Note: This is a revised version of Working Paper 110 which appeared in *The Economic and Social Review*, Vol. 30, No. 4, October, 1999, pp. 403-423

Irish House Price Indices - Methodological Issues

DENIS CONNIFFE and DAVID DUFFY
The Economic and Social Research Institute, Dublin

Abstract: This paper reviews the international literature on house price indices, looking at the basic methodologies employed, the variables chosen for the indices and the functional forms used. Ideas are sought for future improvements to an Irish index and tentative assessments are made of how Irish results correspond to published research findings. Although the recent vintage of hedonic house price indices in Ireland warrants caution in reaching conclusions, there are interesting correspondences with published results. Some issues are identified that will require resolution to achieve further progress.

I INTRODUCTION

It is probably unnecessary to remark that the housing market is currently a topic of considerable interest. The reason for this interest and the importance of the housing market at present is partly illustrated by the following: The number of private houses built per year has increased from 19,300 in 1993 to an estimated 42,350 in 1998; House mortgage finance now accounts for approximately 30 per cent of loans provided by credit institutions and spending on housing and household equipment and operation accounted for 17.7 per cent of personal consumption of goods and services at constant market prices in 1997.

Monitoring the evolution of house prices over time is obviously a necessary requirement if investigating the behaviour of housing markets. But the compilation of an adequate measure of house price changes is not at all a trivial task. The year 1998 saw the first publication of two new series relating to Irish house price indices, one by the Irish Permanent and another by First Active. The ESRI assists with the compilation of the former series (Baker and Duffy, 1998) and so, where data are employed in this paper, they are drawn from that source. Unlike the long-standing series on average house prices maintained by the Department of the Environment, the new indices standardise for characteristics of the houses, so that comparisons over time are, or ought to be, of like with like. Similar indices have existed for many years in some other countries and a considerable literature exists on aspects of the methodology of index construction, although there are still many live issues.

The need for the standardisation, or weighting, incorporated in house price indices arises partly from a general problem with any overall average price, but also from a more specific issue with houses. The general problem is that familiar for the CPI, but true for any category of expenditure based on more than a single homogenous commodity. Even if all commodity prices remained unchanged over a time period, the average price for the category would change if the mix of commodities purchased did. The solution is to choose a fixed mix (often that of the base time period, giving a Laspeyres index) and to price that. But although houses can be sub-classified (for example, into apartments, bungalows, etc.), the difficulty is that this does not result in homogeneous price groups. A house price can depend on a variety of quantitative and qualitative attributes including location. Changes over time need to be assessed from comparisons of prices of houses of equal "quality".

A variety of issues arise when considering the construction of an appropriate index. There is the choice of basic methodology, where considerations of data availability are at least as influential as other factors. Then there is the selection of the set of attributes to be included in the index. This obviously matters greatly, but it is not at all easy to ensure than some characteristics have been adequately taken into account. There are also the more technical (but perhaps somewhat less important) matters concerning the functional forms employed and the methods of estimation. In ideal textbook circumstances, many years of good data would have been assembled before commencing compilation of a house price index , so that all these issues could be resolved. In reality, the data available for constructing the Irish Permanent index date only from 1996 and the most sensible start-up approach was to copy, as far as was practicable, the methodology of a well established UK house price index. The Halifax Index (Fleming and Nellis, 1984), which employs the hedonic approach to be described in the next section, was the chosen model.

However, as data accumulates it will become possible to examine the plausibility of the assumptions underlying the methodology and to make comparisons with other possible approaches. It is reasonable to suppose that, over time, refinements and improvements to the index will be made, partly in the light of data and partly through adoption of ideas appearing in the international literature on house price indices. Since only one year has elapsed since compilation commenced, it would be premature to propose refinements in this paper. What the paper can do, however, is review the extensive literature, examine the relevance of various themes to Irish circumstances and, tentatively, assess how Irish data accords with

findings elsewhere.

II ALTERNATIVE METHODOLOGIES FOR HOUSE PRICE INDICES

Measuring the change in house prices between two time points by comparing the simple average sales price at the second with that at the first has the disadvantage, mentioned in the Introduction, that the two sets of houses will almost certainly differ in their attributes. Nonetheless, such price changes are frequently calculated and the Department of the Environment's series is so computed. As regards the alternatives, it is convenient to commence with the simplest, even if it is probably not the best.

The Median

Clearly, an abnormally large number of very high priced, or very low priced, houses would considerably affect mean price, but it would be expected to have far less, if any, effect on the median price. So use of the median price could be seen as, at least partially, controlling for differences in the quality of houses sold in different periods. In the US, for example, the National Association of Realtors house price index is still based on the median and makes no other adjustments for changes in house quality.

The Hedonic Index

The hedonic approach originated in the development of value indices for manufactured products that combined measures of quantity and quality. The seminal paper by Griliches (1961) derived a hedonic price index for motor cars. The basic technique is, for each time period, to regress price (or log price) on the set of variables measuring quality and take the regression coefficients as the implicit prices of the quality components. (In practice, functional forms are chosen on goodness of fit criteria, as will be discussed in Section IV). The index series is then produced by taking some standard set of frequencies of characteristics (usually those of the base year) and applying the successive sets of prices. Suppose the data extend over t time periods and that the regression in the kth time period is:

$$y_{kj} = b_{k0} + b_{k1}x_{1kj} + b_{k2}x_{2kj} + ... + b_{kp}x_{pkj} + e_{kj}\,, \qquad (1)$$

where the subscript j refers to houses within the time period k; $x_1, x_2, \ldots, x_p$ are the measured characteristics or attributes of the house (which could be in logs, if continuous variables like floor area, rather than binary variables like possession of a garage); y is house price or its log and e represents the remaining 'random' variation affecting price. Then the set of coefficients $b_{k0}, b_{k1}, ..., b_{kp}$, can assign a

price to any type of house, where 'type' is defined by the attributes. Thus a price can be assigned in all time periods corresponding to each house type that occurred in the base period, irrespective of whether or not precisely these types recurred subsequently. The overall price index is usually obtained by weighting the household types by their frequencies in the base period. If y is price, differences from the base period give house price changes as increments, while if y is log price, the antilog of differences give price changes in ratio terms.

The Constrained Hedonic Index

The hedonic model is often simplified by assuming that only the constant term in (1) changes with the time period k. That is, the coefficients $b_{k1}, b_{k2}, ..., b_{kp}$, are constrained to stay the same, equal to $b_1, b_2, ..., b_p$, and these coefficients are estimated from all the data. In effect, the model is then

$$y_{kj} = b_0 + b_1 x_{1kj} + b_2 x_{2kj} + ... + b_p x_{pkj} + \sum_{2}^{t} c_i D_i + e_{kj} \ , \qquad (2)$$

where the D's are dummy variables identifying the period in which a particular house sale occurred and the c's are the t – 1 coefficients determining the attribute adjusted price changes. The terminology adopted here of "Hedonic" and "Constrained Hedonic" is not uniformly employed in the literature on house price indices. It is used by some authors, for example, by Crone and Voith (1992), but Gatzlaff and Ling (1994) call (1) the "Strictly Cross-Sectional" model and (2) the "Explicit Time-Variable" one. Knight, Dombrow and Sirmans (1995) refer to (1) as the "Varying Parameter" model and to (2) as the "Hedonic" approach, which seems inconsistent with the underlying idea of evolving implicit prices for attributes. Other terms also appear in the literature, so that statements in papers about the relative merits of rival indices require careful interpretation.

The Repeat Sales Index and Augmented Repeat Sales Approach

Basing an index on repeated sales holds quality constant (ignoring, for the present, such possibilities as structural improvements, etc.) by getting prices for the same houses in different time periods. The model assumes that if P_j is house price in period j, then price in time period j+1 is $(1+a_j)P_j$, where a_j is the appreciation rate between time points j and j+1. In time period j+2 it is $(1+a_{j+1})(1+a_j)P_j$ and so on. If the house is resold at time j+k at price P_{j+k} , then

$$\log(\frac{P_{j+k}}{P_J}) = \sum_{i=j}^{j+k-1} \log(1+a_i) = \sum_{i=1}^{t} \delta_i D_i \,, \qquad (3)$$

where δ =log(1+a) and the D's are dummy variables equal to 1 for time periods between j and j+k and zero otherwise. By regressing the log ratio on the dummy variables, the δ's and hence the a's can be estimated. The set of appreciation rates gives the repeat sales index.

The approach was initiated by Bailey, Muth and Nourse (1963), lacked popularity relative to the median and hedonic methods for a considerable time, but has received increasing attention in the US literature since the mid-eighties (for example: Case and Shiller, 1987, 1989; Abraham and Schauman, 1991). In the original form (3), the repeat sales index does not require any estimation of implicit prices of attributes and so avoids complications associated with attribute variable selection and specification of functional form. It does have its own difficulties, however, one of which has provided much of the motivation for the Augmented Repeat Sales approach. In most data banks of house sale prices, repeat sales will be a small fraction of the total, so that employing the repeat sales measure (3) could mean working with a small sample and ignoring most of the data.

It is possible to augment repeat sales data in various ways, although assumptions are required and the complexity of the analysis increases. Where a sale price record P_j exists for a past period j, but the house has not been resold, an experienced assessor could estimate a current value and substitute it for P_{j+k} in (3). Some of the data on which the repeat sales index would then be based would be 'pseudo' repeat data, rather than genuine. However, such pseudo data do contain information[1] and even if it would seem wrong to grant them equal status with real repeat data, the analysis method could introduce a weighting to offset that. If a hedonic equation of type (1) has already been estimated, other options will exist. For a recently sold (but not recently built) house, P_{j+k} is available, but no record of past price may be available. Then the appropriate hedonic equation could be used to estimate P_j and substitute into (3). Even when both prices are available in a true resale case, the hedonic equation may still have utility in providing corrections if some attributes of the house or its location have changed between sales. It is easy to see that a combined repeat sales and hedonic (full or constrained) analysis is possible that utilises all the data and in fact it can be

[1] In Ireland, the Sherry FitzGerald Indicator of second-hand house prices is based on repeated assessment of a basket of 150 properties, originally sold through Sherry FitzGerald.

formulated either in a generalised least squares or seemingly unrelated regressions framework. Relevant papers implementing these ideas include Case and Quigley (1991), Clapp and Giaccotto (1992) and Knight et al. (1995).

A similar, although still broader, hybrid model was developed by Can (1992) and Can and Megbolugbe (1997). This commences from a hedonic model, but argues that vitally important location effects can only be adequately incorporated by including the prior sales prices of other houses in its neighbourhood as well as its own past sales price, if any. The model is more general than a combined repeat sales and hedonic model because it also contains a spatial dimension to price. The level of detailed local geographic-price information required would seem formidable, but the authors argue that in an era of GIS's and digital technology, all is feasible.

Comparisons of the Models in the Literature

The literature, especially papers based on US data, contains many comparisons of the rival models, although findings are not all in agreement. For example, the basic repeat sales model has been criticised on the grounds that sample selection biases may operate. The idea is that houses with perceived defects, "lemons", and cheap "starter" homes will change hands more often. Case, Pollakowski and Wachter (1991) and Clapp, Giaccotto and Tirtiroglu (1991) did claim evidence for such biases as they found that houses transacted twice within their survey periods were, on average, cheaper than houses transacted once. However, Gatzlaff and Ling (1994), whose data extended over twenty years, found no such biases and attributed the earlier authors' findings to the fact that their survey periods were relatively short, arguing that starter homes were then more likely to distort the repeat sales index. Steele and Goy (1997) have argued for the existence of a reverse bias in Canadian data, where they say activity by "opportune" buyers – who purchase, renovate and resell houses – causes the repeat sale index to overestimate.

This possibility of sample bias in short period studies has been argued as providing another reason to prefer some composite of the repeat sales and hedonic models. It should be noted, however, that if there is a problem with over-representation of starter homes, the hedonic methods could be adversely affected too, because repeat sales of the same house are treated just like single sales of distinct houses. However, even if insufficient data have been recorded to permit the identification of a house reappearance, an appropriate variable – for example, a first time v repeat buyer categorisation – can be added to the hedonic equation to

try to correct for this. This point will be returned to in Section III. Sample selection bias is not, of course, a phenomenon limited to housing data and corrective methods have been developed in other fields, for example by Heckman (1979), and some authors (see Gatzlaff and Haurin, 1998) have applied these to house price index construction.

Most studies that included both the hedonic model (1) and constrained hedonic model (2) made some tests of the assumption of coefficient stability across periods and usually judged it invalid. For example, Gatzlaff and Ling (1994) used the standard Chow test procedure to compare model (1) fitted to each quarter with model (2) fitted to all data and found the coefficients statistically significantly different. However, statistical significance did not always imply numerical differences between the time series of indices that were large enough to matter. Gatzlaff and Ling built up seven indices (the median, hedonics, basic repeat sales and augmentations) for Miami from 1971 to 1991 and found that, with the exception of the median, the final values of all other indices lay within a quite narrow range. They also made comparisons using a more refined measure of precision – the ratio of the standard error of changes in the index to the average standard of the estimates – and were impressed with the precision of the hedonic approaches, even though an augmented resale index was slightly better.

Findings from some other authors are not greatly different. For example, Crone and Voith (1992) found the hedonic, constrained hedonic and basic resale indices all superior to the mean and the median. This is probably no surprise, but their finding that the median seemed inferior even to the mean is unusual. They found the more sophisticated methods so similar in performance that rankings depended on the choice of measure of precision. With mean squared error as the criterion, the hedonic indices were rated better than the repeat sales, but with mean absolute error as the criterion the situation was reversed. On the other hand, Meese and Wallace (1997), analysing 18 years of Californian data, were not at all impressed by repeat sales measures and preferred hedonic methods. They were also far more positive about the simple median index and felt it compared quite well for spatially aggregated data, although a hedonic index was preferable to measure price change in specific municipalities.

A point made by some authors is the importance of sample sizes in assessments of performance of indices. This could obviously be a key issue in choosing between the hedonic and constrained hedonic models. Even if coefficients were believed to differ from period to period, it might be better to estimate

model (2) if there is too little data to get good estimates of the coefficients of model (1) within each period. For example, a quarterly index could be based on (1), but for a monthly, only model (2) might be feasible. However, small sample sizes could be taken as reasons for not working with too narrow time intervals in the first place. Too large a sample was also once considered a problem before powerful PC's existed, and Fleming and Nellis (1985) said that was why they did not estimate the constrained hedonic model (2).

This particular paper is of interest because Fleming and Nellis (1984) first produced the Halifax House Price Index, on which the Irish Permanent / ESRI index is modelled, and their 1985 paper is an account of the technical side of their work. They compared model (1), which they called the "base weighted" method, with the simple average and found it much better, although they did not compare with the median. They did not consider a repeat sales method, nor, for the reason already given, did they calculate the index based on model (2)[2]. Fleming and Nellis (1992) considered the Nationwide index and replicated the rather limited comparison of hedonic and mean indices.

Evidence from the Irish Permanent / ESRI Index

The high profile of the repeat sales methodologies in the US is, at least partly, due to the extensive data bases assembled by realtors associations. In Ireland, it is relatively unlikely that the first and subsequent buyers of a house would have chosen the same financial institution in obtaining mortgages. So repeat sales approaches have been ruled out of consideration. It might be worthwhile, though, ensuring that houses are identifiable in case a unified data bank ever becomes available in the future. However, the topic is really outside the scope of this paper. Comparisons can be made of indices based on the mean, median, hedonic model (1) and constrained hedonic model (2).

It should be mentioned here that the process of computing the Irish Permanent / ESRI index incorporates identification and exclusion of extreme 'outliers' – house prices that seem way outside the range that could be expected given the recorded attributes of the house. This is to guard against data errors and extremely unusual, but unfortunately undocumented, characteristics of the house. A similar procedure is employed in preparing the Halifax index (see Fleming and Nellis, 1985) and is probably operational in most index compilations, although it is not usually given prominence in the literature. However, excluding cases could

[2] They did calculate a partially constrained coefficients model by pooling data over adjacent time periods, giving a kind of moving average of coefficients and report it gave virtually identical results to the hedonic index.

create biases if carried too far and every effort is made to check or correct suspicious cases and to avoid exclusion. Table 1 shows the sample sizes pre and post exclusions, and the effects on mean and median national house prices, from 1st Quarter 1996 to 4th Quarter 1998.

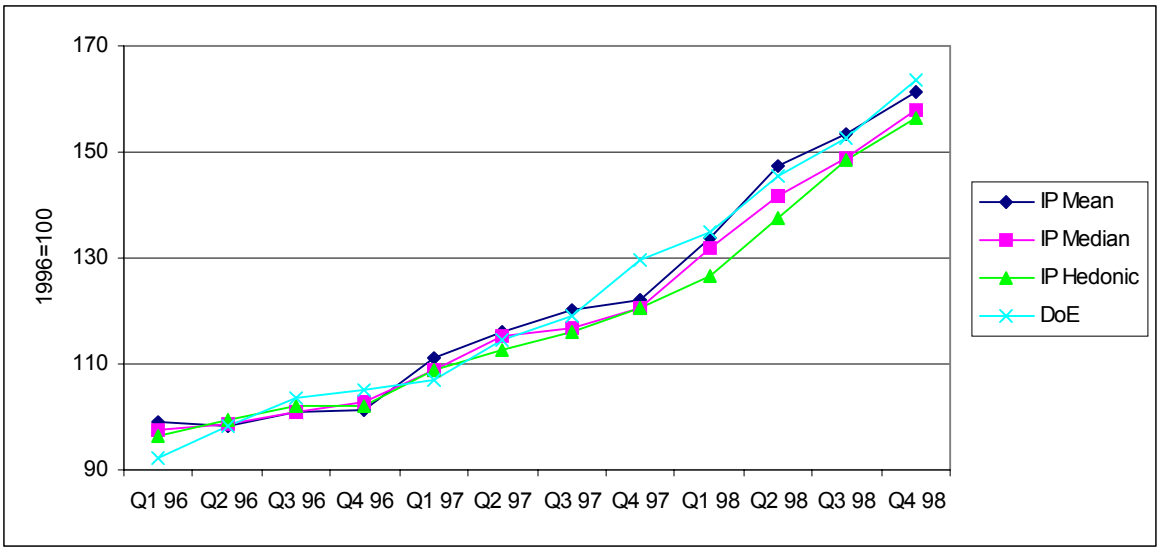
Table 1: National Sample Sizes, Means and Medians, Pre and Post Outlier Removal

	Pre	**Outlier**	**Removal**	**Post**	**Outlier**	**Removal**
Period	Sample Size	Mean	Median	Sample Size	Mean	Median
Q1 96	1484	61.25	55.00	1471	60.60	55.00
Q2 96	2177	61.36	55.95	2154	60.15	55.65
Q3 96	2302	62.25	57.00	2282	61.93	57.00
Q4 96	1987	62.21	58.00	1972	62.20	58.00
Q1 97	1332	68.77	61.62	1320	68.14	61.50
Q2 97	1634	71.86	65.00	1617	71.02	65.00
Q3 97	2366	74.49	66.00	2341	73.74	66.00
Q4 97	1621	76.31	68.00	1592	74.84	68.00
Q1 98	2330	84.74	75.00	2279	81.93	74.50
Q2 98	3022	92.96	80.00	2959	90.20	80.00
Q3 98	3807	95.97	84.50	3721	93.91	84.00
Q4 98	3696	100.87	90.00	3628	98.94	89.25

As can be seen, the number of exclusions is always small relative to sample size. The effect of exclusions is, as would be expected, greater on the mean than on the median, and is generally a reduction. Subsequent tables, analyses and figures will be based on the data purged of outliers.

The indices (1996=100) based on mean, median, hedonic and, for comparison, that maintained by the Department of Environment, are shown in Figure 1. The constrained hedonic index turned out to be extremely close to the hedonic and is not included in Figure 1 because distinguishing them is impossible.

Figure 1: The Hedonic and Other (Mean, Median and DoE) Indices



The indices, except for the DoE initially, move closely together until mid '97. From then on, the mean and the DoE are highest and the hedonic lowest, with the median intermediate and finishing close to the hedonic. The differences between indices are not large relative to the change in all of them from the 1st Quarter of '96 to the final Quarter of '98, but this was a period of continuing rapid price rise. Differences between indices might be more noticeable with a less definite price trend, but it is interesting that the simple median seems to capture much of the downward adjustment the hedonic makes to the mean. But three years is not a long time span.

Another approach to deciding between the hedonic and constrained hedonic is to test for the significance of coefficient changes over time using the standard 'Chow' F test. Table 2 shows the variable coefficients for the quarters of 1996 (model 1) and for model (2), where only the intercepts (not shown) are allowed to vary between quarters. With qualitative variables effects have to be measured relative to a base type. For example, for house type the base is 'Detached', so the negative coefficients for Semi-Detached and Terraced reflect the lower average prices of those house types.

Table 2: Coefficients of Hedonic and Constrained Hedonic for 1996

	Q1	Q2	Q3	Q4	Constrained
First time buyer	-.105	-.106	-.100	-.137	-.111
Semi-Detached	-.036	-.023	-.029	-.062	-.038
Terrace	-.145	-.133	-.160	-.166	-.151
Newly Built	.0984	.106	.0951	.132	.106
Area '000 sq. ft	.453	.483	.481	.310	.432
Garage	.0509	.0347	.0352	.0767	.0463

Type of Central H.	-.130	-.135	-.089	-.156	-.126
Urban Low	-.0698	.0152	-.0234	-.00243	-.0237
Urban Medium	.215	.250	.243	.252	.240
Urban high	.525	.586	.607	.584	.576
Rural Low	-.0734	-.0755	-.0676	-.0376	-.0641
Rural High	.248	.242	.285	.251	.260
F Test for changes	in coeffs.	= 3.64 with 36	and 7827 df	***	

*** = significant at .1% level.

There were 7879 available house records in 1996. A model fitting different coefficients in each quarter requires 36 more parameters than a model fitting the same coefficients but different intercepts in each quarter. The fact that the coefficient changes were so highly significant even within 1996 makes testing over the full three years redundant. Much of the change seems to occur in the final quarter, when, for example, the premium for detached v semi-detached doubles. However, as already said, the indices were virtually identical, showing that when the coefficients are constrained to remain constant, the intercepts are able to absorb most of the effects. This is a very similar finding to that of Gatzlaff and Ling (1994).

III THE VARIABLES IN THE INDEX

A very large number of household attributes or characteristics could be candidates for variables in a hedonic index. Houses differ by age, regional and neighbourhood location, type, design, size, quality of materials, state of repair, possession and size of garage, existence and condition of garden, convenience to transport services and other amenities, and the environmental and 'social' quality of the surroundings. However, data on many characteristics may not have been recorded (convenience to amenities, for example, and even state of repair) and others, while very influential for price, may be difficult to even define objectively, let alone measure (especially the quality of the immediately surrounding location).

As regards the house attributes that usually are available, the joint explanatory power may be much less than the correlations of individual characteristics with price would suggest, because of multicolliniarity. For example, price will be well correlated with total floor area and with number of rooms, but both variables together will explain little more than floor area alone and coefficients will be unstable. The studies in the literature almost invariably start with an assembly of measured variables and reduce them to a parsimonious set of statistically significant and not over correlated variables, while not allowing R^2 to fall appreciably from its maximum with all variables fitted. However, that maximum is often unimpressive, because many important price relevant variables were not available in the first place. These problems are

the reason why the repeat sales approach is so popular in US studies.

How 'Hedonic' need hedonic variables be?

Many indices described in the literature contain variables that are not truly hedonic in that they cannot be considered as direct "quality" measures of house attributes. They are included either because they serve as proxies for important unmeasured variables, or because they are important in their own right, although they do not measure quality alone, or perhaps at all. Regional location variables are an example of the latter. Within a metropolitan area, or even within a region contiguous to such an area, location can most definitely be a quality variable. Identical houses in Foxrock and Finglas will command very different prices and location is one of the qualities a Dublin purchaser takes into account. But a choice between Foxrock and a location in Mayo does not arise for a Dublin purchaser and house price differences between Mayo and Dublin are attributable, not to house quality, so much as to differing levels of economic activity and demographic pressures. So regional location variables can hardly be interpreted as purely hedonic.

Yet regional effects are large and the frequencies of house sales by region fluctuate over time, so some procedure has to be adopted. With enough data, house price indices could be calculated regionally with regions small enough to perhaps justify hedonic interpretation of the variables in each regional index. The national index could be taken as a weighted average of regional indices. However, regressing on regional dummies is one way of standardising with respect to region, which can be implemented even if data are inadequate to estimate indices within each region. In the literature, even when data seem adequate to proceed via regional indices and a weighted average, the regression method seems to be almost universally employed. Fleming and Nellis (1985), for example, used a regional classification of Northern Ireland, Scotland, Wales, and regions of England such as Greater London, South East, South West, etc. and regressed house price on regional dummies, although the Halifax data base was very substantial. Perhaps the reason is that some dummy variables could contain a hedonic element because of the feasibility of commuting from the South East to Greater London etc.

As already said, within region or city location certainly is an important hedonic variable, but there are difficulties to defining and measuring the underlying causitive factors. Some US research (see, for example, Archer and Gatzlaff, 1996; Can and Megbolugbe, 1997) has gone to great efforts to tie location to access to amenities like schools, shopping, recreational facilities etc., but with rather little success. This

might not be surprising in an Irish context either. Much of the perceived value of a high priced house seems to relate strongly to the high value of the surrounding properties and the superior social class of their occupants. Hence Can and Megbolugbe argued that prior house sale prices in the vicinity were the best measures of locality effects. They were not really making a causal argument, but rather claiming prior prices are good proxies for the true causal factors. Other US authors have not taken matters quite this far, but have brought in non-hedonic variables, whether described as such or not. For example, Cropper, Deck and McConnell (1988) included 'hedonic attributes' such as the '% Professional Class' and 'Median Income' of householders in the locality.

Proxies for location effects have appeared in UK indices too. Fleming and Nellis (1992) describe the Nationwide's house price index, which contains regional location variables and immediate neighbourhood variables. The latter are dummy variables classifying the neighbourhood of the house into such categories as: "affluent suburban housing", "high status non-family areas" and "better off retirement areas", at the high income end of the spectrum, and "poorest council estates" and "multiracial areas" at the other end. Clearly what are being indirectly measured here are not characteristics of the houses, but the income status and stage of life cycle of the residents. They prove statistically very significant and are proxies for housing attributes that are important, if perhaps not easily clearly defined. In the Irish Permanent/ESRI index the variable first-time v existing home owner plays a similar role, as will be discussed shortly.

The variables in the Irish Permanent / ESRI index were shown in Table 2. The location variables are always important in the regressions. So is the measure of house size – area in thousands of square feet. The variable first time buyer v existing home owner is usually next in importance. Appendix Table A gives the coefficients, standard errors, t values etc. for a sample quarter to enable assessment of relative importance of the variables. The location variables were chosen to try and capture not only regional, but also at least some of the truly hedonic effects. So, for example, in Dublin the upper price band (Urban High) includes the south inner city, the southern and south eastern suburbs, the coastal northern suburbs from Clontarf to Malahide and a stretch of northern suburbs including Drumcondra and Castleknock. The lower price band contains Ballyfermot, Clondalkin and Jobstown, with the remainder of Dublin city and county constituting the middle price band. A more detailed account of the location variables for the rest of the country and information about the other

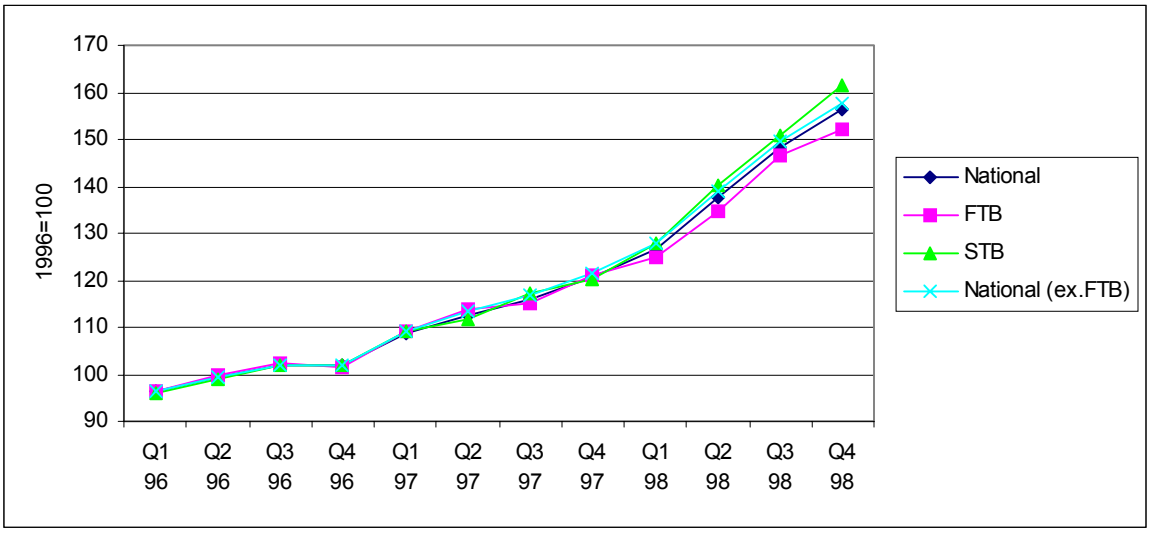
variables in the index is available in Baker and Duffy, 1998.

Obviously, ambiguities about boundaries and the need for appreciable amounts of house sales data for estimation limits the extent of possible subdivision and bands are not really homogenous. Nevertheless, the variables successfully pick up significant variation as Appendix Table A shows. Some of the substantial residual locational variation within bands may be being captured by the first time v other buyer, which has high explanatory power. It seems first time and other buyers usually seek very different properties, so the variable can function as a proxy picking up omitted factors probably including some locational variation within bands. Differences between first time and other buyers are also considered of interest for their own sake by lenders (and the media) and so corresponding sub-indices are published by the Irish Permanent (as they are in the UK by the Halifax and the Nationwide).

However, first time v other buyer, must be suspect as a hedonic measure in an overall index, being a characteristic of the buyer and not the house. It could perhaps be argued that a rise between time points in the proportion of "other" buyers, already property owners, might bid up prices and, if this is then a component of "true" price increase, why should it be standardised for? The same point could, of course, be made about many of the other variables that have appeared in the literature – "median income" as used by Cropper, Deck and McConnell (1988), or the "affluent suburban housing" of Fleming and Nellis (1992) for example. Note too that with the hedonic method, the regression equations are all estimated within periods, so cross-period effects like "bidding up prices" do not invalidate estimation, while the proxy effects benefit it. So in the absence of powerful and truly hedonic measures it would seem unwise to omit any effective price predictor without careful examination Having obtained coefficients though, distinctions could be made in their subsequent use in producing an index. (Although the literature almost always treats all variables, hedonic or not, on the same basis). Given the coefficients, a 'price' can be estimated for any variable set in any time period. Full standardisation takes the same (usually base year) variable set for all time periods, but other options are possible. Some variables could be standardised to base period values and others set to their actual mean (or frequency) in each period. This is equivalent to including these variables in regressions to improve the estimates of other coefficients, but omitting them from the index.

To assess the practical implications, Figure 2 shows an index for first time buyers, for other buyers, for an overall index excluding the first time buyer variable and for the standard hedonic index, already shown in Figure 1.

Figure 2: The National Index: Impact of the FTB Variable (Irish Permanent data)



As is evident, this initial examination suggests there is little, if any, difference between the national index and the national index excluding first-time buyers. However, given the ambiguity of the international literature, and the fact that the Irish Permanent index is so far only based on 4 years data it is too soon to decide about the first-time v second-time buyer variable. Once a longer data series has been collected this topic is one that is worthy of further investigation.

Other Issues Related to Variables: Age Related Hetroscedasticity and Sample Selection Bias

While there is general agreement in the literature on the importance of certain variables such as house size and location measures (or of proxies for the latter), there is less unanimity about other variables. Age of house can serve as an example. In a series of papers Goodman and Thibodeau (1995, 1997, 1998) have

argued that age of house has been treated over simplistically in the literature and that a better approach requires modifications to both the formulation and estimation of the hedonic model. Their idea is that the variance of price (given the variables in the model) will increase with the age of house, because owners may have renovated, or just maintained, or allowed to run down etc. and the older the house, the greater the possible range of condition. They argue that data should not be analysed as homoscedastic (as conventional regression assumes), but as a heteroscedastic model where the variances increase with time. In a sense, this is the issue of omitted variables again, now interpreted as inflating 'random' variance, rather than introducing systematic bias.

However, Goodman and Thibodeau (1995) were criticised for the inadequacy of their location variables and it was suggested that improvements there would remove the heteroscedasticity phenomenon - the argument being that in a 'good' locality an old house would be well maintained, while in a 'bad' one, it might be very run down. Goodman and Thibodeau (1997) admitted that argument carried some weight, but maintained some age related heteroscedasticity remained even after improving the location variables. The topic cannot be taken any further in this paper, although it may deserve future investigation using Irish data.

The idea that sample selection bias can arise from overrepresentation of 'starter' homes, but can perhaps be countered by including an appropriate variable or variables in the regression, was discussed in the previous section. This is another possible justification for introducing a non-hedonic variable like first time buyer, although Gatzlaff and Haurin (1998), who believe sample selection bias likely in short period data, have expressed a contrary view. They believe sample selection bias should be dealt with via Heckman type adjustments, rather than by including non-hedonic variables. So, in relation to the work of Ihlanfeldt and Martinez-Vazquez (1986), who found no evidence of sample selection bias in a short period study (their test method was quite sophisticated, involving censored regression), Gatzlaff and Haurin objected that their regression variables included ".. occupant characteristics; thus, it was not a true hedonic price model". The comment is somewhat inconsistent, because the Heckman procedure can be seen as creating a variable incorporating the prior probability of sample participation and including it in the regression. The generation should employ the extra 'instrumental' variables affecting sample participation and it is difficult to see what these could be if not occupant characteristics. So, in effect, Gatzlaff and Haurin, advocate

adding in a function of non-hedonic variables.

With the Irish data, there are probably more serious sources of sample bias than overrepresentation of 'starter' homes. Using mortgage based data misses out those who do not need to borrow – in particular, 'empty nest' couples trading down. Again, is a sample from any one lending agency equivalent to a random sample of all borrowers? Baker and Duffy (1998) compared the Irish Permanent data set for the year 1996 to that published by the Department of the Environment (DoE) in the 1996 Annual Bulletin of Housing Statistics. They found that new houses account for a smaller proportion of Irish Permanent transactions than in the DoE total. Furthermore the Irish Permanent sample of new houses is more heavily concentrated on semi-detached and terraced houses while having a smaller proportion of bungalows and apartments. Despite this the average price of new houses is very similar. However, the Irish Permanent average price for existing houses is significantly lower than the DoE average, possibly reflecting a different mix of house types and some under-representation of very expensive transactions. In spite of these variations in composition Baker and Duffy concluded that the Irish Permanent appears to be sufficiently comprehensive and representative to provide a suitable base.

Of course, non-random samples do not necessarily imply bias and at worst, findings are valid for the population from which they have been sampled. However, published indices are going to be interpreted as national house price patterns and the topic of possible bias deserves future investigation, either through assembly of a wider database or through Heckman type tests and corrections. Even the latter approach requires assembling some data on the characteristics of currently unrepresented groups.

IV FUNCTIONAL FORM

There do not seem to be any reasons, rooted in economic theory, to prefer any particular functional form to any other one. The regression equation determining the index could perhaps be thought of as one equation of a reduced form of a system of structural equations including price and housing quantity as endogenous variables. However, although some authors commence by mentioning such a formulation, they quickly jump to functional forms reflecting convenience and justified only by statistical goodness of fit criteria. About the most general functional form to appear in the literature was given by Halvorsen and Pollakowski (1981) as

$$\frac{P^{\theta}-1}{\theta} = \alpha_0 + \sum_1^m \alpha_j x_j^{\lambda} + \frac{1}{2}\sum_1^m \sum_1^m \gamma_{jk} x_j^{\lambda} x_k^{\lambda}. \qquad (4)$$

The left hand side of equation (4) is the well known Box-Cox transformation (see Box and Cox, 1964). Taking $\theta = 0$ gives log P as the dependent variable, $\theta = 1$ just gives P and some intermediate value could, at least in theory, be chosen to optimise the fit of the relationship (usually through maximum likelihood). Similar transformations of the attribute variables are possible by varying λ (see Box and Tidwell, 1962). The right hand side of (4) is a quadratic function of powers of the attribute variables, but becomes a linear function of them by taking the γ 's as zero.

In the hedonic approach the equations are fitted period by period, so the functional form, as well as the 'implicit price' coefficients would be changing from period to period. This does not seem particularly plausible and would make an index difficult, if not impossible, to interpret. The approach would be more workable for the constrained hedonic model, but even then there are possible problems. After t periods a value of θ and λ are estimated and the index compiled. However, after the next period the estimates of θ and λ would be different and the entire index would need recompilation. The Halvorsen and Pollakowski paper was theoretical and did not consider such issues, so that several authors (for example, Cassel and Mendelsohn, 1985) were doubtful about the value of approaches based on any general form of (4). Cropper, Deck and McConnell (1988) reported comparisons of various forms of (4) – linear v quadratic with θ and λ either estimated or presumed known. On balance, they favoured a linear model with θ and λ estimated. However, their approach was largely based on simulation – assuming a particular form correct and seeing how inaccurate the others then were – and results depended strongly on assumptions that all relevant variables had been included in the index. They worked with only a constrained hedonic model and did not consider all the practicalities of compiling and interpreting an index.

Almost all the literature has chosen between only four functional forms. In two of these the dependent variable in the regression is price and it is either related to a linear function of the attribute variables or to the logs of these variables where they are quantitative rather than binary (a semilog model). The other two models relate log price to either the linear function of the attribute variables (an exponential model, but mistakenly called semilog in some of the literature), or to the logs of the non-binary variables (a double log model). Choice between models with the same dependent variable has been made on the basis of R^2 or

some other simple criterion, while choice between price and log price has employed Box-Cox testing of θ to see if it is closer to unity than to zero. Gatzlaff and Ling (1994) is a typical example and they ended up favouring an exponential form. So have most other authors, including Fleming and Nellis (1984, 1985) in their work on the Halifax index, and so the functional form

$$\log P_{kj} = b_{k0} + b_{k1}x_{1kj} + b_{k2}x_{2kj} + ... + b_{kp}x_{pkj} + e_{kj} \ , \qquad (5)$$

was adopted for the Irish Permanent / ESRI index.

However, it is worth checking that the Irish data do not contradict this functional form. House area is the only quantitative variable and working with it usually gave slightly better fitting models than working with its log. The differences are *very* slight. For example, for the first quarter of 1996 the R^2 of the model (5) with area was .686 and with log area it was .684. In the fourth quarter of 1996 the R^2 of the log area model was actually larger (.639 as compared to .638). As regards whether the dependent variable should be price or log price, the estimation of θ in the Box-Cox transformation gave values of –0.7 and -.11 for the respective quarters. These values are obviously closer to zero than to unity and, in fact differ significantly from the latter. So log price is a better choice than price as the dependent variable.

Some increase in flexibility over the additive model (5) is obtainable by including interaction terms. An interaction between attributes arises if the effect of one attribute on house price depends on the level of another. For example, could the relationship of house price to size itself change with location? Interaction variables are created by multiplying the pairs of relevant variables and the coefficients of these new variables measure the degree of interaction, if any. Models with interactions can also be considered as sub-cases of the general model (4) and the existence of at least some interactions does not seem implausible. However, with a few exceptions, for example, Grether and Mieszkowski (1974), such models have received little attention in the literature.

Examinations (so far, only performed on the first and fourth quarters of 1996) found some interactions, although not all were statistically significant in both quarters. For example, in the first quarter there was a significant interaction between semi-detached and new houses, the nature being that the price premium between detached and semi-detached houses was less pronounced for new houses. However, the interaction was not significant in the fourth quarter. On the other hand, an interaction between first time buyer and

house area was highly significant in both quarters and its nature was that first time buyers will pay less for increased area than will existing house owners. However, these investigations have only commenced and it is too soon to consider expanding the index to include interactions.

V CONCLUDING REMARKS

The object of this paper was to review the international literature on house price indices, to seek ideas for future developments of the Irish Permanent / ESRI index and to make tentative assessments of how results from Irish data correspond to published research findings. Although the recent vintage of hedonic house price indices in Ireland force emphasis on the word 'tentative', it is interesting that there are correspondences with US papers. For example, there is the statistical rejection of the hypothesis of equality of coefficients over time required for the constrained hedonic model and yet the indices tracked similiarly. There is the relatively good showing of the very simple median based index, which was never considered in UK studies, but which was also found here. However, there are only three years of Irish data and these were years of continuous house price appreciation.

The literature suggests many ideas that can be investigated as data accumulates. These include the refinement of measures of location variables, the introduction of interactions and perhaps even the development of one of the hybrid hedonic and repeat sales models. The prevalence of repeat sales models and the hybrids in the US literature has not occurred without good reason. Even if a hybrid model seems a distant prospect, there is a need now to consider how to identify future repeat sales. There is also a need to be concerned about possible sample biases and how data might be augmented to test for them and, if necessary, correct them. This is very much a concern in the US literature, where the data sets assembled through realtors may well be more representative than Irish data.

REFERENCES

Abraham, J. M. and W. S. Schauman, 1991. "New Evidence on House Prices from Freddie Mac Repeat Sales", *American Real Estate and Urban Economics Association Journal*, Vol. 19, pp. 333-352.

Archer, W. R. and D. H. Gatzlaff, 1996. "Measuring the Importance of Location in House Price Appreciation", *Journal of Urban Economics*, Vol. 40, pp. 334-353.

Bailey, M. J., R. F. Muth and H. O. Nourse, 1963. "A Regression Method for Real Estate Price Index Construction", *Journal of the American Statistical Association,* Vol. 58, pp. 933-942.

Baker, T. and D. Duffy, 1998. *House Price Index: Methodology Explained*, Dublin; Irish Permanent

Box, G. E. P. and D. R. Cox,1964. "An Analysis of Transformations", *Journal of the Royal Statistical Society B,* Vol. 26, pp.211-243.

Box, G. E. P. and P. Tidwell,1962. “Transformation of the Independent Variables”, *Technometrics,* Vol. 4, pp. 531-550.

Can, A., 1992. “Specification and Estimation of Hedonic Housing Price Models”, *Regional Science and Urban Economics*, Vol. 22, pp. 453-474.

Can, A. and I. Megbolugbe, 1997. “Spatial Dependence and House Price Index Construction”, *Journal of Real Estate Finance and Economics,* Vol. 14, pp. 203-222.

Case, B., H. O. Pollakowski and S. M. Wachter, 1991. ″On Choosing among House Price Index Methodologies", *American Real Estate and Urban Economics Association Journal*, Vol. 19, pp. 286-307.

Case, K. E. and R. J. Schiller, 1987. “Prices of Single Family Homes since 1970: New Indexes for Four Cities”, *New England Economic Review,* September-October, pp.45-56.

Case, K. E. and R. J. Schiller, 1989. “The Efficiency of the Market for Single Family Homes”, *American Economic Review,* Vol. 79, pp.125-137.

Case, B. and J. M. Quigley, 1991. “The Dynamics of Real Estate Prices”, *The Review of Economics and Statistics,* Vol.73, pp. 50-58.

Cassel, E. and R. Mendelsohn, 1985. The Choice of Functional Forms for Hedonic Price Equations: Comment”, *Journal of Urban Economics*, Vol. 18, pp. 135-142.

Clapp, J.M., C. Giaccotto and D. Tirtiroglu, 1991. ″ Housing Price Indices based on all transactions compared to repeat subsamples", *American Real Estate and Urban Economics Association Journal*, Vol. 19, pp. 270-285.

Clapp, J.M. and C. Giaccotto, 1992. “Estimating Price Indices for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods”, *Journal of the American Statistical Association,* Vol. 87, pp. 300-306.

Crone, T. M. and R. P. Voith, 1992. “Estimating House Price Appreciation: A Comparison of Methods”, *Journal of Housing Economics*, Vol. 2, pp. 324-338.

Cropper, M. L., L. B. Deck and K. E. McConnell, 1988. “On the Choice of Functional Form for Hedonic Price Functions”, *The Review of Economics and Statistics,* Vol.70, pp. 668-675

Fleming M. C. and J. G. Nellis, 1984. *The Halifax House Price Index - Technical Details*, Halifax: Halifax Building Society.

Fleming M. C. and J. G. Nellis, 1985. “The Application of Hedonic Indexing Methods: A Study of House Prices in the United Kingdom”, *Statistical Journal of the United Nations,* ECE 3, pp. 249-270.

Fleming M. C. and J. G. Nellis, 1992. “Development of Standardised Indices for Measuring House Price InflationIndexing Methods Incorporating Physical and Locational Characteristics”, *Applied Economics*, Vol. 24, pp. 1067-1085.

Gatzlaff, D. H. and D. C. Ling, 1994. "Measuring Changes in Local House Prices: An Empirical Investigation of Alternative Methodologies", *Journal of Urban Economics*, Vol. 40, pp. 221-244.

Gatzlaff, D. H. and D. R. Haurin, 1998. "Sample Selection and Biases in Local House Value Indices", *Journal of Urban Economics*, Vol. 43, pp. 199-222.

Goodman, A. C. and T. G. Thibodeau, 1995. "Age Related Heteroscedasticity in Hedonic House Price Equations", *Journal of Housing Research*, Vol. 6, pp. 22-42.

Goodman, A. C. and T. G. Thibodeau, 1997. "Dwelling Age Related Heteroscedasticity in Repeat Sales House Price Equations", *Real Estate Economics*, Vol. 26, pp. 151-171.

Goodman, A. C. and T. G. Thibodeau, 1998. "Dwelling Age Heteroscedasticity in Hedonic House Price Equations", *Journal of Housing Research*, Vol. 8, pp. 299-317.

Grether, D. M. and P. Mieszkowski, 1974. "Determinants of Real Estate Values", *Journal of Urban Economics*, Vol. 1, pp. 127-146.

Griliches, Z., 1961. "Hedonic Price Indices for Aotomobiles", in Z. Griliches (ed.), *Price Indices and Quality Change: Studies in New Methods of Measurement*, Cambridge: Harvard University Press.

Halvorsen, P. H. and H. O. Pollakowski, 1981. Choice of Functional Form for Hedonic Price Equations, *Journal of Urban Economics*, Vol. 10, pp. 37-49.

Heckman, J., 1979. “Sample Selection Bias as a Specification Error”, *Econometrica,* Vol. 47, pp. 153-161.

Hendershott, P. H. and T. G. Thibodeau, 1990. ″The Relationship between Median and Constant Quality Quality House Prices: Implications for setting FHA loan limits", *American Real Estate and Urban Economics Association Journal*, Vol. 18, pp. 323-334.

Ihlanfeldt, K. R. and J. Martinez-Vazquez, 1986. “Alternative Value Estimates of Owner-Occupied Housing: Evidence on sample selection bias and systematic errors.” *Journal of Urban Economics*, Vol.

20, pp. 357-369.
Knight, J. R., J. Dombrow and C. F. Sirmans, 1995. "A Varying Parameters Approach to Constructing House Price Indexes", *Real Estate Economics*, Vol. 23, pp. 187-205.
Meese, R. A. and N. E. Wallace, 1997. "The Construction of Residential Housing Price Indices: A Comparison of Repeat Sales, Hedonic-Regression and Hybrid Approaches", *Journal of Real Estate Finance and Economics,* Vol. 14, pp. 51-73.
Steele,M. and Got, R., 1997. "Short Holds, the Distributions of First and Second Sales, and Bias in the Repeat-Sales Price Index", *Journal of Real Estate Finance and Economics,* Vol. 14, pp. 133-154.

Appendix A: Quarter 4, 1996 regression results

	Unstandardised Coefficients		Standardised	t
	B	**Std.Error**	**coefficients**	
Constant	10.562	0.022		486.709
FTBDV	-.137	0.011	-.183	-12.423
SDDV	-.06223	0.012	-.081	-5.101
TCEDV	-.166	0.016	-.178	-10.613
NEWDV	.132	0.012	.175	11.184
FOOTAGE (‘000)	.3097	0.000	.380	24.855
GARDV	.07671	0.013	.086	5.852
SOLDV	-.156	0.013	-.183	-12.157
URLDV	.002432	0.041	.001	.060
URMDV	.252	0.015	.268	16.884
URHDV	.584	0.022	.396	27.100
RURLDV	-.03757	0.017	-.032	-2.209
RURHDV	.251	0.015	.255	16.659

Note: The base property against which differences in price is measured is an existing detached house (or bungalow) in a “medium” location (non-Dublin in the case of the national index), with a non-first time buyer, no garage, and gas, electricity or oil heating.