Challenges in Compiling a National House Price Index for Ireland

O'Hanlon, Niall
Central Statistics Office,
Skehard Road,
Cork,
Ireland.
niall.ohanlon@cso.ie

Disclaimer
Any of the views expressed are those of the author and do not reflect the views or policies of the Central Statistics Office.

1. Preface

The Central Statistics Office Ireland (CSO) is currently developing a national house price index using monthly mortgage data supplied to the Ministry of Environment, Heritage and Local Government (DoEHLG) by Irish mortgage lending institutions. As this work is still ongoing the list of challenges contained in this paper is not exhaustive – further challenges will undoubtedly be encountered as we move closer to producing a finished index. Therefore this paper primarily focuses on the challenges we have encountered during the initial phases of development; in securing data, improving its quality and internal consistency and in preparing it for use in index compilation.

2. Introduction

This paper describes the CSO's experience of using an administrative data source designed and compiled by another government ministry. While collaborating with the DoEHLG has given the CSO access to transactions based data it has made us, at least until very recently, quite dependent on that ministry. Consequently the CSO has not always been able to determine the rate of progress of this work.

Although a very rich source of information in its own right, the monthly mortgage file can be considered suboptimal as it does not contain micro location detail. Furthermore the data lacks a consistency that might be assumed inherent in data generated from administrative records. This is compounded by the fact that it originates from multiple mortgage lenders, each of whom operates different systems and practices. Data of this type cannot therefore be assumed to be

homogenous in respect of quality. This adds a degree of complexity to the treatment or cleaning of data prior to its use in index design and calculation.

3. Utilising an existing data source

Irish mortgage lenders are required under Section 13 of the Housing (Miscellaneous Provisions) Act 2002, to submit monthly mortgage returns to the DoEHLG, containing data on both mortgage approvals (occurring where a formal letter of mortgage offer has issued) and mortgage drawdowns (where the loan has been drawn down). This data requirement was set up primarily to generate a national mix-adjusted house price index but also to produce other relevant statistics that would inform housing policy generally.

The data consist of an individual record for every loan approval and loan drawdown made by the lender in the reference month. The data is anonymised – neither the individual borrower nor the property to be purchased is identifiable. Each record contains 67 variables of which; 2 relate to the financial institution, 32 to the borrower(s), 18 to the loan details and status and 15 to details of the property to be mortgaged. These variables are set out at Annex I. Those variables relating to the property to be purchased, which could be used directly in the compilation of a House Price Index are;

- Transaction type – private purchase or with government subsidy
- Agreed purchase price of the property
- County of location (26 administrative regions)
- City indicator (for 4 cities excluding Dublin)
- Postcode, where relevant (for Dublin only)
- Newly built property indicator
- Year of Build
- Dwelling Type (detached, semi detached, terraced, flat or bungalow
- Construction type (brick/block, timber frame or pre-cast concrete)
- Floor area
- Plot size (land)
- Number of rooms
- Number of bedrooms
- Use of Property
- Price at drawdown

It should be noted the national location descriptors are limited to administrative county, city indicator and postcodes where relevant. Ireland does not yet have a national postcode system. Postcodes are limited primarily to Dublin City which is divided into 22 postal districts. The CSO did not have any input into the design of this reporting requirement. The absence of

detailed micro-location does at very least present a considerable challenge for index compilation but could ultimately severely limit the usefulness of these data for the purposes of constructing a house price index. However the dataset does, at present, offer the best potential source of data in respect of a transactions based index. Other possible sources such as administrative data generated by the taxation and property registration processes are not suitable for reasons of completeness and timeliness. Furthermore the CSO recognises the considerable investment made by the DoEHLG and the mortgage lenders in developing this dataset and the potential usefulness of it beyond index compilation. The CSO is anxious therefore to use it if at all possible. We are not at this time contemplating introducing a new reporting requirement for mortgage lenders but may seek to make changes to the current specifications at some future date. A national postcode system is currently being developed and it is expected to be implemented in 2011. It is likely that at that time the CSO will seek to have this new national postcode added to the current address requirements.

4. Collaborating with the data owner

The DoEHLG developed a House Price Statistical System (HPSS) to process data and produce statistical analysis, including a national house price index. It was envisaged that mortgage approvals data would be used as monthly price observations as these represent the earliest formal recording of agreed price. The original proposal for a mix adjusted design followed a fine stratification approach where each month mortgage approvals were stratified into 288 cells (8 geographic regions * 3 house size categories (based on number of bedrooms) * 3 house types * 2 buyers status types (first time buyer or otherwise) * 2 age of property (new or previously lived in)). Mortgage drawdowns would be used to internally weight each of the 288 cells.

The DoEHLG was not happy with the results of its initial attempts at developing indices however. For the years 2005 and 2006 prices grew at a much higher rate than those in the best known index published in Ireland (the PTSB/ESRI index compiled using data from one mortgage lender).

In mid 2007 the DoEHLG asked the CSO to become formally involved in the design and build of a national house price index. In response to this request and the impending requirement for Owner Occupied Housing (OOH) indices in the context of the EU Harmonised Index of Consumer Prices, the CSO assigned a statistician to work on HPI and OOH indices in 2008. During 2008 the DoEHLG transmitted data from its HPSS for the period 2005-2007 to the CSO. These data contained only those records which had passed the HPSS validation system and so were deemed “clean”. The DoEHLG was of the opinion that the HPSS should continue to be the primary processing system. Once the CSO had successfully designed an index it would be produced on a monthly basis using the HPSS. The early phase of work for the CSO therefore involved testing and refining the fine stratification approach, using the 288 cell design and numerous variants of it. New edits and outlier checks based on tukey and median absolute deviation approaches were set up, and a series of prototype indices were calculated. These

indices, of both monthly and quarterly periodicity and with lesser and greater degrees of fine stratification than the original 288 cell design all produced indices which gave implausible results. In particular the indices all failed to identify falling prices that the market experienced from early 2007. Investigation of price movement at the elementary aggregate level showed that the largest increases tended to be for those cells with higher value houses – so larger price increases at the high end of the market heavily influenced the index.

An in-depth analysis of data quality revealed a number of significant weaknesses including systematic errors in data supplied by individual mortgage lenders. As the CSO was only receiving a "validated" file it was not possible to determine the true level of data quality of all records supplied by the lenders. The CSO requested access to original files and eventually these were delivered thereby allowing the CSO to fully analyse the quality of data supplied by each mortgage lender. This analysis showed marked differences in data quality across institutions, between variables and over time.

Although the DoEHLG had been periodically liaising with the lenders on data quality issues the CSO was initially provided with very little information in respect of these contacts – and usually only as a response to specific questions which it raised. However in late 2009 the DoEHLG agreed that the CSO should discuss data quality issues directly and bilaterally with the individual mortgage lenders. Importantly it was also agreed that the CSO could propose and agree with the lenders alternative data reporting arrangements (such as making some problematic variables non-mandatory) where it deemed necessary so as to improve overall data quality. This agreement has allowed the CSO to focus its data quality improvement efforts on those variables relating to the property.

5. Challenges presented by data quality issues

Overall the quality of data is not as good as might be expected given that they are drawn from administrative records generated by formal mortgage approvals and loan drawdowns. The table at Annex 2 shows the percentage of approvals records which fail basic edit checks (mainly missing values) on each of those variables that might be used for index calculation. In each of the quarters presented less than 50% of all records pass these checks.

It is also clear from the table that quality is not consistent over time. Quality improves markedly in respect of construction type and size of property in Q1 2007 while an improvement in the quality of year of construction does not occur until Q2 2008. Missing plot size however remains very high and is by far and away the single biggest problem. Rather than excluding records with missing plot size, the impact of this error could be reduced by using the plot size variable only in respect of detached houses as it is less likely to be a significant price determinant in other house types. Similarly it may be less significant in urban areas where plot sizes tend to be more uniform in size. The quality of variables is therefore an important determinant in the choice of characteristics in a hedonic model. However this can be complicated by variation in quality over time and between different mortgage lenders. One mortgage lender does not provide detail on the

number of rooms, otherwise the quality of its data is excellent. Excluding records on the basis of missing number of rooms would result in the entire set of records for that lender being disregarded. Alternatively imputation of missing values might be deemed acceptable in the case of some variables, perhaps for earlier periods when data quality was poorer.

The variation in data quality from lender to lender is considerable and further complicates efforts to secure consistency of quality. The tables at Annex 3 and Annex 4 show the percentage of approvals records which fail basic edit checks for 2 different lenders. The substantial differences in quality between the 2 show that bilateral cooperation with individual lenders is required. In the case of the lender whose data is described at Annex 4 - discussions between the CSO and lender revealed that the approval records are generated from provisional loan offers. As the lender does not complete valuation reports prior to these offers (instead relying on the applicant to supply detail on property characteristics), the quality of data provided is much poorer. The CSO and lending institution have agreed that a separate file generated on the basis of the formal valuation report will be provided to the CSO. This file will contain just those variables listed previously that relate to the property to be purchased only. This arrangement would not have been possible without the CSO first having permission to negotiate solutions with individual lenders.

6. The collapse in Irish housing market

The value of new mortgage lending for residential property purchased peaked at just over €7.7billion in the 3rd quarter of 2006. By the same quarter of 2009 this had fallen to €1.6billion, a decrease of 80%. The number of mortgage approvals (price observations for the index) also fell by approximately 80% from just over 35,000 approvals in the 3rd quarter of 2006 to around 7,100 in the same quarter of 2009. Such a dramatic collapse in market activity has given rise to two important points in respect of the development of a house price index. Firstly, the fine stratification approach as originally envisaged cannot be supported by the greatly reduced number of observations generated (at least on a monthly basis) and so the hedonic method of index construction has become the sole workable approach.

Secondly, the decline in new lending has not been uniform across lenders. Of the 11 lenders that have supplied data since 2005, 6 generated less than 10% of the number of approvals in Q3 2009 as they did for the same period 2 years previously. In Q3 2009 6 lenders accounted for 95% with the remaining market share distributed among 3 other lenders. This change in market composition has given rise to a number of issues to be considered during the index design stage. Is it worthwhile to attempt to engage in bilateral efforts to improve historical data with lenders who have effectively stopped generating new mortgage lending? Can we reasonably expect these lenders to engage fully with us and to resource efforts to improve the quality of historical

data? For the sake of practicality is it appropriate to focus on those 6 lenders dominating the market?

The CSO has decided to concentrate its initial efforts in improving data quality to those 6 mortgage lenders generating the vast majority of new business. The early stages of the design and build of a hedonic index will initially focus only on data from these lenders but over time data from the other lenders will be included, quality permitting.

7. Lack of address detail

The lack of micro location detail on the DoEHLG dataset and the absence of a national postcode system in Ireland further complicate the design of a hedonic index. It is not possible to identify location beyond administrative county, city or postcode (where they exist). Where typically a broad geographic indicator (such as region) and a micro location indicator (such as neighbourhood classification) might be used, the CSO has during the early design stages used a single geographic indicator identifying, where relevant, administrative county, city or postcode. This results in almost 70 different classes.

The Census Division of the CSO calculates various socio-economic measures by small area. It is currently aggregating these measures to Dublin postcode level which will allow for testing within a hedonic model of a quality of location characteristic. However there can be considerable variation in socio-economic quality within certain Dublin postcodes.

8. Conclusion

The CSO is in a somewhat privileged position in that it has access to a rich set of data on mortgage transactions, albeit without detailed location of property information. However the quality of these data varies considerably between variables, over time and between reporting institutions. Data originally provided by mortgage lenders should therefore be considered as coming from different administrative data sources and so direct cooperation with individual mortgage lenders is vital. Flexibility in data preparation and model specification is required so that the best use of heterogeneous data can be achieved. As price observations are based on actual transactions as opposed to offer prices (as is typically the case for other price indices) the house price index must be designed such that it can cope with substantial falls in the number of observed prices.

Annex 1. Variables collected in DOEHLG dataset

Financial Institution	Borrower	Loan	Property
Institution Code	Number of Male	Amount of Approval	**Transaction Type**
Sequence Number	Number of Female	**Property Acquisition**	**Price of Property**
File Month	Age of Main	PA Nature	**Location County**
	Age of Second	NPA Nature	**Location Dublin**
	Gender Main	Loan Term	Location Abroad
	Gender Second	Initial Gross IR	**New or Second Hand**
	Marital Status Main	Rate Type	**Year Built**
	Marital Status Second	Years Fixed	**Dwelling Type**
	Employment Status Main	Loan Type	**Construction Type**
	Employment Status Second	Means of funding Gap	**Floor Area**
	Employment Sector Main	LA Clawback	**Plot size**
	Employment Sector Second	Loan Approval	**Rooms**
	Occupation Main	LA Date	**Bedrooms**
	Occupation Second	Drawdown	**Price of Property at Drawdown**
	Buyer Status Main	DD Date	
	Buyer Status Second	Total Loan Amount	**Use of Property**
	Gross A Income Main	Other Costs	
	Gross A Income Second	Indemnity Bond	
	Net M Income Main		
	Net M Income Second		
	Other Non Rental Income Main		
	Other Non Rental Income Second		
	Rental Income Main		
	Rental Income Second		
	Current Tenure Main		
	Current Tenure Second		
	Location County Main		
	Location Dublin Main		
	Location Abroad Main		
	Location Dublin Second		
	Location County Second		
	Location Abroad Second		

Annex 2. Percentage of error records by type – all lenders

Quarter	20051	20052	20053	20054	20061	20062	20063	20064	20071	20072	20073	20074	20081	20082	20083	20084	20091	20092	20093
% Error Records	71.5	72.1	68.5	74.7	67.5	72.3	70.4	71.9	57.5	59.5	60.3	59.9	55.8	58.1	61.4	61.6	62.3	58.1	54.5
wrong month	0.0	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
no price	1.4	1.6	1.0	1.8	1.6	1.4	1.0	0.9	0.3	0.4	0.4	0.6	0.3	0.4	0.3	0.1	0.0	0.0	0.0
no county	0.1	0.4	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.2	0.0	0.0	0.1	0.1
no exact location	2.8	2.6	2.5	1.3	1.4	1.2	1.4	2.0	1.8	1.8	1.5	1.5	2.2	2.1	1.7	1.8	1.6	2.1	2.1
incorrect Dublin postcode	1.7	1.2	1.4	1.9	1.7	1.5	1.3	1.3	0.9	1.0	1.0	1.4	1.2	1.1	0.9	1.0	0.7	0.7	1.1
no buyerstatus	0.0	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	1.2	2.2	1.1	0.9	1.8	2.9	3.3	2.3	1.5	1.2
no property use	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
no housestatus	0.6	0.8	0.4	1.0	0.7	0.6	0.5	0.7	0.6	0.9	0.7	0.6	0.5	0.2	0.1	0.0	0.1	0.0	0.0
no year construction	20.4	21.1	21.1	17.5	19.5	22.6	21.4	19.5	16.9	17.0	15.1	12.2	9.4	5.3	3.8	4.2	3.7	4.1	4.9
not built yet	3.3	5.2	0.4	2.6	0.0	0.0	0.1	2.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	5.0	0.0	0.0	0.2
secondhand new	1.7	2.2	2.6	2.0	1.8	1.6	2.2	2.1	1.9	1.8	2.0	2.1	1.5	4.5	3.8	2.1	1.9	1.8	2.3
no housetype	2.0	3.5	1.7	2.4	1.6	1.7	1.6	2.0	1.7	1.7	1.9	1.8	2.0	2.0	2.2	1.3	1.1	1.2	1.6
no construction type	32.5	30.0	24.5	33.1	28.8	30.1	28.3	30.2	9.8	9.2	11.8	8.5	7.0	7.5	8.4	5.8	2.7	3.8	5.3
no number beds	27.8	22.1	21.0	30.7	26.4	27.6	25.9	27.2	3.3	3.2	4.7	5.0	4.5	4.8	5.2	3.4	2.5	2.8	3.6
no floor area	29.8	25.2	22.9	32.5	27.3	28.5	26.8	28.7	8.0	8.0	11.4	9.4	7.2	9.4	10.8	8.8	3.8	3.2	5.2
no number rooms	33.5	28.1	26.0	34.2	30.7	31.2	29.2	30.5	7.5	7.8	7.0	6.4	5.0	5.2	5.2	3.2	2.0	2.8	3.1
more bedooms than rooms	2.6	2.1	2.3	2.9	2.2	2.9	2.6	2.3	6.7	6.6	8.0	8.3	6.1	6.3	6.6	6.8	4.6	4.7	7.1
no plot size	40.6	40.4	37.7	42.9	36.6	40.4	40.3	40.6	29.0	32.7	35.5	34.8	39.6	40.5	44.9	43.9	50.7	45.9	38.7

Annex 3. Percentage of error records by type – lender A

Quarter	20051	20052	20053	20054	20061	20062	20063	20064	20071	20072	20073	20074	20081	20082	20083	20084	20091	20092	20093
% Error Records	18	40	20	36	21	19	19	31	21	18	17	30	14	26	17	25	10	12	11
wrong month	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no price	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no county	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no exact location	6	6	6	6	6	6	6	7	7	7	8	7	7	8	6	5	5	8	7
incorrect Dublin postcode	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no buyerstatus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no property use	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no housestatus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no year construction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
not built yet	0	14	0	21	0	0	0	14	0	0	0	15	0	0	0	16	0	0	0
secondhand new	7	6	9	6	9	8	8	6	8	6	6	6	4	18	10	4	3	3	4
no housetype	0	6	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
no construction type	0	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no number beds	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
no floor area	0	6	0	1	1	0	0	0	0	1	1	1	0	1	0	0	0	0	0
no number rooms	0	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
more beds than rooms	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no plot size	5	5	5	6	6	6	5	7	7	5	3	4	3	2	1	0	0	0	0

Annex 4. Percentage of error records by type – lender B

Quarter	20051	20052	20053	20054	20061	20062	20063	20064	20071	20072	20073	20074	20081	20082	20083	20084	20091	20092	20093
Error Records	98	98	98	97	98	98	97	97	75	74	76	68	69	71	69	63	70	60	57
wrong month	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no price	5	5	5	5	5	4	3	3	1	2	1	2	1	2	1	0	0	0	0
no county	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no exact location	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
incorrect Dublin postcode	3	3	2	3	3	3	2	3	2	2	2	3	3	3	2	3	3	3	4
no buyerstatus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no property use	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
no housestatus	2	2	2	3	3	2	2	2	3	4	3	2	2	1	0	0	0	0	0
no year construction	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0
not built yet	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
secondhand new	1	1	2	2	1	1	2	2	0	1	1	1	0	1	1	2	1	2	2
no housetype	6	5	5	5	4	4	5	5	6	6	4	4	5	4	4	2	3	2	1
no construction type	92	92	90	89	92	90	89	90	31	26	33	22	19	20	23	16	11	10	11
no number beds	90	89	89	87	91	89	87	87	12	10	12	13	13	12	13	8	8	7	7
no floor area	92	92	91	89	92	90	89	90	29	28	34	27	23	25	26	19	14	11	12
no number rooms	90	88	88	86	91	89	87	87	12	11	13	13	13	14	14	8	9	8	7
more beds than roo	3	3	3	4	3	3	4	3	23	22	22	22	20	19	19	20	20	18	19
no plot size	76	77	76	70	75	73	74	72	31	32	36	28	26	29	29	27	23	22	21