therefore underestimate the positive effects of regeneration in those areas if the index includes few long-stay properties.

Understanding the role of frequency of sale is not just of relevance to policy makers, however. The bias it implies for house price indices has the potential to distort private sector investment decisions. The lack of an appropriately adjusted house price index will be a source of uncertainty for potential investors in either new construction or potential landlords. To make appropriate financial decisions, such investors need to be able to readily compare the performance of the housing sector with that of other tradable assets such as stocks and bonds and so lack of information in the housing asset market relative to other asset markets will further reduce the attractiveness of housing construction as a destination for investment funds. This applies to the macro decisions of institutional investors, but also to the small scale decisions of thousands of smaller investors and landlords who also have to make rational choices about how best to use their funds.

Correcting for Sample Selection Bias

Hedonic Regression

The heterogeneity of housing means that even if data were available on the whole housing stock, any estimate of house price change would need to control for differences in the attributes of dwellings. The hedonic method regresses house price on dwelling attributes, location and time variables. Estimated coefficients can then be used to predict the value of a constant quality dwelling for a given time period and location. Although a theoretical basis for the hedonic approach has been well established (Rosen, 1974), a range of specification issues have emerged in the literature (see Linneman 1980; Halvorsen and Pollakowski 1981; Butler 1982 and Malpezzi 2004). A more recently voiced concern has been the fact that 'house value indices derived from the conventional hedonic method are subject to bias if the sample of houses is not a random sample of the stock' (Gatzlaff and Haurin, 1998).

Correcting Bias

Gatzlaff and Haurin (1998, p.199) conclude that "Correction requires joint estimation of the probability that a house will sell and the sale price" (see also Quan 1993). This follows in the time honoured tradition of viewing the sample selection problem as one of omitted variable bias (Heckman 1979) where the omitted variable in the house price equation is the probability of the property coming onto the market. Gatzlaff and Haurin (1998) take this view and use porbit regression to estimate the probability of a property coming onto the market. This estimated probability[2] can then be entered into the sale price equation to correct for sample selection bias. More specifically, let S_{it}^* be the latent variable that drives the decision of household i of whether or not to sell in period t:

$$S_{it}^* = P_{it}^O - P_{it}^R \qquad [1]$$

where P_{it}^O is the offer price and P_{it}^R is the sellers reservation price. Offer prices are determined as follows,

$$P_{it}^O = V_{it} + e_{it}^O \qquad [2]$$

where e_{it}^O refers to the seller's preferences/information (assumed to be random normally distributed) and V_{it} is the market value of the house, which is a function of attributes and location,

$$V_{it} = \Sigma \alpha X^H{}_{\mathrm{ijt}} + \Sigma \beta D_t \qquad [3]$$

where $X^H{}_{\mathrm{ijt}}$ is a vector of j property and neighbourhood characteristics of the ith property at time t, and D_t is a dummy variable equalling 1 in the period that V_{it} is observed (t = 1 is omitted).

The seller's reservation price, P_{it}^R , depends on the opportunity cost of waiting for a better offer, and on the seller's knowledge of the determination of the value of the property, and the distribution of potential offers:

$$P_{it}^R = V_{it} + e^R{}_{it} \qquad [4]$$

It follows from [1], [2] and [4] that the latent variable determining whether the owner sells or not is given by the difference between the random two random variables associated with buyers and sellers respectively,

[2] More precisely, the inverse Mills ratio is calculated.

$$S_{it}^{*} = P_{it}^{O} - P_{it}^{R} = e_{it}^{O} - e_{it}^{R} \qquad [1]'$$

Gatzlaff and Haurin (1998) argue that these two error terms will determined by personal attributes x^p, and local or national macro economic factors z_t:

$$e_{it}^{O} - e_{it}^{R} = \mathrm{f}(x^p, z_t) \qquad [5]$$

Although S_{it}^{*} is not observable, we can observe the outcome S_{it} of the household decision of whether or not to sell, where $S_{it} = 1$ if the household decides to sell, and = 0 otherwise. These observed values can be explained using a probit regression with S_{it}^{*} as the dependent variable, and x^p, z_t as the explanatory variables. This estimated regression can be used to derive, λ_{it}, the hazard of non-selection ('inverse Mills ratio')

$$\lambda_{it} = \lambda_{it}(x^p, z_t) \qquad [6]$$

Applying Heckman's (1979) result to hedonic estimation, it can be shown that the omitted variable bias associated with having a non-random sample in the OLS estimation can be overcome by including λ_{it} as an explanatory variable. So in,

$$P_{it} = \Sigma\alpha X^{H}{}_{\mathrm{ijt}} + \Sigma\beta D_t + \gamma\lambda_{\mathrm{it}} + u_{it} \qquad [7]$$

α and β are unbiased, where P_{it} is the observed selling price on properties that actually sell.

Duration Dependence

There is, however, a crucial missing ingredient in Gatzlaff and Haurin's (1998) specification of the seller's reservation price in [4], the current length of stay, L_{it}. If L_{it} is included,

$$P_{it}^{R} = V_{it} + L_{it} + e^{R}{}_{it} \qquad [4]^{\#}$$

then,

$$e_{it}^{O} - e_{it}^{R} = \mathrm{f}(x^p, z_{t,,} L_{it}) \qquad [5]^{\#}$$

and,

$$\lambda_{it} = \lambda_{it}(x^p, z_{t,} L_{it}) \qquad [6]^{\#}$$

which suggests that the estimation of the hazard of non-selection also has to include the length of stay. This raises issues of duration dependence which are most appropriately dealt with using time-to-event estimation techniques such as Cox Proportional-Hazard estimation and Log-logistic regression. Applying duration analysis to the estimation of λ_{it} would therefore overcome an important potential weakness in the Gatzlaff and Haurin (1998) analysis. By applying probit rather than a duration model approach that allows for duration dependence, Gatzlaff and Haurin implicitly assume that there is no "duration dependence" in the process by which properties come onto the market. That is, the longer a property remains off the market does not in any way affect the probability of it entering the market in the next period. This is akin to saying that a person who moves into a property is equally likely to put the house on the market the following day as he is in ten years time. However, given the emotional upheaval and transactions costs associated with moving, it is highly unlikely that this will be the case.

Data

In addition to the duration dependency problem, the applicability of Gatzlaff and Haurin (1998) approach to the UK context is limited since comprehensive data on unsold properties are rarely available. (Note that the Gatzalff and Haurin data is "limited to single-family detached homes with between 600 and 6000 square feet of living area and less than five acres of land" op cit, p.209, which presumably precludes the potentially important variation in frequency of sale and house-price inflation due to variation in property type). An alternative approach would be to make use of the duration of stay information that could potentially be gleaned from Scottish Land Registry data ("SASINES"). If fifteen to twenty years of Land Registry data can be compiled for a region, while information on properties that do not sell at all would not be available, we would be able to examine the nature of any property that sells at least once during that period. Duration modelling techniques could then be applied to explain the length of time the property remains off the market using techniques that control for "censored" observations -- properties that sold once but currently remain off the market. Since duration until resale varies with market buoyancy, this bias could be controlled for by predicting the hazard rate for each property for a set of "controlled" market conditions. This hazard rate could then be entered into the house price equation to control for sample selection bias.

To some extent the distinction between properties that sell and those that do not is a false dichotomy, or at least an incomplete one. The real issue is frequency of sale. A window of ten years of all property transactions, for example, will not include all properties since some will not sell at all. However, it will most probably include all types of properties. Even though a there may exist a type of property that sells once in twenty years, provided this class of properties is of reasonable size, it is likely that a number of these properties will trade within the ten year window. So the ten year window should give a random sample of all levels of frequency of sale. Application of censored duration techniques should adequately control for those properties that sell only once in this period.

Evidence of Transactions Bias: A Simple Example

To illustrate the kind of biases endemic in existing price indices, consider the following table which lists the number of properties in each West of Scotland local authority that sold either once, twice, three times, four times or five or more times in the 1991 to 2000 period. The table also presents the proportion of sales in each area that fall into each of these repeat sales categories. The data are drawn from SASINES (i.e. land registry) records on the West of Scotland. There is clearly considerable variation in repeat sales even within the West of Scotland. In the City of Glasgow, for example, nearly 30% of properties transacted sold twice, and 10% sold three times. This contrasts with Argyll and Bute where less than 18% sold twice and only 3.6% sold three times. Overall, 63.3% of properties that sold came on the market only once, 25.9% sold twice, 8.4% sold three times, 1.9% sold four times and 0.5% sold five or more times. It is likely that there are similar intra and inter settlement disparities in the proportion of properties sold at all.

Bias in rates of change?

For house price indices to be distorted by frequency of sale, however, there would have to be different patterns of house price values for different rates of turnover. Is there evidence for this kind of bias in a UK context (most studies of these issues are from a US context – for example: Gatzlaff and Haurin 1994, 1997, 1998; Fisher et al 2003; Hwang and Quigley 2004)? In **tables 1 and 2** it can be seen that the mean house price tends to be lower for properties which frequently sell (categories with small samples – less than 200 observations – should excluded because of the high variation in prices and the absence of any mix-adjustment). A notable exception is the City of Glasgow which is a very heterogeneous area and likely to be biased by the West End which is a generally considered a separate submarket (it is a high value area with high turnover). Further analysis would be needed to ascertain whether the effect was caused by the heterogeneity of the City of Glasgow (we would like to know, for example, whether *within* the West End, more frequently traded properties tend to be of lower value). Nevertheless, it is clear that there is good reason to believe that house prices vary systematically by frequency of sale and that grouping all properties together without accounting for this non-randomness is likely to result in house price indices giving a biased picture of the level of prices at a given point in time. There is also evidence here to support the argument that in many areas, properties that remain off the market for long periods yield higher yields of "satisfaction" (whether due to location, size or quality) as they tend to sell for a higher price than frequently sold dwellings. This might reflect simple lifecycle patterns or it might also be the result of information asymmetries in the housing market (buyers know less than sellers about the true quality of the dwelling and the desirability of its location) and this can result in the stock of dwellings for sale at a given point in time being characterised by a disproportionate number of poor quality properties (a process called "adverse selection – see Akerlof's 1963 seminal theoretical paper on the "Market for Lemons").

Table 1 Variation in the Frequency of Sale of Properties

Glasgow Submarkets:

	Number of times a dwelling has sold in the 1991-2000 period:					
	1	**2**	**3**	**4**	**5+**	**All**
Glasgow City Centre	2,999	1,269	484	135	49	4,936
	60.8%	25.7%	9.8%	2.7%	1.0%	100.0%
Glasgow East End	7,713	3,624	1,322	377	114	13,150
	58.7%	27.6%	10.1%	2.9%	0.9%	100.0%
Glasgow North Side	2,879	1,191	491	134	41	4,736
	60.8%	25.1%	10.4%	2.8%	0.9%	100.0%
Glasgow South Side	19,905	9,685	3,835	1,010	328	34,763
	57.3%	27.9%	11.0%	2.9%	0.9%	100.0%
Glasgow West End	17,023	7,979	3,180	803	266	29,251
	58.2%	27.3%	10.9%	2.7%	0.9%	100.0%

Local Authorities in Strathclyde:

	Number of times a dwelling has sold in the 1991-2000 period:					
	1	**2**	**3**	**4**	**5+**	**All**
Argyll & Bute	14,815	3,422	687	96	73	19,093
	77.6%	17.9%	3.6%	0.5%	0.4%	100.0%
City of Glasgow	83,971	40,255	14,199	3,089	838	142,352
	59.0%	28.3%	10.0%	2.2%	0.6%	100.0%
East Ayrshire	16,136	5,396	1,509	332	35	23,408
	68.9%	23.1%	6.5%	1.4%	0.2%	100.0%
East Dunbartonshire	13,796	6,774	1,967	432	65	23,034
	59.9%	29.4%	8.5%	1.9%	0.3%	100.0%
East Renfrewshire	13,696	5,944	1,989	486	165	22,280
	61.5%	26.7%	8.9%	2.2%	0.7%	100.0%
Inverclyde	13,521	4,560	1,232	305	124	19,742
	68.5%	23.1%	6.2%	1.5%	0.6%	100.0%
North Ayrshire	21,235	6,484	1,839	352	52	29,962
	70.9%	21.6%	6.1%	1.2%	0.2%	100.0%
North Lanarkshire	41,634	16,570	5,685	1,388	349	65,626
	63.4%	25.3%	8.7%	2.1%	0.5%	100.0%
Renfrewshire	27,292	10,742	3,211	677	140	42,062
	64.9%	25.5%	7.6%	1.6%	0.3%	100.0%
South Ayrshire	18,310	6,364	1,731	328	65	26,798
	68.3%	23.8%	6.5%	1.2%	0.2%	100.0%
South Lanarkshire	41,467	18,552	6,643	1,747	744	69,153
	60.0%	26.8%	9.6%	2.5%	1.1%	100.0%
West Dumbartonshire	874	403	126	30	5	1,438
	60.8%	28.0%	8.8%	2.1%	0.4%	100.0%
West Dunbartonshire	11,015	4,398	1,278	302	53	17,046
	64.6%	25.8%	7.5%	1.8%	0.3%	100.0%
Total	317,762	129,864	42,096	9,564	2,708	501,994
	63.3%	25.9%	8.4%	1.9%	0.5%	100.0%

Table 2 Average House Prices in 1991 by no. times sold in previous 10 years

	City of Glasgow			**East Renfrewshire**			**North Ayrshire**		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	£ 37,288	£ 43,320	9192	**£ 65,173**	£ 42,210	1235	£ 34,165	£ 26,116	2399
2	£ 37,676	£ 27,346	4138	£ 61,638	£ 35,940	607	**£ 35,754**	£ 21,518	755
3	**£ 38,028**	£ 27,370	1492	£ 57,829	£ 28,508	221	£ 32,625	£ 18,211	236
4	£ 37,325	£ 18,648	343	£ 53,135	£ 24,692	54	£ 35,084	£ 29,721	39
5+	£ 38,934	£ 18,816	103	£ 38,345	£ 18,689	15	£ 22,550	£ 8,603	4
All	£ 37,477	£ 37,627	15268	£ 62,912	£ 38,900	2132	£ 34,405	£ 24,732	3433

	Renfrewshire			**East Ayrshire**			**East Dunbartonshire**		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	**£ 41,103**	£ 72,331	2895	**£ 33,254**	£ 27,616	1688	**£ 64,404**	£ 41,465	1406
2	£ 38,038	£ 22,466	1159	£ 32,814	£ 18,042	634	£ 59,634	£ 36,380	742
3	£ 37,969	£ 23,037	363	£ 33,258	£ 21,741	200	£ 55,912	£ 31,089	244
4	£ 37,008	£ 18,206	75	£ 39,025	£ 20,607	43	£ 48,358	£ 28,829	60
5+	£ 35,874	£ 22,070	17	£ 22,848	£ 6,030	3	£ 46,475	£ 58,754	6
All	£ 39,975	£ 59,502	4509	£ 33,230	£ 25,017	2568	£ 61,686	£ 38,970	2458

	South Lanarkshire			**Argyll & Bute**			**South Ayrshire**		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	**£ 38,627**	£ 34,187	4018	**£ 43,248**	£ 39,470	1618	**£ 47,515**	£ 33,222	1783
2	£ 37,164	£ 26,054	1999	£ 44,813	£ 31,904	343	£ 47,282	£ 29,145	703
3	£ 36,781	£ 22,691	782	£ 48,289	£ 33,436	81	£ 40,610	£ 20,368	210
4	£ 35,590	£ 18,257	200	£ 49,023	£ 33,865	13	£ 38,817	£ 16,339	44
5+	£ 34,908	£ 15,532	86	£ 33,197	£ 11,162	24	£ 41,729	£ 12,090	16
All	£ 37,880	£ 30,398	7085	£ 43,622	£ 37,861	2079	£ 46,757	£ 31,165	2756

	North Lanarkshire			**West Dumbartonshire**			**Inverclyde**		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	£ 30,933	£ 25,665	3807	**£ 32,372**	£ 23,967	1309	**£ 36,581**	£ 34,611	1296
2	£ 31,494	£ 18,954	1583	£ 32,760	£ 19,008	558	£ 34,371	£ 34,460	416
3	**£ 31,565**	£ 16,507	587	£ 30,861	£ 15,828	181	£ 33,627	£ 22,444	123
4	£ 30,832	£ 15,438	142	£ 36,015	£ 18,041	46	£ 31,976	£ 14,070	28
5+	£ 34,503	£ 21,817	31	£ 33,500	£ 18,053	7	£ 59,811	£ 46,267	7
All	£ 31,153	£ 23,109	6150	£ 32,428	£ 21,989	2101	£ 35,913	£ 33,762	1870

Table 3 Average House Prices in 2000 by no. times sold in previous 10years

	City of Glasgow			East Renfrewshire			North Ayrshire		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	£54,716	£ 50,744	9382	**£98,288**	£106,418	1548	**£ 48,074**	£ 43,868	2165
2	**£57,202**	£ 51,253	4991	£82,881	£ 53,830	579	£ 44,403	£ 26,711	840
3	£55,827	£ 40,150	1929	£81,285	£ 45,267	173	£ 42,366	£ 24,266	258
4	£52,912	£ 35,444	409	£64,299	£ 43,925	47	£ 39,101	£ 34,378	45
5+	£50,802	£ 35,421	107	£53,746	£ 23,727	23	£ 42,873	£ 24,966	8
Total	£55,512	£ 49,402	16818	£92,177	£ 91,529	2370	£ 46,565	£ 38,785	3316

	Renfrewshire			East Ayrshire			East Dunbartonshire		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	**£59,377**	£ 49,949	2777	**£49,320**	£167,645	2039	**£ 89,050**	£ 68,914	1403
2	£46,410	£ 29,625	1251	£42,165	£ 21,329	672	£ 81,549	£ 56,465	718
3	£46,528	£ 27,457	402	£41,306	£ 15,805	189	£ 79,464	£ 51,879	201
4	£40,986	£ 22,838	79	£39,881	£ 16,689	56	£ 62,744	£ 27,121	47
5+	£35,475	£ 21,167	16	£30,456	£ 15,055	7	£ 65,669	£ 58,961	9
Total	£54,245	£ 43,514	4525	£46,963	£139,552	2963	£ 85,366	£ 63,578	2378

	South Lanarkshire			Argyll & Bute			South Ayrshire		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	**£57,961**	£ 46,401	4867	**£70,301**	£193,571	1330	**£ 63,964**	£ 55,403	1876
2	£53,297	£ 63,628	2168	£62,922	£ 52,241	320	£ 59,780	£ 38,398	750
3	£44,992	£ 27,455	802	£67,252	£ 65,024	64	£ 50,969	£ 26,532	232
4	£45,693	£ 24,788	222	£99,245	£ 66,275	11	£ 47,442	£ 19,694	47
5+	£40,167	£ 18,165	95	£62,800	£ 49,241	7	£ 44,235	£ 9,214	7
Total	£54,904	£ 49,763	8154	£68,979	£171,676	1732	£ 61,537	£ 49,356	2912

	North Lanarkshire			West Dumbartonshire			Inverclyde		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
1	**£46,240**	£ 37,045	4939	**£46,491**	£ 30,615	1294	**£ 49,914**	£ 50,833	1521
2	£42,189	£ 24,198	2008	£42,313	£ 21,620	558	£ 48,035	£ 33,953	553
3	£39,942	£ 20,878	725	£39,408	£ 18,942	175	£ 47,349	£ 26,385	166
4	£36,475	£ 17,385	191	£40,053	£ 11,624	42	£ 37,815	£ 17,469	39
5+	£35,187	£ 17,821	41	£39,961	£ 19,237	5	£ 37,372	£ 14,939	17
Total	£44,340	£ 32,590	7904	£44,623	£ 27,388	2074	£ 48,977	£ 45,271	2296

The really crucial question, however, is whether there are different rates of house price change across the different rates of property turnover. If so, frequently sold properties could be characterised as a different submarket. In **Table 4** it can be seen that the increase in prices tends to be greater for properties that sell only once (where the sample is less than 200 in either 1991 or 2000 the figures should be treated with caution as the standard deviation of house prices is so large that very large samples are needed to give reliable estimates). As a result, for most of the local authorities listed, using the change in average of all properties under estimates the rate of growth of houseprices. Note that if a repeat sales index were used, in most Las this would result in even greater bias. In some areas the difference is enormous. In Renfrewshire, for example, the per centage increase in average prices from 1991 to 2000 was double that of either properties that sold twice or three times (44.5% compared with 22% and 22.5%), and four times that of properties sold four times (44.5% compared with 10%). An exception to the rule is again the City of Glasgow, the figures for which are probably distorted by the West End.

It could be argued that, to some extent, repeat sale bias may be mitigated by standard methods for controlling for the mix of dwellings coming onto the market (hedonic techniques, for example). For this mitigation to be effective, repeat sale patterns would have to fall along dwelling attribute lines, and the relevant attributes would have to be adequately controlled for in the mix-adjustment procedure. It is highly unlikely that existing mix adjusted indices adequately capture this effect, however, since there is likely to be a sub-city level spatial dimension to the repeat sales process and existing indices do not have this level of spatial refinement (see chapter 2 of Meen and Andrew, 1998, for a summary of how existing house price indices are computed).

Table 4 % Change in Average House Price by No. Times sold in 1999-2000 Period

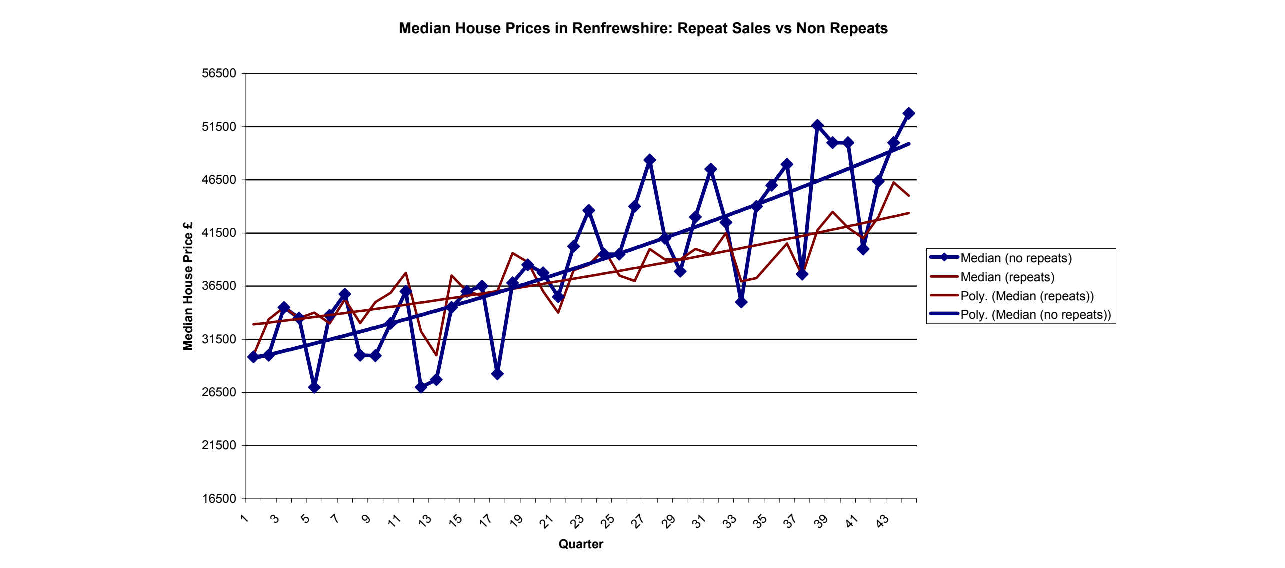
	City of Glasgow		East Renfrewshire		North Ayrshire	
	Mean	**min n**	**Mean**	**min n**	**Mean**	**min n**
1	46.7%	9192	**50.8%**	1235	**40.7%**	2165
2	**51.8%**	4138	34.5%	579	24.2%	755
3	46.8%	1492	40.6%	173	29.9%	236
4	41.8%	343	21.0%	47	11.4%	39
5+	30.5%	103	40.2%	15	90.1%	4
All	48.1%	15268	46.5%	2132	35.3%	3316

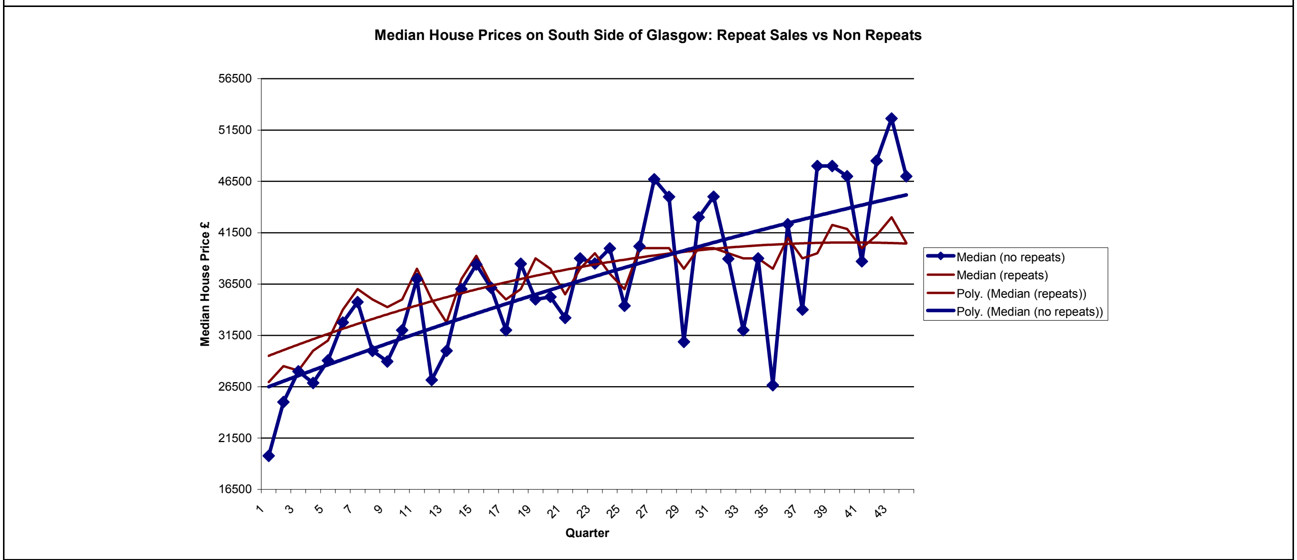
	Renfrewshire		East Ayrshire		East Dunbartonshire	
	Mean	**min n**	**Mean**	**min n**	**Mean**	**min n**
1	**44.5%**	2777	**48.3%**	1688	38.3%	1403
2	22.0%	1159	28.5%	634	36.7%	718
3	22.5%	363	24.2%	189	**42.1%**	201
4	10.7%	75	2.2%	43	29.7%	47
5+	-1.1%	16	33.3%	3	41.3%	6
All	35.7%	4509	41.3%	2568	38.4%	2378

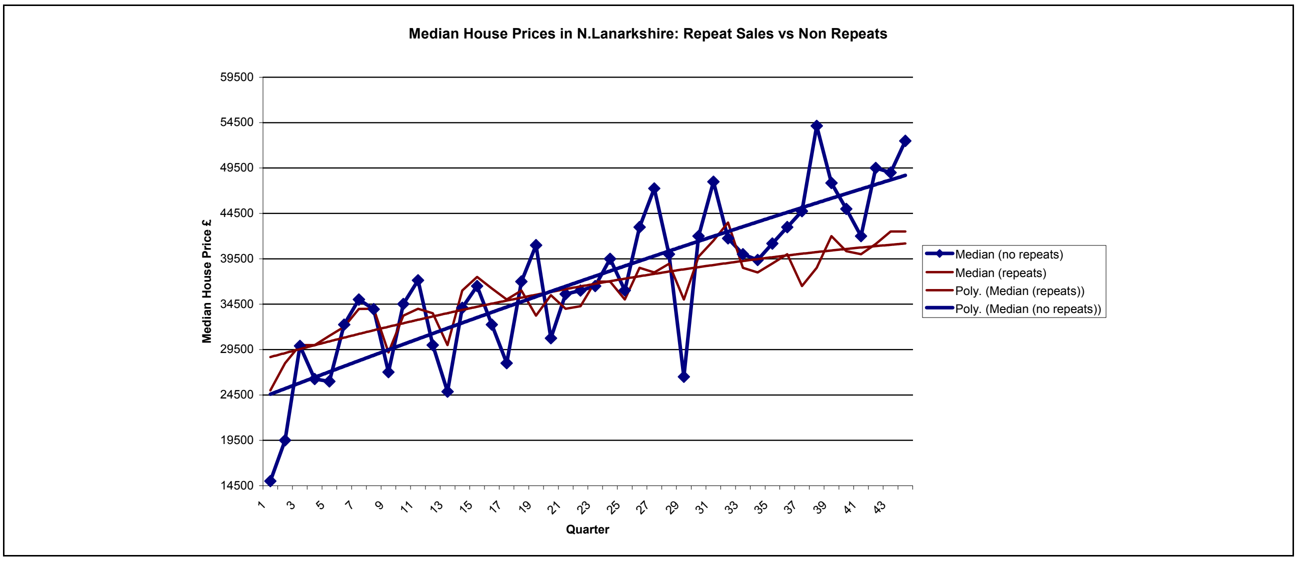
	South Lanarkshire		Argyll & Bute		South Ayrshire	
	Mean	**min n**	**Mean**	**min n**	**Mean**	**min n**
1	**50.1%**	4018	**62.6%**	1330	**34.6%**	1783
2	43.4%	1999	40.4%	320	26.4%	703
3	22.3%	782	39.3%	64	25.5%	210
4	28.4%	200	102.4%	11	22.2%	44
5+	15.1%	86	89.2%	7	6.0%	7
All	44.9%	7085	58.1%	1732	31.6%	2756

	North Lanarkshire		West Dumbartonshire		Inverclyde	
	Mean	**min n**	**Mean**	**min n**	**Mean**	**min n**
1	**49.5%**	3807	**43.6%**	1294	**36.4%**	1296
2	34.0%	1583	29.2%	558	39.8%	416
3	26.5%	587	27.7%	175	40.8%	123
4	18.3%	142	11.2%	42	18.3%	28
5+	2.0%	31	19.3%	5	-37.5%	7
All	42.3%	6150	37.6%	2074	36.4%	1870

Figure 1 Comparison of Repeats and Non-Repeats For Different Submarkets



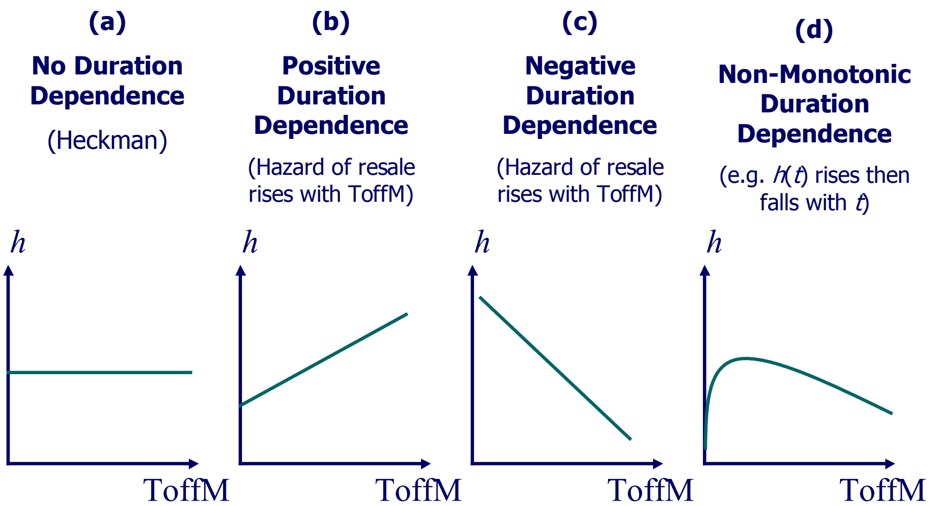




Econometric Model: Evidence of Duration Dependence

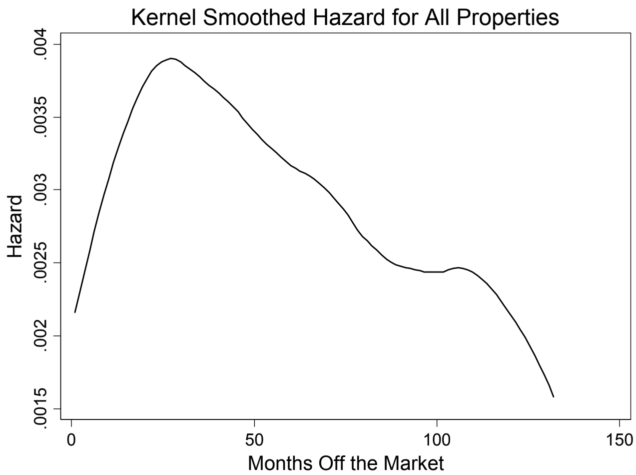
In order to choose the appropriate form of sample selection correction, we need to ascertain whether there exists some degree of duration dependence. If there is zero duration dependence then the probability of a property coming onto the market estimated can reliably estimated from a simple probit regression (as in the Heckman two step approach adopted by Gatzlaff and Haurin). If this were the case, we would expect the hazard of a property coming onto the market to remain unaffected by how long it had been off the market. This would result in a horizontal hazard function as depicted in panel (a) of Figure **2**. If, however, the hazard of a property coming onto the market increases the longer remains off the market, then we would expect the hazard curve to be upward sloping, as represented in panel (b) of Figure 1. The converse is depicted in panel (c). Finally, the hazard curve may be non-linear, rising (falling) initially, reaching a zenith (trough) and then declining (increasing) or undulating in a regular or irregular pattern. Panel (d) depicts a simple hill-like hazard function.

Figure 2 Hazard Function Types for Different Levels of Duration Dependence



Based on over quarter of a million transactions in the 1990-2000 period the kernel-smoothed hazard curve drawn in Figure 3 is constructed assuming no variation in hazard rates between properties of different types or location. The curve is clearly both non-linear and non-monotonic suggesting a high degree of duration dependence.

Figure 3



Semi-Parametric Multiple Regression Estimates

To arrive at a more robust confirmation of the existence of duration dependency, we need to control for other factors. For example, a general market decline in turnover over the period or the collapse of a high turnover area in the second half of the 1990s might be the cause of the fall-off in the hazard function. Cross-sectional variation might occur because certain areas are less attractive and/or have thinner markets, or because some areas attract an older population. So in the regressions that follow we include:

- the **deprivation score** of an area and,
- the **proportion of the population** that are over 54.
- **Population** changes and,
- cycles in **unemployment** are both factors that might vary both over time and across space.
- Also the number of **newbuild** is a potentially important determinant of the probability of a property coming onto the market.
- Also, the **distance travelled by buyers** might be a factor given that areas that predominantly attract buyers from well outside the area will be characterised by households that have few local ties to discourage resale.

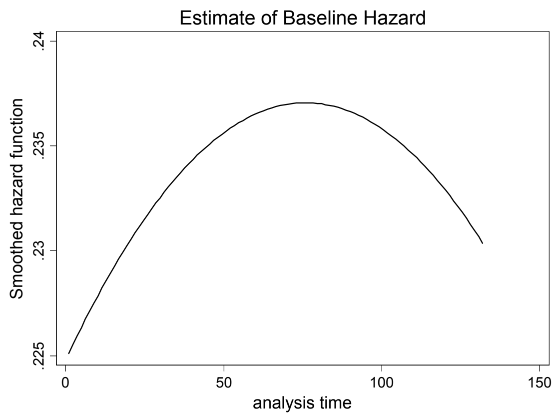
The following is a proportional hazards Cox regression that takes into account the possibility of "repeat failure" (i.e. a property repeatedly coming onto the market). It can be seen that higher deprivation scores are associated with lower hazard rates, and that an increase in the claimant count has a similar (but much greater) effect. Also, the more newbuild in an area, the less likely a property is to come onto the market. Factors that increase the hazard of entering the market include the distance travelled by the buyer and positive changes in the population of a district.

No. of subjects =	270695	Number of obs =	286,575
No. of failures =	62445		
Time at risk =	17213106		
		LR chi2(16) =	317750.24
Log likelihood =	-596064.56	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deprivation	.9948036	.0015893	-3.26	0.001	.9916935	.9979233
Δ unemp	.4306751	.0196333	-18.48	0.000	.3938635	.4709272
buyer_distance	1.038548	.0024437	16.07	0.000	1.033769	1.043349
newbuild	.9991922	.0001364	-5.92	0.000	.998925	.9994595
prop_gt54	.0155844	.0040706	-15.93	0.000	.0093403	.0260027
Δ pop	825.7914	992.6918	5.59	0.000	78.27674	8711.801

Estimating this regression also allows us to derive the *baseline hazard* (a way of examining the shape of the hazard function holding other factors constant), plotted in Figure 4 below. This clearly shows a concave and highly duration dependent hazard function.

Figure 4



Although this curve is estimated using non-parametric kernel density estimation, it shows a remarkably regular quadratic shape. Running a quadratic OLS regression of the estimated baseline hazard against analysis time (time off the market) confirms this as the following results show (Adj R-squared = 0.6521):

haz2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Analysis time	.0275415	.0000922	298.72	0.000000	.0273608	.0277222
Analysis time^2	-.0002018	8.75e-07	-230.70	0.000000	-.0002035	-.0002001
Constant	.2300911	.0018797	122.41	0.000000	.2264069	.2337752

For the hazard of entering the market to be non-duration dependent, the coefficients on both the linear and quadratic analysis time variable would have to equal zero and the t tests report that we can reject both these null hypotheses with infinitesimal chance of false rejection (note how narrow the confidence intervals are for both coefficients).

Parametric Estimation:

Further confirmation of the duration dependency of the hazard function is found from considering parametric estimation. Gompertz, Weibull and Log-logistic regressions were run with the results and interpretation summarised below (regression tables follow the summary). The overwhelming evidence for duration dependence presented here precludes simple probit estimation of the probability of a property entering the market since such an approach has to assume that the length of time a property has been off the market has no bearing on whether it will reenter the market in the next period. This means that a property that has just been purchased has as much chance of entering the market again as a property that has been off the market for several years. Clearly, this is not the case and so a more robust approach to correcting the sample selectivity bias intrinsic to transactions based hedonic estimation is needed. That is, the estimation of the hazard of selection (or non-selection) needs to allow for duration dependency.

Gompertz Distribution:

if gamma > 0 then positive duration dependence;
if gamma = zero then no duration dependence;
if gamma < 0 then negative duration dependence.

Estimated gamma: CI 95% (.3046027, .3086786**) Clearly greater than zero**
=> duration dependence

Weibull Distribution:

if p > 1 then positive duration dependence;
if p = one then no duration dependence;
if p < 0 then negative duration dependence.

Estimated p: CI 95% 4.735841, 4.800477**) Clearly greater than one**
=> duration dependence

Log-logistic Distribution:

if gamma = 1 then the hazard is monotonic and negative duration dependence

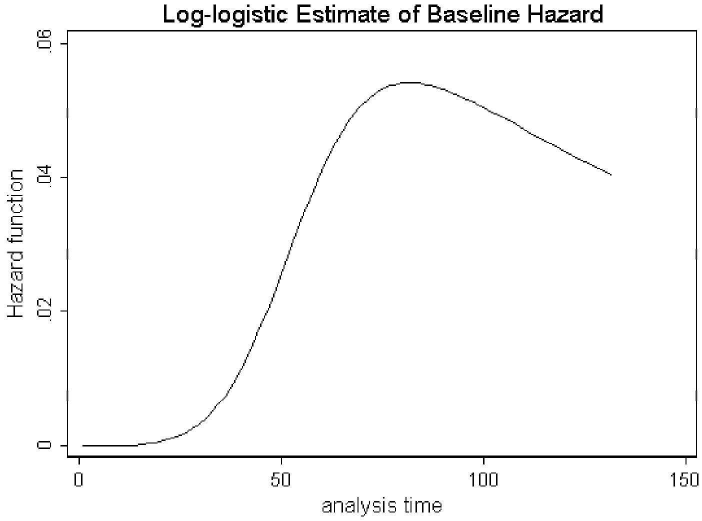
if 0.5 < gamma < 1 then the hazard rises steeply but declines shallowly indicating highly positive duration dependence at the outset, gradually becoming slightly negative duration dependent.

if gamma < 0.25 then the hazard initially rises but declines steeply indicating gradually increasing duration dependence, which at some point rapidly becomes highly negatively duration dependent.

Estimated gamma: CI 95% (.183254 .1857785**) Clearly less than one**
=> non-monotonic duration dependence

All of these results confirm duration dependency and the log-logistic estimation confirms the non-monotonic nature of the hazard function (the baseline hazard from which is plotted in Figure 5 below).

Figure 5




```
Gompertz regression -- log relative-hazard form
No. of subjects =       270695                    Number of obs   =    286,575
No. of failures =        62445
Time at risk    =     17213106
                                                  LR chi2(16)     = 324595.53
Log likelihood  =    -51570.394                   Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deprivation	.9946122	.0015796	-3.40	0.001	.991521	.997713
Δ unemp	.3710097	.0167684	-21.94	0.000	.3395578	.4053748
buyer_dist	1.039365	.0020061	20.00	0.000	1.035441	1.043305
newbuild	.9991295	.0001353	-6.43	0.000	.9988642	.9993947
prop_gt54	.0150625	.0039341	-16.06	0.000	.0090277	.0251314
Δ pop	1343.006	1614.87	5.99	0.000	127.2239	14177.1
yr1991_d	.0252044	.0005183	-178.99	0.000	.0242087	.026241
yr1992_d	.0005459	.0000156	-262.63	0.000	.0005161	.0005774
yr1993_d	9.78e-06	3.89e-07	-290.31	0.000	9.05e-06	.0000106
yr1994_d	1.88e-07	9.73e-09	-299.85	0.000	1.70e-07	2.08e-07
yr1995_d	3.33e-09	2.14e-10	-303.21	0.000	2.93e-09	3.77e-09
yr1996_d	8.22e-11	6.32e-12	-301.73	0.000	7.07e-11	9.55e-11
yr1997_d	1.02e-12	9.24e-14	-305.54	0.000	8.57e-13	1.22e-12
yr1998_d	1.47e-14	1.54e-15	-302.44	0.000	1.19e-14	1.80e-14
yr1999_d	4.09e-16	4.77e-17	-303.88	0.000	3.25e-16	5.14e-16
yr2000_d	4.44e-18	5.95e-19	-298.45	0.000	3.42e-18	5.77e-18
gamma	**.3066406**	**.0010398**	**294.91**	**0.000**	**.3046027**	**.3086786**

Weibull regression -- log relative-hazard form

No. of subjects = 270695 Number of obs = 286,575
No. of failures = 62445
Time at risk = 17213106
LR chi2(16) = 249620.73
Log likelihood = -89724.535 Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deprivtn	.9959746	.0015824	-2.54	0.011	.9928781	.9990809
Δ unemp	.1728571	.0078458	-38.67	0.000	.1581437	.1889394
buyer_dist~n	1.037106	.0019985	18.91	0.000	1.033197	1.041031
newbuild	1.000612	.0001332	4.60	0.000	1.000351	1.000873
prop_gt54	.0155921	.0040743	-15.92	0.000	.0093429	.0260212
Δ pop	753.0472	906.0247	5.51	0.000	71.23667	7960.508
yr1991_d	.0256331	.0005078	-184.96	0.000	.024657	.0266478
yr1992_d	.0022976	.0000553	-252.59	0.000	.0021918	.0024085
yr1993_d	.0003595	.0000101	-281.78	0.000	.0003402	.0003798
yr1994_d	.0000768	2.45e-06	-297.48	0.000	.0000721	.0000817
yr1995_d	.00002	7.20e-07	-299.92	0.000	.0000186	.0000214
yr1996_d	1.00e-05	3.92e-07	-293.73	0.000	9.26e-06	.0000108
yr1997_d	2.45e-06	1.06e-07	-298.66	0.000	2.25e-06	2.67e-06
yr1998_d	1.16e-06	5.55e-08	-285.44	0.000	1.06e-06	1.27e-06
yr1999_d	5.22e-07	2.71e-08	-278.78	0.000	4.71e-07	5.78e-07
yr2000_d	1.22e-07	7.92e-09	-244.40	0.000	1.07e-07	1.38e-07
/ln_p	1.561937	.0034582	451.66	0.000	1.555159	1.568715
p	**4.76805**	**.0164888**			**4.735841**	**4.800477**
1/p	.2097294	.0007253			.2083126	.2111557

Log-logistic **regression** -- accelerated failure-time form

No. of subjects = 270695 Number of obs = 286,575
No. of failures = 62445
Time at risk = 17213106
LR chi2(16) = 247246.05
Log likelihood = -90459.799 Prob > chi2 = 0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
deprivtn	.0008969	.0003591	2.50	0.013	.000193	.0016007
Δ unemp	.3825159	.0099812	38.32	0.000	.3629531	.4020786
buyer_dist	-.0083315	.0004451	-18.72	0.000	-.0092039	-.0074591
newbuild	-.0000407	.0000296	-1.38	0.169	-.0000987	.0000173
prop_gt54	.8925794	.0586084	15.23	0.000	.777709	1.00745
Δ pop	-1.367138	.2708366	-5.05	0.000	-1.897968	-.8363082
yr1991_d	.8763299	.0040902	214.25	0.000	.8683133	.8843465
yr1992_d	1.39071	.0039356	353.37	0.000	1.382997	1.398424
yr1993_d	1.786662	.0036836	485.02	0.000	1.779442	1.793881
yr1994_d	2.110045	.0036691	575.08	0.000	2.102854	2.117237
yr1995_d	2.387568	.0041142	580.32	0.000	2.379504	2.395632
yr1996_d	2.528824	.0045165	559.91	0.000	2.519972	2.537677
yr1997_d	2.812375	.0050296	559.16	0.000	2.802517	2.822233
yr1998_d	2.960935	.0059856	494.67	0.000	2.949204	2.972667
yr1999_d	3.121442	.0067735	460.83	0.000	3.108167	3.134718
yr2000_d	3.402063	.0098221	346.37	0.000	3.382812	3.421314
_cons	2.045325	.0157124	130.17	0.000	2.014529	2.076121
/ln_gam	-1.690041	.0034905	-484.19	0.000	-1.696882	-1.6832
gamma	**.1845119**	**.000644**			**.183254**	**.1857785**

Omitted Variable Bias in Submarket Hedonic Estimation

To illustrate how the estimated hazard of resale could be incorporated into submarket hedonic regressions the following regresses the log of selling price against distance to CBD, the estimated hazard of sale for each property transaction, the interaction of this hazard with year dummies and stand alone year dummies.

Looking at the results it can be seen that the stand alone hazard variable has statistically significant positive effect on house prices in the West End submarket (interpreted as capturing the cyclical variation in market buoyancy), though when interacted with year dummies, it can be seen that the house price is generally lower for higher the hazard rates (i.e. for a given period, cross sectional variation in the hazard will reflect quality differences in properties/neighbourhoods and so higher hazards of resale will correspond to lower quality). The next stage in our analysis will be to bootstrap these hedonic regressions since it is not clear whether Heckman's analytical formulae for computing the appropriate standard errors still applies when the hazard of selection is used (rather than the hazard of non-selection) and when other violations apply (see Greene 2003).

West End: SASINES

Source	SS	df	MS		
				Number of obs =	2539
				F(22, 2516) =	7.35
Model	82.4536294	22	3.74789225	Prob > F =	0.0000
Residual	1282.29594	2516	.509656575	R-squared =	0.0604
				Adj R-squared =	0.0522

price_ln	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cbd_glas_ln	-.1196089	.023692	-5.05	0.000	-.1660669	-.073151
haz3_ln	2.520464	1.175102	2.14	0.032	.216199	4.824729
haz31991	-2.910495	1.439026	-2.02	0.043	-5.732292	-.0886989
haz31992	-1.110732	1.422626	-0.78	0.435	-3.90037	1.678907
haz31993	-1.985175	1.31037	-1.51	0.130	-4.554689	.5843385
haz31994	-1.950866	1.641273	-1.19	0.235	-5.169251	1.267518
haz31995	-1.509138	1.294221	-1.17	0.244	-4.046985	1.02871
haz31996	-2.476553	1.797042	-1.38	0.168	-6.000385	1.047279
haz31997	-2.024863	1.355783	-1.49	0.135	-4.683429	.6337025
haz31998	-1.680718	1.274523	-1.32	0.187	-4.17994	.8185034
haz31999	.1051627	1.58252	0.07	0.947	-2.998012	3.208337
haz32000	-2.390052	3.319275	-0.72	0.472	-8.898842	4.118738
yr1991_d	-.0096691	.1573177	-0.06	0.951	-.3181545	.2988162
yr1992_d	.1641667	.1529758	1.07	0.283	-.1358046	.464138
yr1993_d	.1666161	.1467194	1.14	0.256	-.1210871	.4543193
yr1994_d	.0131218	.1847797	0.07	0.943	-.349214	.3754576
yr1995_d	.1169831	.1540642	0.76	0.448	-.1851225	.4190887
yr1996_d	.1480436	.2271113	0.65	0.515	-.2973007	.5933878
yr1997_d	.2140097	.1767039	1.21	0.226	-.1324903	.5605096
yr1998_d	.2811318	.1649122	1.70	0.088	-.0422458	.6045094
yr1999_d	.7411036	.2085149	3.55	0.000	.3322251	1.149982
yr2000_d	.4650536	.232966	2.00	0.046	.0082288	.9218784
_cons	11.51992	.2584922	44.57	0.000	11.01304	12.0268

```
Strathclyde: SASINES
      Source |       SS       df       MS              Number of obs =   19746
-------------+------------------------------           F( 40, 19705) =   19.53
       Model |  401.441724    40  10.0360431           Prob > F      =  0.0000
    Residual |  10127.9887 19705   .51398065           R-squared     =  0.0381
-------------+------------------------------           Adj R-squared =  0.0362
    price_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 cbd_glas_ln |  -.1102142    .008715   -12.65   0.000    -.1272963    -.093132
     haz3_ln |   1.410915   .4006454     3.52   0.000     .6256165    2.196214
    haz31991 |  -.2317122   .4797832    -0.48   0.629    -1.172128    .7087034
    haz31992 |    .199651   .4809911     0.42   0.678     -.743132    1.142434
    haz31993 |  -.6540363   .4480152    -1.46   0.144    -1.532184    .2241113
    haz31994 |   .1832509   .5732268     0.32   0.749    -.9403219    1.306824
    haz31995 |   -.873003   .4449185    -1.96   0.050    -1.745081   -.0009252
    haz31996 |  -.6295961   .5926881    -1.06   0.288    -1.791315    .5321226
    haz31997 |   -1.58916   .4386051    -3.62   0.000    -2.448863   -.7294573
    haz31998 |  -.9623314   .4241663    -2.27   0.023    -1.793733   -.1309298
    haz31999 |  -.5294323   .5816505    -0.91   0.363    -1.669516    .6106517
    haz32000 |   .6022788   1.445904     0.42   0.677    -2.231816    3.436373
   yr1991_d |   .1785661    .053211     3.36   0.001      .074268    .2828642
   yr1992_d |   .1884636   .0519785     3.63   0.000     .0865814    .2903457
   yr1993_d |   .1584145   .0501274     3.16   0.002     .0601606    .2566684
   yr1994_d |     .18224   .0644637     2.83   0.005     .0558857    .3085943
   yr1995_d |    .089394   .0542609     1.65   0.099    -.0169619    .1957499
   yr1996_d |    .198499   .0781477     2.54   0.011     .0453229     .351675
   yr1997_d |   .1194774   .0566712     2.11   0.035     .0083972    .2305577
   yr1998_d |   .3018836   .0554045     5.45   0.000     .1932861    .4104811
   yr1999_d |   .4492976   .0789457     5.69   0.000     .2945572    .6040379
   yr2000_d |   .4958358    .099402     4.99   0.000     .3009995    .6906721
+Area Dummies|
       _cons |   11.34761   .0971513   116.80   0.000     11.15719    11.53804
```

The findings for the West End are generally confirmed for when the regression is run on the whole of Strathclyde using area dummies to capture spatial variation. Neither of these regressions include dwelling attribute variables (hence the very low adjusted R^2) since the SASINES data do not record such information. However, the hazard regressions could be used to predict hazard rates for other data sets and result in a more robust estimation of constant quality price.

Variations in Price Trends and Repeat Sales across submarkets

The foregoing analysis has practical implications for the way house price indices are calculated but also has some bearing on the theoretical conception and definition of submarkets. I have already hinted at the possibility of treating high turnover areas as different submarkets to low turnover areas. An empirical problem emerges, however, as to how to distinguish *intrinsically* different rates of turnover across areas from *temporary/cyclical* imbalances (e.g. one area might face a rise in unemployment due to a factory closure; or a rise in new build – these factors in themselves do not characterise differences in submarkets but may induce differences in turnover rates). A more sophisticated approach would be to distinguish submarkets by their different baseline hazard functions. Following on from the structural break approach used in the "law of one price" definition of submarkets using (this entails testing for structural breaks in hedonic price equations – see chapter 2 of Pryce 2004 for a critical review) one possibility would be to test for structural breaks in hazard regression models (using either the Cox or Log-logistic approach since these capture the non-monotonicity of the hazard of resale). A grid search approach could be used of the form suggested by Pryce 2004, where, rather than a single point of structural break being tested for, tests are run at each point in the sample. A spatial surface plot of the probability of structural break is derived to see where the breaks are most likely to occur.

Movements in Adjusted Prices

There are also submarket implications for the sample selection corrected price indices that emerge out of the duration analysis. Theil (1954) specifies three conditions for aggregation – that is, the conditions under which units might be legitimately combined across space into a single unit of analysis. Parameters have to be first homogenous; second, dependent; and third, convergent within the area of aggregation. These requirements are closely related to the dynamic aspects of the question discussed in earlier in the report regarding the means by which submarkets should be defined and demarcated. Meen (1996) makes use of these conditions to examine the unity of housing markets at the regional level but they might legitimately be applied to the sub-city segmentation of the housing market, though there has been very little UK work done at this level.

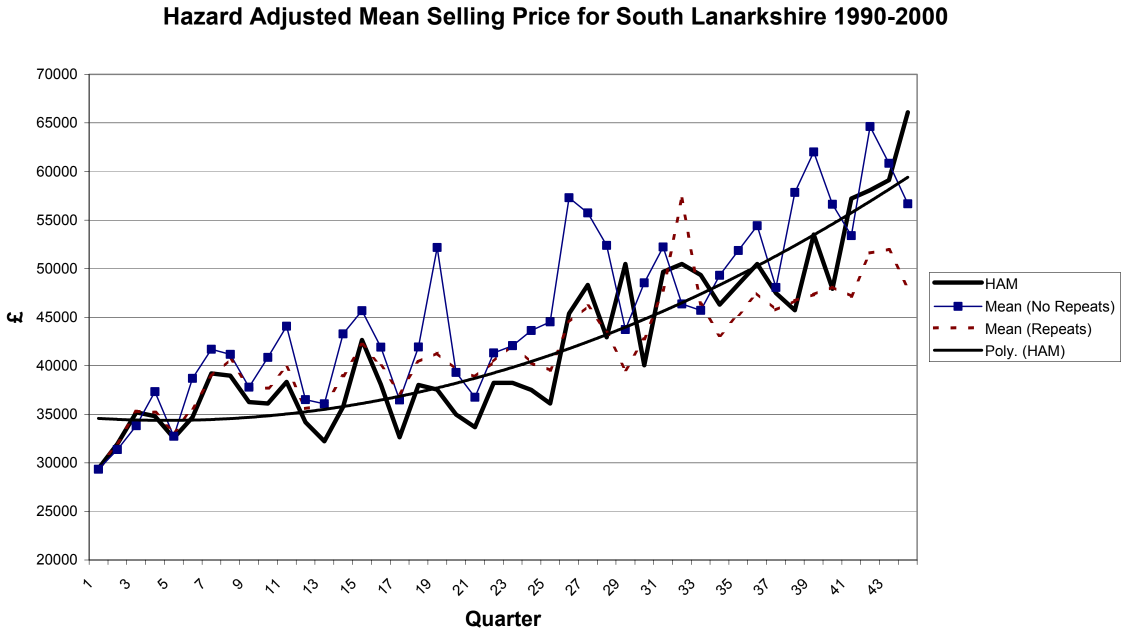
One aspect of Theil's conditions that has been explored at the submarket level is that of dependence. Jones et al (2003) construct repeat sales indices for six Glasgow submarkets (as defined by Watkins 2001) and apply cointegration techniques to determine whether these submarkets remain distinct overtime. The criterion for independence is the absence of a cointegrating relationship between the repeat sales indices for the different housing market segments. Two pairs of comparisons fail this test and as a result the six pre-defined submarkets collapse to four when this dynamic definition of dependence is used. Although the Jones et al (2003) paper offers many innovations, there a number of limitations to the method used. First there are problems associated with the initial delineation of submarkets (see the critique in chapter 2 above). Second, the authors employ repeat sales indices as the basis for the analysis and the corollary of the discussion of frequency of sale earlier in this paper is that there are likely to be several problems in applying the repeat sales approach to computing price indices. First, the proportion of repeat sales is likely to vary between submarkets. Two, this variation is not random but correlated with nature and quality of the properties. Third, repeatedly sold properties will, as a result, have different average price levels than infrequently sold properties. Fourth, the rate of change of prices in the repeat sales group may well be quite different to the price inflation of properties that sell only once or not at all. These problems could be rectified, however, by applying the sample selection approach suggested above to hedonic price equations.

Unfortunately, attribute information is not available at submarket levels for long periods for most areas of the UK. The alternative proposed here is to use the estimated hazard rates to weight observations. So price indices could be derived on the basis of hazard-weighted means (**HAM = hazard adjusted mean**) for each time period. The weighted average in a given period would be:

$$\text{Hazard adjusted mean in period } t = \mathbf{HAM_t} = (\Sigma_i\, \boldsymbol{h_t}\text{.}\ \mathbf{price}_{it}) / \Sigma_i\, \boldsymbol{h_{it}}$$

An example of this approach is depicted in Figure 6 where the hazard adjusted mean selling price for South Lanarkshire is compared with the Repeats and Non-Repeats means for each quarter in the period 1990-2000 (though there is a question here as to whether a "constant hazard" should be predicted rather than the simple predicted hazard for each observation).

Figure 6



An important element of the next phase of the research will be to compare HAM with a selection- adjusted hedonic price index. If HAM compares favourably it could be a more useful approach in the UK context since it does not rely on attribute information and could therefore be more easily applied to other areas of the country.

Avenues for Further Research

By controlling for variation in repeat sales (using duration analysis as described) it may be possible to use this adjustment to develop consistent indices for each submarket within a city. The cointegration tests used by Jones et al would then give more meaningful results. Movements in the submarket boundaries should also be investigated since tests for Theil's dependence requirement become meaningless if there are shifts in the homogeneity conditions. A practical solution in future analysis would be to run the grid search procedure suggested in **chapter two** of Pryce (2004) on selected years of the data to verify that submarket boundaries have not shifted. Where they have, it may be possible to identify irreducible cores of each submarket that remain distinct over time and it would be for these submarket cores that adjusted price indices could be designed and compared.

The Role of Credit in Determining Volatility

A potentially important factor in shaping the dynamics of local housing markets is the nature and availability of credit, and in particular, spatial variations in the structure of mortgage finance. Since the great majority of house purchases require a mortgage of some type (variations in the proportion of non-mortgage transactions is itself of interest), differences in LTVs (loan to value ratios) and other mortgage characteristics can potentially cause significant asymmetries in the impact of interest rate changes and can themselves influence local housing market dynamics. Stein (1995), for example, develops a model of trade in the housing market where purchasers require a down-payment to purchase a new home. Stein shows how variations in loan to value ratios can influence the volatility of house prices and time-to-sale of properties. Lamont and Stein (1999) investigate the Stein (1995) hypotheses using US city-level data on the relationship between homeowner borrowing patterns and house-price appreciation rates. They find that in cities where a greater fraction of homeowners have high loan-to-value ratios do indeed appear to have house prices that react more sensitively to changes in income. No research has been done in the UK to date on this issue.

A related topic also of interest is that of the impact of differences in the average loan to value ratios at the local level on the time-to-sale of owner occupied housing. Work by Stein (1995), for example, suggests that households with high loan to value ratios need to make a larger capital gain and they tend to put their dwelling on the market at a higher asking price and as a result are likely to face longer times-to-sale. Since loan to value ratios are related to income, a closely allied question is that of whether lower income groups face greater capital loss and time-to-sale risks.

Added to this complex mix is the effect of future increases local supply on time to sale, and the corollary for local possession rates. For example, if an expansion in housing supply increases the time to sale in an area, this might eliminate one of the possible escape routes for mortgage borrowers facing repayment difficulties. As such, homeowners that would otherwise have been confident of rapidly selling their home (with a view to downsizing or switching to rental accommodation) in the event of mortgage repayment difficulties, in a supply-rich world may actually find themselves unable to sell rapidly and hence facing possession (and possible homelessness) or long term reliance on Income Support for Mortgage Interest. That such households may already be more likely to have high loan to value ratios and (according to Stein) higher reservation prices as a result, the local effect of new supply may be to compound the repossession risks facing vulnerable households. Thus submarkets may not only be defined in terms of product mix or attribute valuation shifts, but also in terms of credit market behaviour, which in turn impinges upon the shape of housing demand. Different credit rationing patterns across space will influence the spatial contours of demand elasticities, which in turn may determine the impact of new supply.

There are also ways in which the structure and nature of submarkets might be expected to shape lending policy. For example, if there are significant differences in price volatility and liquidity (see below) then the corresponding variation in risk across submarkets should, in an efficient market for credit, be reflected in variations in the risk premiums embodied in mortgage rates. Controlling for spatial variations in loan to value ratios, one would expect there to be variation in mortgage rates across submarkets according to risk. On the other hand, if there are adverse selection consequences of risk pricing, as anticipated by Pryce (2003), then such a pattern may not emerge. Risk pricing by submarket, if it exists, would of course have sociological implications if those submarkets facing the highest risk and highest mortgage rates were predominantly occupied by low income families (which might well be the case if the higher price volatility is indeed associated with higher loan to value ratios – one would expect low income groups to have higher debt gearing[3] – and if liquidity is lower in deprived areas[4]).

Duration of Stay and Liquidity Bias

I have so far referred to time-off-the-market as being synonymous with duration of stay. This assumes, of course, that the time taken to sell a property is negligible, or a fixed constant for all dwellings. This, of course, is unlikely to be the case. We can, in fact, decompose duration of stay into these two phases:

Duration of stay = ToffM + TOM

Put another way, we need to consider not only "time off the market" bias but also the "time to sale" bias in house price indices. Strictly speaking, therefore, to control for the sample selection bias intrinsic to transaction based price indices, we should combine these two durations and speak of "time to resale", since it is only properties that are sold (rather than those that simply enter the market) that are included in databases of transactions.

A further question of interest is the extent to which time-off-the-market is independent of time-on-the-market. If, for example, TOM is in fact endogenous,

$$\text{TOM} = f(\text{ToffM})$$

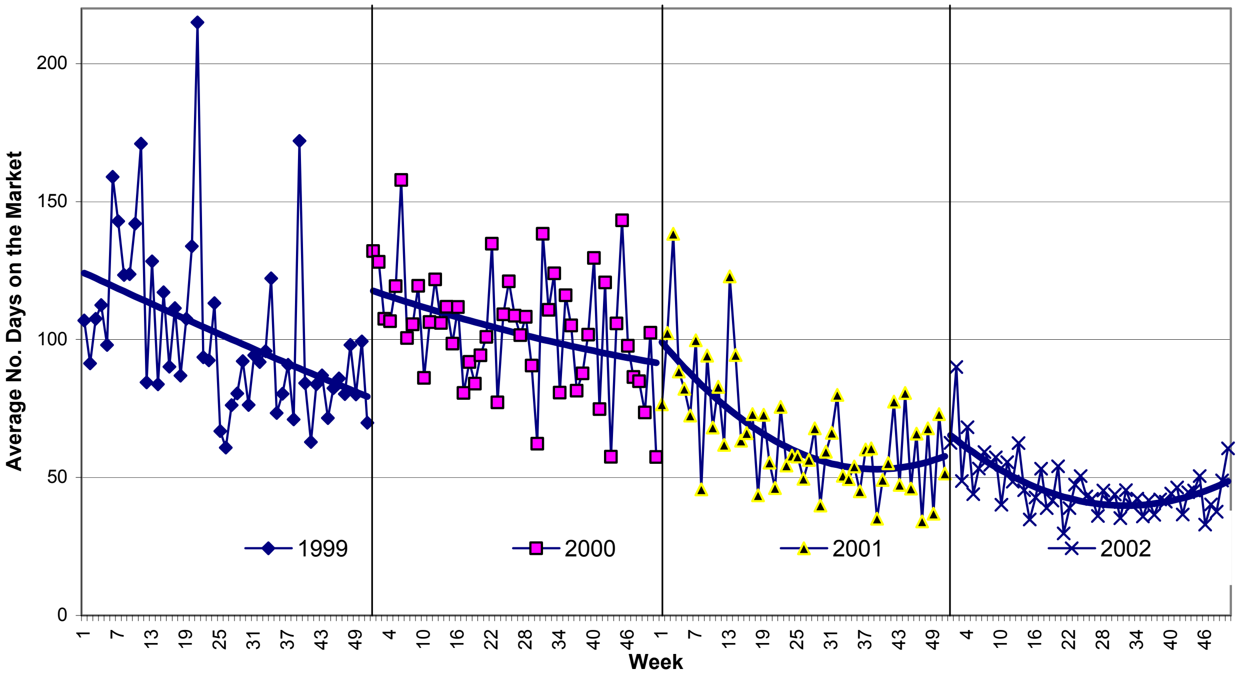
then we can further decompose duration of stay into a reduced form equation made up of ToffM and the additional determinants of TOM.

[3] Hendershott, Pryce and White (2003) find LTVs in the UK to be negatively correlated with income.

[4] See Pryce and Gibb (2004a,b).

To illustrate the extent to which the probability of a property selling varies even within a city, analysis from Gibb and Pryce (2004a) based on Glasgow transactions data is summarised below. Consider first how liquidity varies over the housing market cycle. In Figure 1 it can be seen that there is a seasonal effect on liquidity, with properties taking less time to sell in quarter 3. This seasonal effect is dominated, however, by cyclical movements in the housing market, to the extent that as the market accelerates in 1999 and 2000, quarter 4 properties actually sell more quickly than quarter 3 properties. In 2002, as the boom approaches its zenith, the seasonal "U" shape in time on the market becomes more apparent, and the overall level of time-to-sale is at its lowest.

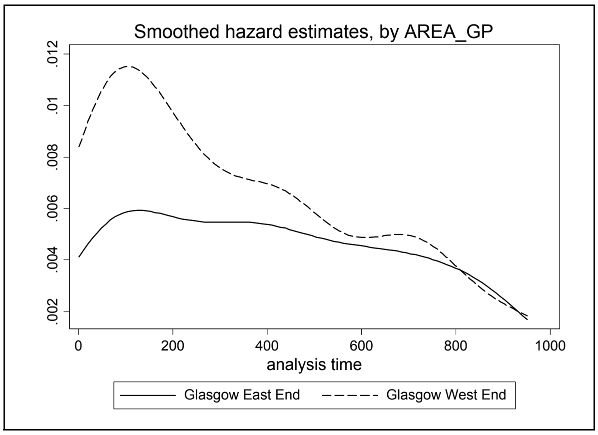
Figure 7 Average Time on the Market in Glasgow Over Time



Submarket Variations In Liquidity And The Duration Dependence Issue

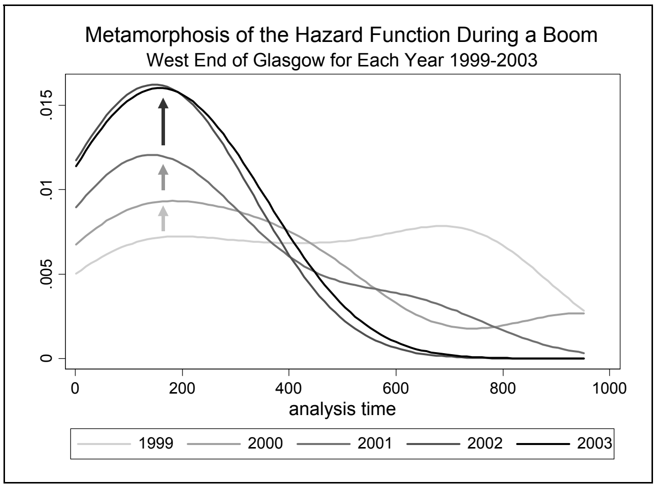
An under-researched element in the time-to-sale literature is that of variations between submarkets. One would expect that large, buoyant submarkets with efficient dissemination on properties for sale and good transport facilities available to house-searchers would enjoy greater levels of liquidity (shorter time on the market) than niche properties in less desirable markets with an inefficient estate agency sector and poor transport facilites/remote location. Pryce and Gibb (2005) find significant non-proportional shifts in the hazard function for TOM across submarkets and over time (see Figures 8 and 9 below). An interesting question, therefore, is whether similar shifts will be evident in the hazard function for ToffM.

Figure 8 Comparison of Hazard Functions for Different Submarkets



Source: Pryce and Gibb (2005)

Figure 9 Changes in the Hazard Function Over Time in the West End of Glasgow



Source: Pryce and Gibb (2005)

The most obvious implication of the variations in time on the market over time and across submarkets is that it suggests that TOM should be included as a determinant of λ:

$$\lambda_{it} = \lambda_{it}(x^p, z_t, L_{it}, TOM_{it}) \qquad [6]^{\#}$$

where TOM_{it} is the average time on the market in the submarket to which property i belongs.

Using ToffM to Define Submarkets:

An alternative track to the one taken here would be to conceive of long-stay vs short-stay properties as different submarkets. In other words, the property types and locations are sufficiently different in the eyes of purchasers that they cannot be conceived of as close substitutes. If so, it might be more appropriate to estimate separate price regressions for long-stay vs short-stay properties. As such, the duration analysis described above would be used to categorize properties into those that frequently sell, and those that infrequently sell. Grid-search procedures could be used to test for "structural breaks" (i.e. shifts) in the house price parameters. Perhaps more usefully, different rates of house price change could be investigated amongst properties of different rates of turnover. It might be, for example, that properties that sell infrequently experience a higher rate of growth than those that frequently come onto the market.