1352.0.55.093

Australian Bureau of Statistics

Research Paper

Refining the Stratification for the Established House Price Index

www.abs.gov.au

New Issue

Research Paper

Refining the Stratification for the Established House Price Index

Alexa Olczyk and Steve Lane

Analytical Services Branch

Methodology Advisory Committee

13 June 2008, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 21 AUG 2008

ABS Catalogue no. 1352.0.55.093

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Alexa Olczyk, Analytical Services Branch on Canberra (02) 6252 5854 or email <analytical.services@abs.gov.au>.

REFINING THE STRATIFICATION FOR THE ESTABLISHED HOUSE PRICE INDEX

Alexa Olczyk and Steve Lane
Analytical Services Branch

QUESTIONS FOR THE COMMITTEE

1. Does the use of price as a stratification variable introduce a conceptual discrepancy?
   i.e. Is this a case of price explaining price?

2. Does our specification of the suburb price level as the long term median price make sense?
   i.e. Is the use of the mean of quarterly medians over a four and a half year period (March quarter 2002 to June quarter 2006) appropriate?

3. Does the use of distance variables in the locational principal component (or as stand-alone stratification variables) make theoretical sense?
   i.e. Do the centroid distances from the suburb to the CBD, the nearest hospital and the nearest shops etc. seem reasonable as stratification variables? Do they represent attributes that determine the similarity of suburbs?

4. Are our methods of assessing stratification results appropriate?
   i.e. Is the Quality Assessment Framework (QAF) a robust approach for determining which stratification results are most effective when controlling for compositional change?

5. Are there alternative methods that could be applied when assessing the effectiveness of stratification results?

CONTENTS

REFINING THE STRATIFICATION FOR THE ESTABLISHED HOUSE PRICE INDEX

Alexa Olczyk and Steve Lane
Analytical Services Branch

ABSTRACT

The Australian Bureau of Statistics publishes quarterly estimates of the change in the price of established houses in Australia based on a method that has developed over time and has always involved some form of stratification. The established house price index (HPI) methodology is currently based on attributes that can be broadly defined as the structural, locational and neighbourhood characteristics of suburbs. This approach to constructing the HPI has been assessed and the compilation process has brought to light issues concerning the limited ability of the current stratification to measure the pure price evolution of the housing stock. The price-based stratification approach introduced by the Reserve Bank of Australia in 2006 has also paved the way for an investigation into potential improvements to the current stratification method through the use of the long-term median price of a suburb as a stratification variable. In light of these developments, this paper explores the possibility of refining the stratification used to construct the HPI. A three-phase Quality Assessment Framework, supported by index sensitivity analysis, is used to analyse the effectiveness of competing stratification methods. Preliminary findings based on applying this framework indicate that a simpler stratification method – based on the long-term median price and the neighbourhood characteristics of a suburb – can provide a more accurate measure of the pure price evolution of the housing stock.

1. INTRODUCTION

1.1 Difficulty in measuring house price movements

In Australia, as in many other countries, movements in house prices have significant social and economic impacts and therefore their measurement is very important. However, accurately measuring the changes in house prices over time is neither simple nor straightforward. It involves complex conceptual, methodological and data issues. Unlike other price indexes, house price indexes are inherently difficult to construct and thus continue to be a challenging area for statistical agencies.

Price indexes can be constructed by one of two approaches. Standard price index methodology selects a sample of representative items and re-prices them through time using transaction prices and activity weights (based on the number of items that are sold) to measure change in the selling price of representative items. The alternative, asset price index methodology, re-prices the stock of representative items using transaction prices and stock weights (based on the number of items in the stock) to measure change in the price of the stock of representative items.

It is important to realise that the standard price index methodology does not apply in the case of house price indexes for several reasons. First, no two houses are absolutely identical – houses are heterogeneous goods with differing physical characteristics and varied locations. Second, the number of transactions (house sales) is low relative to the housing stock. Third, infrequent turnover means that instead of being able to observe the price of each house at every point in time, the price of a house is revealed only when the house is sold. Thus, even if a representative sample of houses could be assembled, very few of them would be transacted in any period.

A feasible alternative to using standard price index methodology is to construct an asset based house price index (HPI) that measures the change in the price of the housing stock. However, since only a fraction of the total stock of housing is transacted in any period, HPI measures based on transaction prices are dependent on the mix of houses sold in each period. Given the heterogeneous nature of the housing stock, changes in the average price of house sales from one period to the next will reflect changes in the types of houses being transacted, as well as pure price changes. For example, if transactions in one period consist mostly of larger houses or houses in more expensive areas, average transaction prices will appear to have increased, even though the price for a given type of house may remain unchanged. This problem arises because the sample of houses selling in any period is self-selected and transactions are driven by the erratic buying and selling patterns within the housing market, rather than controlled by the usual random sampling techniques.

The problem of sample bias arises as the mix of dwellings in any given period could vary with changing economic conditions and changing consumer preferences – a phenomenon known as *compositional change*.

1.2 Stratification as a solution

The challenge, therefore, is to construct a measure of house price changes given these complications by abstracting from compositional change and capturing only the pure price evolution of the housing stock. The most commonly used method for compositional adjustment is the simple weighted average method, sometimes called mix-adjustment or stratification. The stratification approach involves grouping transactions in such a way as to combine the 'most similar' houses.

Generally, stratification groups objects according to a set of defining characteristics, so that objects within each group are relatively closer to each other than to objects belonging to other groups. The objective is to minimise the heterogeneity of objects within each group.

The Australian Bureau of Statistics (ABS) uses the stratification approach to construct a price index of established residential dwellings (including land) that abstracts from compositional change. The method for constructing the ABS HPI has developed over time, but has always applied some form of stratification. The original approach was based on geographically stratified median prices and evolved into a stratification method based on geography and structural, locational and neighbourhood (SLN) characteristics. Hedonic methods have also been investigated. This paper examines the possibility of refining the stratification currently used to construct the ABS HPI. The key question to be addressed is whether a refined method of stratification can lead to improvements in the reliability of the index.

This paper starts by providing a brief history of the ABS HPI (Section 2), followed by a short description of the method used when constructing the HPI (Section 3). Section 4 outlines the approach used to refine the stratification and Section 5 discusses the quality assessment framework (QAF) applied to establish the effectiveness of various stratification results. Preliminary findings are discussed in Section 6 and the final section (Section 7) provides some concluding remarks and future research directions. The questions that we would like the Methodology Advisory Committee to particularly focus on are outlined on the first page of this document.

2. HISTORY OF THE ABS HOUSE PRICE INDEX

Since 1986, the ABS has compiled an established house price index for each of Australia's eight capital cities as well as an index at the national aggregate level (a weighted average of the city level indexes). The established HPI is compiled and published quarterly in *House Price Indexes, Eight Capital Cities* (ABS cat. no. 6416.0).

The HPI was originally designed to meet the specific data requirements for the construction of a price measure for mortgage interest charges, which were included in the Consumer Price Index (CPI) from 1986 to 1998. When mortgage interest was removed from the CPI in 1998 as a result of the 13th Series CPI Review, the ABS continued to publish the price index of established houses because of user interest in the series.

2.1 Geographic stratification

Initially, the HPI was based on a median measure of price, with a partial attempt to control for compositional change made by stratifying houses within each capital city by region. Depending on the size of the city, transactions were stratified into several geographic regions, with each region encompassing a statistical subdivision (SSD).[1] There were seven regions each for Sydney, Brisbane and Perth, six regions for Melbourne, five regions for Adelaide, four regions for Hobart and three regions each for Darwin and Canberra.

For Sydney, Melbourne, Brisbane and Adelaide, a 'trimean' method[2] was used in the calculation of prices used to construct region indexes. A simpler method[3] was used in the calculation of region price indexes for Perth, Hobart, Darwin and Canberra. Region prices and representative region weights – based on the number of established houses and their average value – were used to construct capital city indexes. The indexes for each of the eight capital cities were arithmetically weighted together[4] to form a national HPI.

---

1 A statistical subdivision (SSD) is a general purpose spatial unit of intermediate size in the Australian Standard Geographical Classification (ASGC). In aggregate, SSDs cover Australia without gaps or overlaps. SSDs are defined as socially and economically homogeneous regions characterised by identifiable links between the inhabitants (ABS 2005).

2 After 'extreme' or 'unrepresentative' prices in each region were removed from the sample, the prices were divided into three quantiles (representing low, medium and high valued houses). An unweighted average price was calculated for each quantile and aggregated to a regional price index through the use of a weighted arithmetic mean (the medium quantile was weighted by 0.5 and the low and high quantiles were weighted by 0.25 each).

3 After removing 'extreme' or 'unrepresentative' values, an unweighted average price was calculated for each region.

4 The city weights were estimated based on the value of secured (individual) finance commitments for the purchases of newly erected and established houses.

2.2 Hedonics

The surge in Australian house prices from 1999 until late 2003 engaged the close attention of policy makers and commentators on the HPI and its ability to indicate (or even predict) changes in economic activity. This heightened interest in the methodology for HPI construction, in particular the difficulty in controlling for quality and compositional changes, led the ABS to experiment with hedonic methods.

In 2003, the ABS undertook a study[5] to investigate the feasibility of applying hedonic methods to construct the HPI. The main motivation for the study was to test the efficacy of hedonic methods to account for the impact of housing attributes on the HPI. The analysis was based on data from Hobart and Adelaide only – data required for the application of hedonic methods to other cities were not available.

Results from the hedonic regressions were consistent with *a priori* expectations. Unfortunately, firm conclusions as to the usefulness of hedonic methods in HPI construction could not be drawn due to limitations caused by the data. In particular, the following issues remained unresolved:

- Only data from Hobart and Adelaide were available. Different results may be obtained when hedonic methods are applied to large cities (such as Sydney and Melbourne) where the situation (in terms of the housing market, socio-economic conditions and consumer preferences) is more complex.
- The data only covered a short period of time, when housing markets were booming everywhere. The results could be unique to this stage of the housing market cycle.
- A hedonic function depends on the availability of well-defined attribute variables. The attribute variables used in estimating the hedonic functions may require improvement in terms of coverage and definition.

5 Chen, Zhao, Romanis and Lim (2004).

2.3 Current stratification method

Policy maker and media interest in the HPI's ability to accurately measure Australian house price movements motivated the ABS to re-examine the construction of the HPI. Issues such as housing affordability, changing interest rates and growing private debt had increased the pressure for more accurate and timely measures of house price changes.[6] In 2004 the ABS initiated a review of the HPI, focussed on determining the best means of stratification given available data sources and looking into other data sources that could improve timeliness.

The main outcome of the review was the change in the objective of the HPI. Its focus became one of providing a more accurate measure of the contemporary rate of change in the prices of the stock of established houses.[7] The index was made more timely through the use of contract exchange rather than settlement to determine transaction dates, and more accurate by controlling as far as possible for compositional change by stratifying the house price observations (based on suburbs) by defining characteristics within a number of regions within each city.

House price (transaction) data was historically sourced from State and Territory Valuer General's Offices (VGs) and it took several months for all of the transactions relating to a particular quarter to be settled, recorded by the relevant land title's office and passed on to the ABS. This made it impossible to produce a timely HPI from this administrative data source, although it was the preferred source as far as coverage and comprehensiveness were concerned. However, it was found that loan approval documents from mortgage lenders could be used to produce preliminary estimates of the HPI. As a large percentage of house sales involve mortgages, loan approval documents created by mortgage lenders are a source of timely house price data and although they do not cover all house sales, they have sufficient coverage to be used as a means of estimating the preliminary movements in house prices.

Supplementing the VGs' data with loan approval data from mortgage lenders improved the timeliness of index. The VGs' data are used to compile the price index up to the point for which a complete set of data can be obtained on an exchange date basis. Substantially complete sets of data are available for all cities up to the quarter ending two quarters prior to the most recent quarter. The series based on this data is referred to as the benchmark series. The supplementary mortgage lenders' data

6 Various commentators criticised the timeliness and reliability of measures of house prices published by various organisations, including the ABS. There were several different measures of house prices published regularly, each based on different methodologies, scope, coverage and timing. Users were primarily concerned about the conflicting signals that the different measures were delivering and, in the case of the ABS series, its lack of timeliness and the point in the transaction cycle at which it recorded prices (the date of final settlement rather than contract exchange date – a timing difference of approximately two months).

7 Implementing the changes outlined in this section enabled the index to be published five weeks after the reference quarter instead of the previous nine-week lag.

combined with early VGs' data is used to project the HPI for the two most recent quarters to provide a more timely indicator of changes in house prices. As the full set of VGs' benchmark data becomes available, it is used to replace the leading indicator component. As a result, estimates of the HPI for the two most recent quarters are preliminary and subject to revision.

The current HPI is the result of stratification based on attributes that can be broadly defined as the structural, locational and neighbourhood (SLN) characteristics of suburbs. Analysis determined that four structural variables, four locational variables and one neighbourhood variable were the most relevant in determining the similarity of suburbs for stratification purposes.

Structural characteristics of each suburb were derived from the household file of the 2001 Census, summarised at the suburb level to create the following four variables capturing the general structural attributes of homes within each suburb:

- the percentage of dwellings in the suburb with two or fewer bedrooms;
- the percentage of dwellings with four or more bedrooms;
- the percentage of dwellings that are owner-occupied; and
- the percentage of dwellings in suburb that are houses.

Locational variables were obtained from data sets created by the Geography section of the ABS and include the distances of the midpoint of any given street to the CBD; the nearest shops; and the nearest hospital. Data at the street level were used to calculate the average distances of streets within suburbs to derive average distances for suburbs.

The 2001 Socioeconomic Index for Areas (SEIFA) was used to represent neighbourhood characteristics.[8] This ABS-produced index is the result of principal components analysis (PCA) on a wide set of variables from the 2001 Census that capture aspects of the socioeconomic status of areas, such as the proportion of families with high incomes, people with a tertiary education, and employees in skilled occupations.[9]

Visualising the solutions to cluster analysis is near impossible when dealing with more than three variables. In total, the above attributes offered nine variables for the clustering process. However, an important part of the cluster analysis is the ability to visualise the solution *post hoc* to assess its plausibility. PCA[10] was therefore used to reduce the non-SEIFA variables into two principal components, one each for the structural variables and the locational variables.

8 In particular, the Index of Relative Socioeconomic Advantage/Disadvantage was applied in the stratification.

9 For more information on SEIFA, please see ABS (2006a).

10 PCA finds a linear transformation of a set of variables so as to maximise the variance of the derived variable. It can be shown that doing so maximises the sum of the squared correlations between the derived variable and the original variables, resulting in a derived variable that is a close analogue of its constituents.

Cluster analysis was applied at the SSD level [11] to preserve the geographical homogeneity of suburbs and to produce robust groupings using the two principal components and SEIFA (in effect, three principal components) as the stratification variables. The outcome of this cluster analysis was the current method – the SLN stratification.

The HPI based on the SLN stratification is compiled using weights relating to the stock of established houses. The weights are expressed in terms of stock values (currently derived from the 2001 Census) with an initial value of the established housing stock in each cluster estimated by aggregating suburb counts to clusters and valuing them at March quarter 2002 mean prices. The ratio of the observed median price of each cluster for the current and previous quarter (price relative) is used to move forward the stock values for each cluster in each city.

The ABS first published the HPI based on the SLN stratification and the supplementary data from mortgage lenders in December 2005, with the series backdated to the March quarter of 2002. An assessment of the method's strengths and practical limitations has now been made and several key issues, such as the index not controlling adequately for compositional change, have been uncovered and require investigation. A recent paper published by the Reserve Bank of Australia (Prasad and Richards, 2006), on price based stratification (explained in the next section), has also paved the way for investigating potential improvements to the current stratification method.

2.4 Price stratification

In 2006, the RBA proposed a simple measure of house price growth that addressed the problem of compositional change by stratifying individual transactions based on house prices. The RBA suggested grouping suburbs into strata based on the variable most likely (on an *a priori* basis) to explain the price in any transaction – the long-term level of prices for the suburb in which the house is located. The RBA argued that a stratification based solely on this long term median house price of suburbs led to robust estimates of housing indexes, in the sense that they were as good as those resulting from more sophisticated measures. The price-based result was considered by some to be an improvement over the simple (unstratified) median approach and performed better in real time with limited data samples. The growth rates produced by the price based stratification also lined up closely and were highly correlated with more advanced regression-based measures. [12]

11 The stratification was restricted by the SSD level constraint so that clusters would be geographically contiguous (only suburbs within the same SSD could be grouped together and clusters could not cross SSD boundaries). This restriction was driven by wishing to publish the HPI at a lower than city level (which is currently not a viable option).
PCA was conducted at the SSD level – that is, different structural and locational principal components were constructed for each SSD.

12 For more information on regression-based measures, see Hansen (2006).

The price-based stratification method appears to have several advantages: it is operationally very simple to implement and, on the surface, achieves the statistical objective of minimising within-group variation in prices. However, it makes two implicit assumptions – first, that houses which are homogeneous in prices are homogeneous in reality, and second, that houses which are homogeneous in price levels are homogeneous in price *movements*.[13] On the surface, it might appear that the ABS method (SLN stratification) is making these assumptions as well. However, it is more reasonable to presume that suburbs which are grouped together on the basis of similarity in housing fundamentals, rather than prices, will have more similar price movements.

2.5 Refining the stratification

The SLN stratification method has been used to construct the HPI since the March quarter of 2002, so there has been plenty of time to assess its 'use-ability'. The process of compiling the HPI has brought to light a significant problem concerning the number of clusters per capital city. In some cities, there seem to be too few clusters and the sheer number of observations leads to a large range of values, affecting the volatility of the overall HPI (the distribution of transaction prices by cluster is irregular). In other cities, the stratification suffers from too many clusters and as a result, several clusters have too few sales observations in some quarters and are sensitive to outliers (the cluster medians are volatile and non-robust). These observations mean that the current method does not account well enough for compositional changes.

In light of the above and other developments, the current research described in this paper aims to improve the HPI stratification. We assume that stratification works and we are seeking an incremental improvement (in effect, we are refining the current stratification to better control for compositional change). It has been hypothesised that relaxing the SSD constraint and allowing stratification across SSD boundaries would go some way to improving compositional adjustment. It is hoped that stratification without the SSD restriction will lead to a more acceptable number of clusters in each city and that each cluster will contain a relatively broader grouping of similar suburbs. This would lead to a sufficient number of sales observations per

13 The price based stratification assumes that expensive homes in the inner-city, typically smaller in size and relatively older in age, will experience the same price movements as homes of equivalent value in newer, outer suburban areas. This is unlikely at a time when rising expectations of fuel costs will lead to increased premiums on homes located close to the city.
From a statistical perspective, the instability of the stratification over time will impound variability into the index, as no constraints are imposed to prevent suburbs from jumping between strata in subsequent years. There is also no guarantee that suburbs that have house sales in one period will have sales in other periods – such suburbs will appear in the price based stratification in one period and not in another, causing irregularities in the weighting scheme leading to index volatility.

stratum per quarter and result in stable stratum medians. Thus, the first step in improving HPI stratification is removing the SSD restriction.

The *second step* to improving the HPI involves exploring the inclusion of the long-term median house price of a suburb [14] as a stratification variable and examine its impact on the HPI. While the RBA approach of simple price based stratification is based on the premise of economic strata that control for compositional change, we hope to add value by also including socioeconomic characteristics as stratification variables and so improve the level of compositional adjustment.

The *third step* will include reviewing all of the variables in the current (SLN stratification) method. We will investigate the variables used in the cluster analysis and determine if better stratification results can possibly be achieved by using a simpler method (for example, without the application of principal components analysis).

The *final step* to improving HPI stratification will involve utilising better data sets as there have been significant improvements in the availability and coverage of data since the last stratification review was completed. The availability of more complete historical sales data may lead to better stratification through better geographical coverage and fewer suburbs being excluded from the analysis. The recent release of the 2006 Census data and SEIFA indexes also means that we will be in a position to update the stratification to better reflect the current situation and express the contemporaneous relationships between housing fundamentals and house price changes.

14 We define the long term price level indicator of a suburb as the mean of the quarterly median prices within the suburb over a four and a half year period (from March quarter of 2002 to June quarter of 2006).

3. METHODOLOGY FOR CONSTRUCTING A HOUSE PRICE INDEX

3.1 Applying stratification

We use cluster analysis to devise a stratification of the housing stock in order to maximise the homogeneity of houses within each stratum whilst also ensuring that a sufficient number of sales observations are available to defensibly construct price relatives. [15] Stratifying the housing stock is based on the idea that the next best alternative to comparing the price movements of identical houses is to compare the price movements of houses that are similar in their attributes. The stratification approach achieves this objective by combining suburbs into strata according to a set of fundamentals that determine house prices. The aim is to group houses that are similar (rather than identical) according to these attributes, to (1) maximise the level of homogeneity of the suburbs [16] contained in the same stratum, and (2) have a sufficient number of sales observations in each stratum in each reference quarter [17] to safely construct price relatives. Achieving this involves balancing a trade-off between these conditions: namely that the level of homogeneity achieved in the final solution is limited by the importance of having a sufficiently large number of sales observations to calculate price relatives. Individual houses could be grouped together on the basis of attributes to enable a comparison of like against like. However, the absence of adequate unit-level information, in addition to the computational burden, makes this approach not feasible.

A compromise to the above issue is to treat larger areas, for which we have sufficient information, as the building blocks for the stratification. The stratification for the HPI is therefore based on clustering suburbs. [18] The success of the clustering algorithm is contingent upon choosing the right attributes to group the suburbs, as these attributes constitute a frame of reference to establish the clustering and as such, they must relate to the classification being sought. Choosing irrelevant variables can lead to clusters that do not distinguish between the objects being clustered and including too many variables may obscure the cluster structure.

15 Cluster analysis is a method that uses mathematical algorithms to devise classifications of units based on a set of rules that aim to maximise the degree of homogeneity within each classification. The Appendix explains the theory of cluster analysis.

16 In our case homogeneity is achieved by reducing the range of house prices in each stratum, and hence decreasing the variation of medians over time.

17 In our case a sufficient number of sales is 100 transactions per quarter.

18 This involves the implicit assumption that houses within a suburb are homogeneous. Naturally, using any large area as a unit for clustering involves an implicit assumption that homes within the area are homogeneous, an assumption that becomes increasingly tenuous as the size of the area increases.

In this instance, the objective is to group suburbs that have similar price levels and price movements in order to stabilise the city-wide movements over time and capture the pure price evolution of the housing stock. The variables that are chosen must therefore bear a strong relation to house prices. The number of groups and the properties of the groups are exogenously determined: the clustering algorithm groups objects together into a pre-specified number of groups, using the attributes provided to minimise the error sum of squares in the sample. Adjustments to the solution may be necessary to ensure that other desirable conditions, which are not (or cannot be) included in the clustering process, are satisfied. In this case, we want stratification results that minimise the range of prices in each stratum, eliminating the influence of compositional change and giving stable and reliable summary price measures based on a sufficient number of observations.

3.2 Applying the price index formula

Traditionally, the HPI is calculated using the Laspeyres price index formula. The formula can be written as:

$$I_t^L = \frac{\sum_{i=1}^{n} p_{it} q_{i0}}{\sum_{i=1}^{n} p_{i0} q_{i0}} \times 100 \quad = \quad \sum_{i=1}^{n} w_{i0} \times \left(\frac{p_{it}}{p_{i0}} \right) \times 100 \qquad (1)$$

where

p_{it} is the price in cluster i, t quarters after the base period;

p_{i0} is the price in cluster i during the base period;

q_{i0} is the Census count of houses in cluster i in the base period; and

w_{i0} is the value expenditure weight used to weight the price relatives for each cluster to form the Laspeyres Index at higher geographical levels, calculated as:

$$w_{i0} = \frac{p_{i0} q_{i0}}{\sum_{i=1}^{n} p_{i0} q_{i0}} \qquad (2)$$

The price relatives are based on median prices because they are considered to be the best measure of central tendency and are less affected by outliers than other summary measures, such as means.

Using a Laspeyres price index for the measurement of house price inflation is slightly different to its use in constructing the consumer price index (CPI). In this instance, we are not deriving a weighted average of the price relatives of multiple commodities over time, but rather a weighted average of the price relatives of multiple clusters over time. The clusters correspond to the homogeneous groups of suburbs formed as a consequence of applying the stratification procedure.

The prices (p_{it}) used in the calculation are obtained from the sales data (transaction data from the Valuer General's Department) and are combined with Census information on house counts (q_{it}) to construct the HPI. It is important to realise that in order to preserve the Laspeyres index form, a weighted average of price relatives must involve the use of stock value weights rather than quantity or activity weights. [19]

19 Activity weights are based on the number of houses sold in a reference quarter, rather than the actual number of houses in the housing stock (quantity weights) or the total estimated value of the housing population (stock value weights). The danger of using activity-based weights is that the effects of compositional change can be built into the index by placing a greater weight on those areas that have high sales volume than areas with low sales volume. Stock value weights based on Census counts, on the other hand, remain constant until the stratification is revised with the release of new census data. One drawback of this strategy is that the value of housing used for each stratum can quickly become outdated, leading to increasingly unrepresentative weightings as time passes.

4. APPROACH TO REFINING THE STRATIFICATION

Our work on refining the stratification is largely based on improving the current method used in constructing the HPI. The current method is the result of stratification based on structural, locational and neighbourhood (SLN) principal components (PCs) restricted by a geography level constraint, such that only suburbs within the same statistical subdivision (SSD) are grouped together. Based on the developments outlined in Section 2.5, the process of refining the stratification explores four avenues:

- removing the SSD level restriction;
- adding the long-term median house price of a suburb as a stratification variable;
- determining if better stratification results [20] can be achieved by using a relatively simpler stratification method; and
- exploiting the availability of superior data. [21]

Based on feedback from the HPI team (the ABS staff responsible for compiling the HPI every quarter), our analysis of stratification methods is carried out using the benchmark [22] transaction data and the choice of the number of clusters in each city is guided by the direction provided by the HPI team, as outlined in table 4.1.

4.1 Current and suggested numbers of clusters for each capital city

	Number of clusters	
City	*Current*	*Suggested*
Sydney	55	fewer
Melbourne	39	fewer
Brisbane	51	fewer
Adelaide	27	about right
Perth	14	more
Hobart	8	fewer
Darwin	5	more
Canberra	14	fewer

20 A stratification result is based on the choice of a particular stratification method and a specific number of clusters.

21 The data are superior in the sense that transaction data is more complete (has better geographic coverage) and recent data releases (of the 2006 Census and SEIFA) mean contemporaneous stratification results are possible.

22 Benchmark data refers to the substantially complete sets of data on an exchange date basis available from the VGs.

Our analysis is also constrained by the availability of potential stratification variables. Utilising available data sources, the variables considered as candidates for inclusion in the stratification method are listed in table 4.2.

4.2 Potential stratification variables

STRUCTURAL:

Percentage of dwellings in the suburb that are:
- Houses
- Townhouses
- Units

Percentage of dwellings in the suburb that have:
- 2 or fewer bedrooms
- 3 bedrooms
- 4 or more bedrooms

Percentage of dwellings in the suburb that are:
- Owned
- Rented

LOCATIONAL:

Distance from the centroid of the suburb to the nearest:
- Central business district (CBD)
- Shops
- Hospital
- School (primary or high or combined) or tertiary institution

NEIGHBOURHOOD:

Socio-Economic Index for Areas – Index of Relative Advantage/Disadvantage

LONG-TERM MEDIAN HOUSE PRICE:

Mean of the quarterly median prices within the suburb over a four and a half year period

The analysis used to refine the HPI stratification has two distinct paths: one based on extending the known methodology and the other on investigating new methodology. Extending the known methodology includes removing the SSD restriction (stratifying all suburbs within a capital city based on the SLN principal components (PCs)) and adding the suburb long-term median house price as a stratification variable. Investigating the new methodology includes examining simple stratifications based on combinations of the SLN variables without applying PCA. The variables include: the suburb level SEIFA score (the Index of Relative Advantage/Disadvantage), the percentage of dwellings within the suburb that have three bedrooms, the percentage of dwellings within the suburb that are owned and the suburb's long-term median house price.

All in all, there are seven different stratification methods to analyse:

- applying PCA at the city-level (rather than at the SSD level as for the current stratification) to create structural and locational PCs [23], then stratifying suburbs based on the three PCs (SEIFA is itself a PC);
- applying PCA at the city-level to create the structural and locational PCs and then stratifying suburbs based on the three PCs plus the suburb's long-term median house price;
- stratifying suburbs based on their SEIFA score only;
- stratifying suburbs based on their long-term median house price only;
- stratifying suburbs based on their SEIFA score and long-term median house price;
- stratifying suburbs based on their SEIFA score, percentage of three bedroom dwellings and percentage of owned dwellings; and
- stratifying suburbs based on their SEIFA score, percentage of three bedroom dwellings and percentage of owned dwellings and long-term median house price.

23 The structural PC based on the percentage of dwellings within the suburb that have: two or fewer bedrooms; four or more bedrooms; are owned; and are houses. The locational PC is based on the centroid distance of the suburb to: the CBD; the nearest hospital; and the nearest shops.

5. QUALITY ASSESSMENT FRAMEWORK

We began our analysis with a list of the seven stratification methods to be examined and a vague idea about the 'right' number of clusters for each city. To guide our decision making process on the way to choosing the most effective stratification method and the optimal number of clusters, we applied a three-phase Quality Assessment Framework (QAF) and supported it with sensitivity analysis.

The first phase of the QAF looked at goodness of fit statistics at the city level and allowed us to determine the (approximately) appropriate number of clusters for each method in each city. [24]

The second phase examined each of the stratification methods combined with the appropriate numbers of clusters to determine if they fulfilled our criteria – namely, that the stratification (1) maximised the level of homogeneity of the suburbs contained in the same stratum, and (2) led to a sufficient number of sales observations in each stratum in each reference quarter to safely construct price relatives. These criteria were assessed through the use of quality dials – visual representations of cluster level information. Homogeneity was assessed by using boxplots to examine the distribution of transaction prices in each cluster every quarter, while the sufficient number of sales criterion was assessed based on histograms of the sales counts in each cluster every quarter. Stratification results (combinations of a particular stratification method and an appropriate number of clusters) that fulfilled these criteria were then subjected to the final phase of the QAF.

The third phase of assessment allowed us to look at detailed cluster level information – examining histograms of the frequency distributions of transaction prices (density plots) at the cluster level by quarter. Such in-depth analysis allowed us to closely examine if the compositional adjustment we were seeking was actually being achieved by the stratification results.

Throughout this three phase process, we applied mocked-up versions of the city-level HPIs to examine the sensitivity of the index results to the choice of stratification method and number of clusters. This sensitivity analysis allowed us to determine if our efforts at improving the stratification were actually having a visible impact on the final result.

24 Numerous statistical tests exist to determine the optimal number of clusters, however these do not factor in the requirement that each cluster must have a sufficient number of sales observations in every quarter. This is why such tests (R^2, *RMSSTD*, pseudo-*F* and pseudo-t^2) were only used to decide upon an initial number of clusters and subsequent analysis was used to satisfy the aforementioned constraint.

At the end of this labour-intensive and time-consuming process, we could finally conclude which stratification method (and how many clusters for each capital city) would lead to the most efficient results when constructing the HPI. The final step before implementing the most efficient stratification result requires refining the stratification results to account for excluded [25] and problem [26] suburbs. This final step ensures that irregularities in the stratification result do not confound the final index.

5.1 Goodness of fit measures

5.1.1 R^2

We have established that the desired outcome of the stratification process is to produce fine groupings of suburbs in order to remove the volatility in medians that results from changing compositions in strata that are too broad. The effectiveness of any stratification result can therefore be measured by its ability to remove as much variation in prices that is explained by the heterogeneity of homes in the sample as possible. To measure how well a particular stratification result does this, we calculated an R^2 value [27], which measures the proportion of the total variation in prices explained by the stratification.

Let x_i be the i-th observation ($i = 1, \ldots, n$), C_k be the set of observations belonging to cluster k i.e. $x_i \in C_k$ for $k = 1, \ldots, K$, $\bar{x}$ be the mean of all observations, and $\bar{x}_k$ be the mean of cluster k. Then:

$$\begin{aligned} W_k &= \sum_{i \in C_k} \left\| x_i - \bar{x}_k \right\|^2 \\ W_K &= \sum_{k=1}^{K} W_k \\ T &= \sum_{i}^{n} \left\| x_i - \bar{x} \right\|^2 \end{aligned} \qquad (3)$$

and

$$R^2 = 1 - \frac{W_K}{T} \qquad (4)$$

25 Excluded suburbs arose because of incomplete data – some suburbs were missing SEIFA scores due to low populations at the CD level, while others had unstable long term median price levels (where unstable was defined as based on fewer than 15 observations over the 4.5 year period and having high variability in levels).

26 Problem suburbs are those that were frequently assigned to their own clusters (due to their uniqueness in terms of the stratification variables) or formed clusters with very few other suburbs and consequently failed to meet the homogeneity and sufficient sales criteria.

27 R^2 values are a good measure of the homogeneity achieved by the clustering solution, as are two-way scatter plots of the variables to show whether suburbs within the same cluster are grouped together and whether clusters of suburbs are noticeably differentiated over the variable dimensions used.

R^2 is a measure of the amount of variation in the stratification variables explained by the current stratification result, which means that stratification results with an R^2 closer to 1 are more effective. Note ,however, that as the number of clusters increases, R^2 approaches 1 and so we cannot use R^2 alone to judge the quality of the clustering or to provide us with an optimal choice of the number of clusters. However, we used the R^2 to compare different stratification methods for a given number of clusters (note: when comparing between methods, differences will occur due to the stratification variables used). To guide the choice of the optimal number of clusters, we plotted R^2 against the number of clusters. We chose to use a threshold value of R^2 to decide on the optimal number of clusters: for example, pick the number of clusters to be such that $R^2 > 0.95$. However, when plotting R^2 values against the number of clusters, there is often a noticeable plateau; the beginning of this plateau may also be used as a guide to the choice of the optimal number of clusters.

5.1.2 RMSSTD

The Root-Mean-Square Standard Deviation (*RMSSTD*) provided us with a measure of volatility for a particular stratification method and a given number of clusters.

$$RMSSTD_k = \sqrt{\frac{W_k}{\upsilon(n_k - 1)}} \tag{5}$$

where υ is the number of stratification variables and n_k is the number of observations in cluster k.

The calculation is based on the most recently formed cluster, giving an indication of the change in volatility when the number of clusters decreases by one through the fusion of two observations. We plotted *RMSSTD* values against the number of clusters to give an indication of how well the stratification is performing. When looking at a particular stratification method, we preferred to choose the number of clusters that has a low *RMSSTD* and when comparing between stratification methods, we preferred those in which *RMSSTD* approaches zero the fastest.

5.1.3 Pseudo-F and pseudo-t^2 statistics

The pseudo-F statistic measures the performance of a stratification method with a given number of clusters (K) – giving us an idea of the separation among the clusters.

$$pseudo-F = \frac{T - W_K}{\upsilon(K-1)} \Big/ \frac{W_K}{\upsilon(n-K)} \tag{6}$$

Once again, we informed our choice of the optimal number of clusters by plotting the values of pseudo-F against the number of clusters. As pseudo-F is a ratio of the 'between cluster error' and the 'within cluster error' we look for a (locally) large value of the statistic – a maximum (or peak) on the pseudo-F plot.

Pseudo-t^2 gives a measure of the performance associated with joining clusters k and l at a given stage of the clustering process. The calculation is based on the two most recently joined clusters and provides us with a measure of the separation between them.

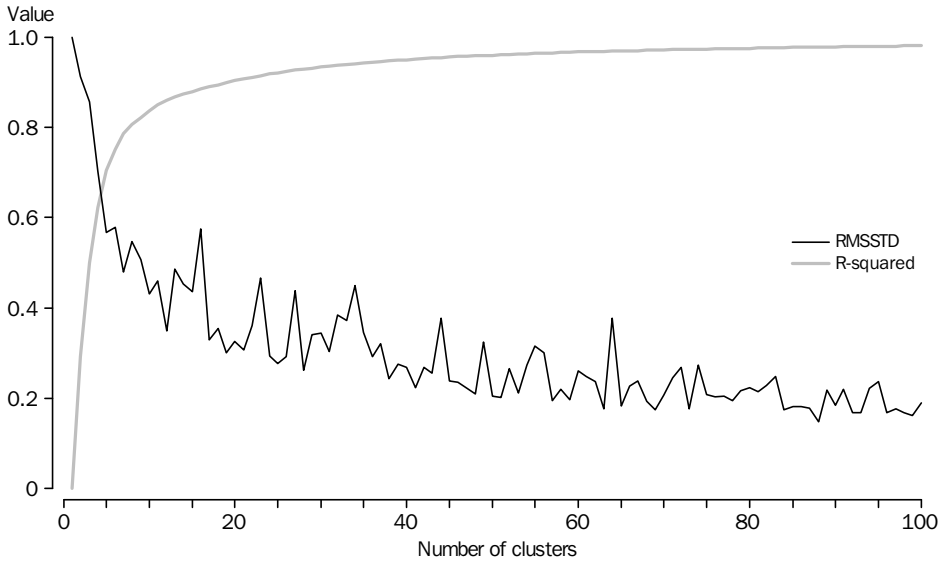
$$pseudo-t^2 = \frac{B_{kl}}{(W_k + W_l)/(n_k + n_l - 2)} \tag{7}$$

$$B_{kl} = W_m - (W_k + W_l) \tag{8}$$

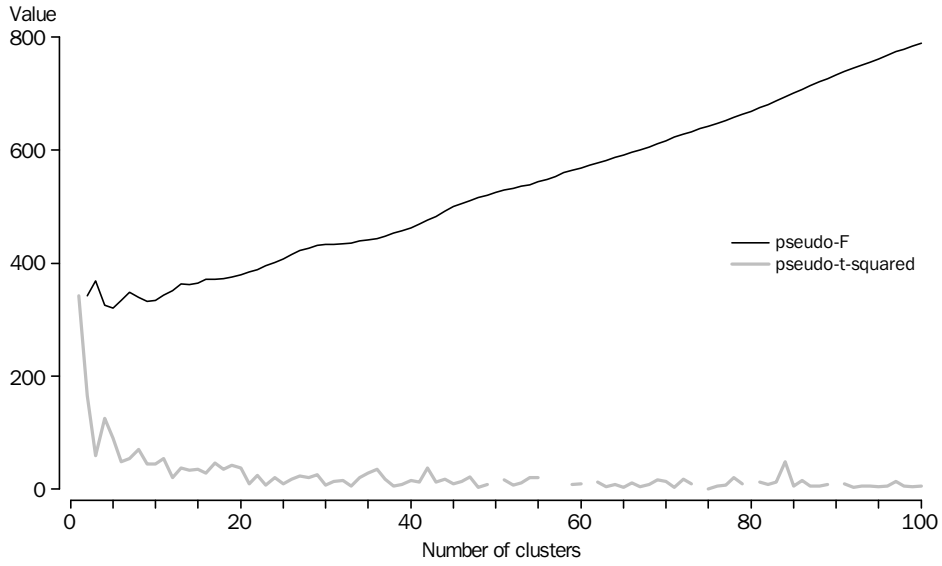
We plotted the pseudo-t^2 values associated with each level of the clustering process against the number of clusters to give us an idea of the optimal number of clusters. To choose the optimal number of clusters, we looked for a large (positive) change in the statistic as the number of clusters decreases.

When analysing stratification results it is important to use all of the available information and apply all of the tools described above to decide on the most effective stratification method and choose the optimal number of clusters. Looking at figures 5.1 and 5.2 we can see the difficulty in deciding on an optimal number of clusters (ignoring the method choice altogether). The R^2 and *RMSSTD* plots in figure 5.1 suggest that 20 clusters may be appropriate, whilst the pseudo-F and pseudo-t^2 plots in figure 5.2 suggest that fewer than twenty clusters may be appropriate – perhaps five or six.

5.1 Example plot of R^2 and RMSSTD values against number of clusters



5.2 Example plot of pseudo-F and pseudo-t^2 values against number of clusters

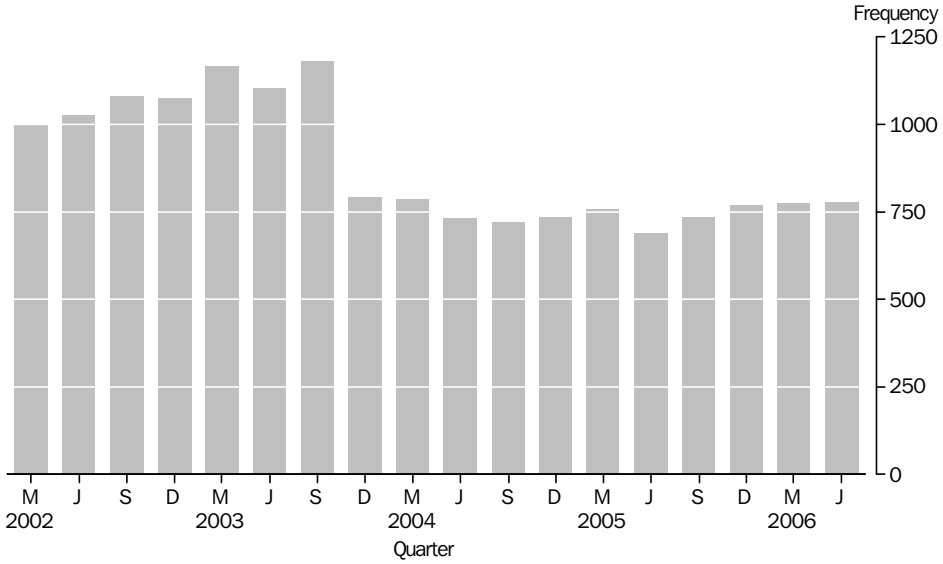


5.2 Quality dials

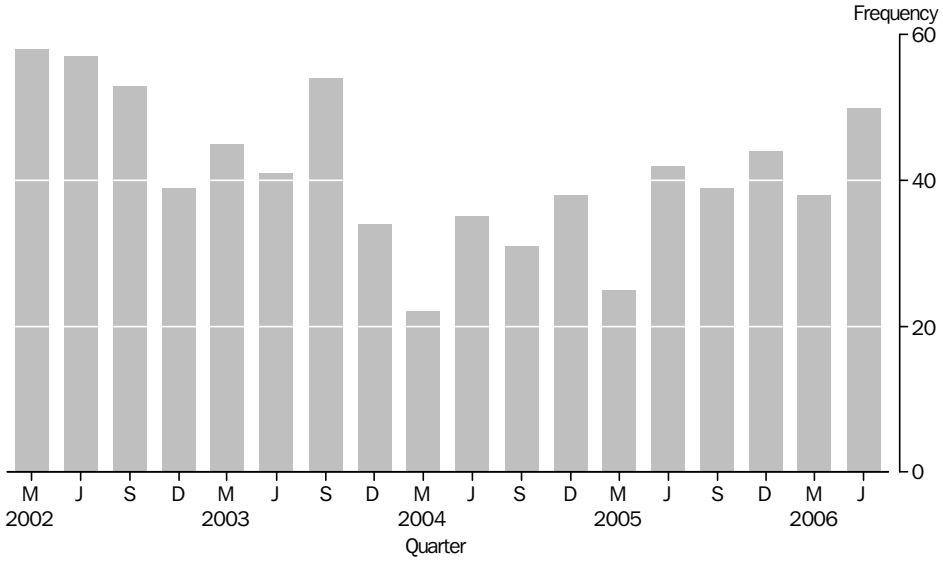
We constructed quality dials – visual representations of cluster level characteristics – to assess whether a particular stratification result fulfils the homogeneous price and sufficient number of sales criteria. Examining cluster level transaction histories (counts of house sales per quarter) allowed us to determine if the stratification result leads to clusters that lend themselves to constructing robust median price measures by satisfying the sufficient number of sales criterion. Specifically, we wanted to ensure that each cluster had at least 100 sales observations in every quarter. Looking at boxplots of transaction prices at cluster level over time allows us to investigate the distributional properties of strata and establish whether the variation in medians and ranges indicates homogeneity. In particular, we looked for similar distributions of prices within a cluster over time – most easily seen by focussing on the five number summary statistics – and consistent changes in levels from quarter to quarter.

Figures 5.3 and 5.4 give examples of 'acceptable' and 'unacceptable' sales histograms – used for assessing the sufficient number of sales criterion. In the 'acceptable' example, the cluster has at least 100 sales per quarter, while in the 'unacceptable' example, there are not enough sales per quarter (note that this is an extreme case – often, clusters will have approximately 100 sales early in the period, but the number of sales drops off over time).

5.3 Example of an acceptable histogram

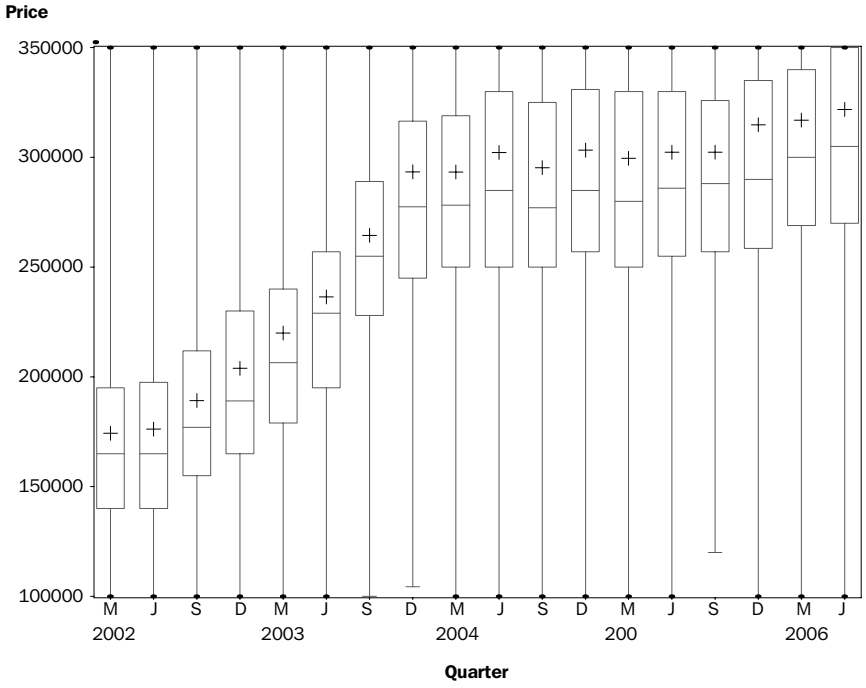


5.4 Example of an unacceptable histogram

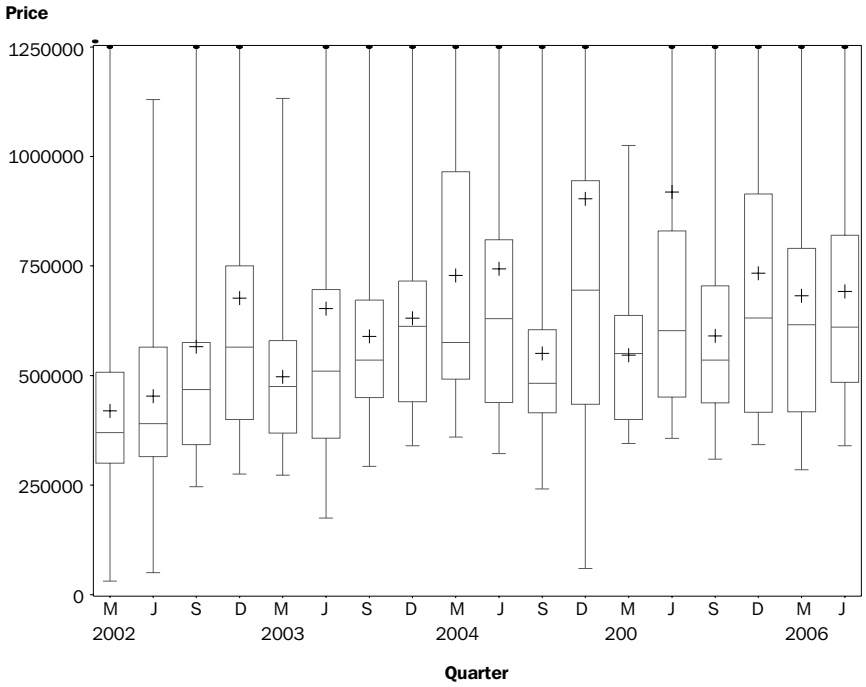


Figures 5.5 and 5.6 give examples of 'acceptable' and 'unacceptable' transaction price boxplots, used for assessing the homogeneity criterion. The 'acceptable' example is characterised by almost symmetrical price distributions from quarter to quarter, stable median prices and a visible trend over time. On the other hand, the 'unacceptable' example has volatile and asymmetric price distributions, very unstable median prices and no visible trend.

5.5 Example of acceptable boxplots



5.6 Example of unacceptable boxplots



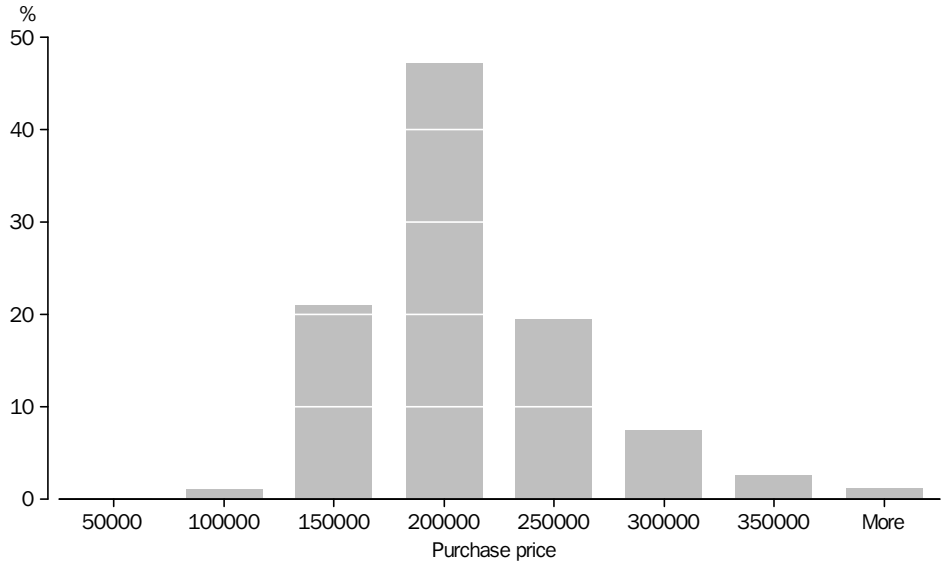
5.3 Density plots

As a quality assessment tool, density plots provide the most detail about cluster level characteristics. The study of histograms showing the frequency distributions of transaction prices within a cluster over time is a tool often employed by the HPI team when constructing the HPI. Such detailed analysis allows a thorough investigation of price change within a cluster over time to determine if the stratification has sufficiently adjusted for compositional change. An effective stratification result leads to regular distributional properties at the cluster level over time.[28] Specifically, 'regular' properties exclude distributions that are highly skewed, bi-modal, or have large spreads (evidenced by outliers).

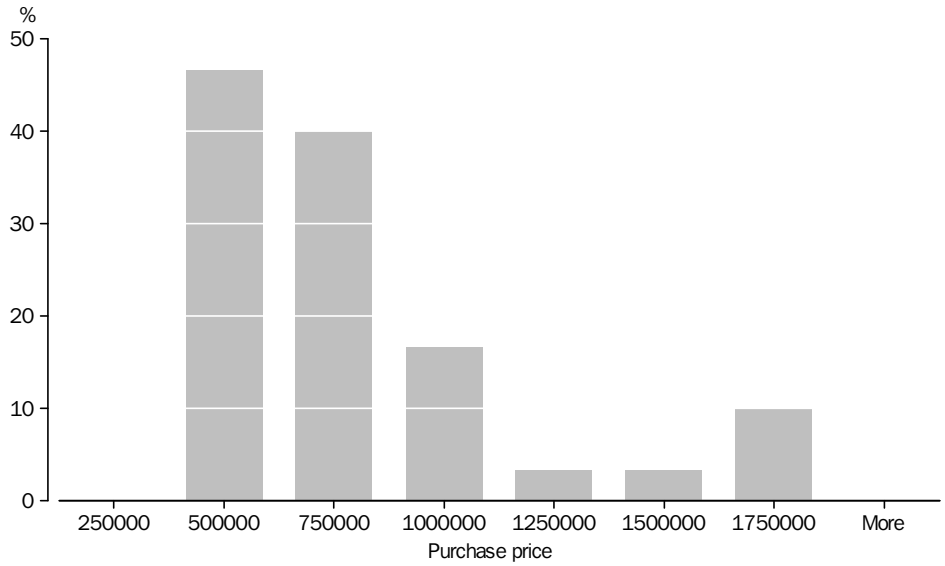
Figures 5.7 and 5.8 give examples of 'acceptable' and 'unacceptable' density plots – used to determine whether compositional change is still affecting price change. The 'acceptable' density plot clearly shows an almost normal distribution with a slight right skew, while the 'unacceptable' density plot suffers from a bi-modal distribution and quite large outliers. It is also important to note that volatile changes in distributions over time (i.e. clusters which chop and change between regular and irregular distributions from quarter to quarter) are indicative of unacceptable cluster characteristics and poor performance of the chosen stratification method.

28 In effect, we seek to isolate various homogenous sub-populations in the housing stock.

5.7 Example of an acceptable density plot



5.8 Example of an unacceptable density plot



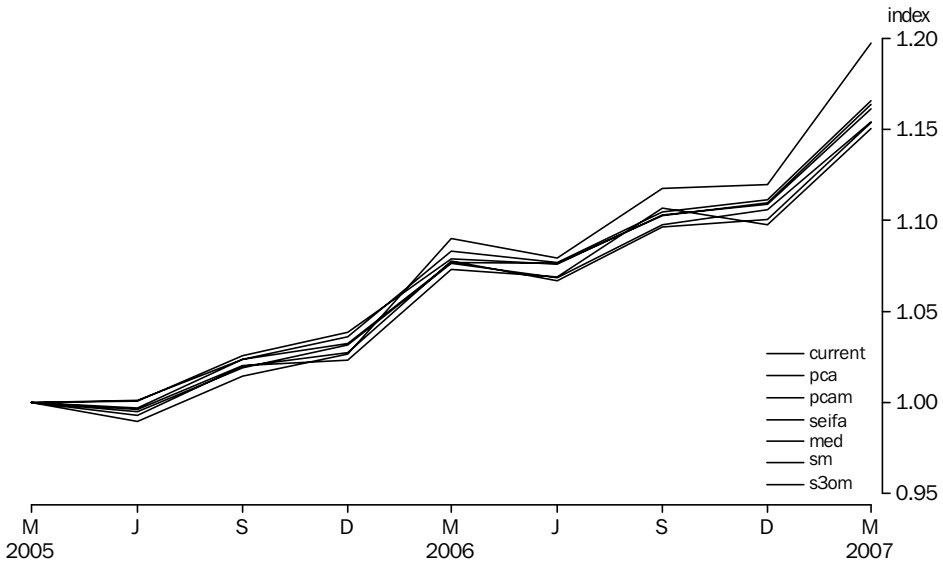
5.4 Indexes for sensitivity analysis

An investigation into the sensitivity of the HPI to various stratification methods was carried out by constructing mock city-level indexes based on the competing stratification methods and differing numbers of clusters. Unlike the other tools in the QAF, the indexes for sensitivity analysis use the most recent transaction data – from June quarter 2005 to June quarter 2007 – to gain a contemporaneous appraisal of the effect that refining the stratification has on the final result. As a means for comparing competing stratification methods and varying number of clusters, current clusters (based on SLN stratification) and current cluster weights (based on stock values

derived from the 2001 Census and initial values based on March quarter 2002 mean prices) were used to construct an unedited benchmark index.[29] The competing stratification results were based on both the known and new stratification methodologies – all seven methods and various numbers of clusters can be compared. The weights for these competing stratification results used stock weights derived from the 2006 Census and values based on the median cluster price in June quarter 2005.

The indexes allow for a comparison of price movements between stratification results over time. It is hoped that by better controlling for compositional change, the competing stratification methods will result in a smoother trend than the benchmark index. The important questions to be answered by this quality assessment tool include: do the benchmark and competing indexes follow each other closely? Do individual methods or particular numbers of clusters give significantly different results in terms of index movements over time? Figure 5.9 gives an example plot of competing stratification methods (all with the same number of clusters) – it is clear that many of the competing methods give quite similar results.

5.9 Example of index plots



29 As a consequence, this benchmark index does not correspond to published HPI data.

6. PRELIMINARY FINDINGS

The process of refining the stratification for constructing the HPI is well under way, but has not yet been concluded. The first two phases of the QAF have been completed and we are in the process of examining detailed cluster level information to determine which stratification method and what number of clusters will be the most effective at controlling for compositional change. Our preliminary analysis is supported by the use of mock indexes for sensitivity analysis and has been compared to the results obtained from applying the QAF to the current stratification. The assessment results for the current clusters are a used as a tool for comparison, to see if our refinement of the stratification method and alternative choices of the number of clusters are actually an improvement on the current method. The results for this preliminary analysis are obtained without the use of post-stratification editing which is routinely implemented by the HPI team when compiling the index for publishing.

According to the first phase of the QAF, we investigated the goodness of fit statistics for each of the proposed stratifications (see Section 4) to determine an approximate number of clusters for each city. The statistics were not unanimous in their suggestion of the optimal number of clusters per city for each method – the results were occasionally contradictory and the pseudo-F and pseudo-t^2 plots were often unable to indicate an optimal choice of the number of clusters. However, as Table 4.1 suggests, the number of clusters proposed by the statistics for each method was in fact lower than that which is currently used in each city. For example, the statistics indicated that Sydney (which currently has 55 clusters) should have:

- 10 or 18 or 20 clusters when stratifying suburbs based on their SEIFA score only;
- 10 or 15 clusters when stratifying suburbs based on their long-term median house price only; and
- 15 or 18 clusters when stratifying suburbs based on their SEIFA score and long-term median house price.

The appropriate number of clusters for each method in each city was then subjected to the second phase of the QAF, involving an examination of quality dials for each stratification result-by-city combination. These quality dials suggested that methods based on PCA did not control for compositional changes as well as the other methods (even with the addition of the long-term median house price of a suburb as a stratification variable). The methods did not result in sufficient numbers of sales observations per cluster and suffered from unstable median prices within clusters. Of the remaining methods, those that included the SEIFA score and the percentage of three bedroom dwellings as stratification variables did not perform as well as the others.

Resulting from this assessment, only stratifications based on the SEIFA score of a suburb, the long-term median house price of a suburb, and the SEIFA score and long-term median house price of a suburb were subjected to the third phase of the QAF. Examination of the in-depth distribution of sales prices within each cluster by city identified the stratifications based on variables that included the long-term median house price of a suburb as having the most regular properties. Further, when we looked at changes in the estimated densities over time, the stratifications based on the SEIFA score and long-term median house price of a suburb were the most stable, indicating acceptable cluster characteristics.

Underpinned by the three phases of the QAF, the supporting evidence provided by the indexes constructed for sensitivity analysis, and the significant improvement in the characteristics of the proposed stratification method over the current stratification method, we were able to conclude that stratification based on the SEIFA score and long-term median house price of a suburb was the most efficient for constructing the HPI. Based on this decision regarding the optimal stratification method and the preliminary analysis carried out thus far, table 6.1 outlines the proposed optimal numbers of clusters for each capital city.

6.1 Proposed number of clusters for each capital city – stratifying suburbs based on their SEIFA score and long-term median house price.

City	*Proposed numbers of clusters*
Sydney	15 or 18
Melbourne	13 or 17
Brisbane	15
Adelaide*	N/A
Perth	15
Hobart	9 or 12
Darwin	8 or 10
Canberra	13 or 16

* At this stage, complete data for Adelaide are not available, and as a result no stratifications have been performed.

7. CONCLUDING REMARKS AND FURTHER RESEARCH

The investigation into refining the stratification used for constructing the established HPI is still in progress. Preliminary findings based on applying the three-phase QAF and supported by index sensitivity analysis indicate that more accurate measurement of the pure price evolution of the housing stock is possible. Instead of applying the current stratification based on attributes that can be broadly defined as the structural, locational and neighbourhood characteristics of suburbs, the framework puts forward a simpler stratification method – based on the long-term median price and the neighbourhood characteristics of a suburb. The preliminary findings also suggest that it may be more appropriate to compile the HPI based on a significantly smaller number of clusters than is currently used for each capital city.

At present the scope of the HPI is limited to detached houses located in Australia's capital cities. However, key users of the HPI have indicated interest in an expanded scope. Of priority is extending the coverage of dwellings to include apartments, units and townhouses. The ABS plans to investigate the feasibility of this extension in the near future. Extension of the coverage of the HPI to regional cities and rural areas is also of interest, but of a lower priority.

REFERENCES

Australian Bureau of Statistics (2005) *Australian Standard Geographical Classification*, cat. no. 1216.0, ABS, Canberra.

—— (2006a) *Information Paper: An Introduction to Socio-Economic Indexes for Areas (SEIFA)*, cat. no. 2039.0, ABS, Canberra.

—— (2006b) *House Price Indexes, Eight Capital Cities*, cat. no. 6416.0, ABS, Canberra.

—— (2006c) *A Guide to House Price Indexes*, cat. no. 6464.0, ABS, Canberra.

Branson, M. (2006) *The Australian Experience in Developing an Established House Price Index*, OECD–IMF Workshop on Real Estate Price Indexes, 6–7 November, Paris.

Chen, L. and Samy, L. (2004) *Controlling for Compositional Change in HPI Measurement Through Improved Stratification*, Internal paper, Australian Bureau of Statistics.

Chen, L.; Zhao, S.; Romanis, P. and Lim, P.P. (2004) "Exploring Hedonic Methods for Constructing a House Price Index", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.067, Australian Bureau of Statistics, Canberra.

Everitt, B.S. (1993) *Cluster Analysis*, Halstead Press, New York.

Hansen, J. (2006) "Australian House Prices: A Comparison of Hedonic and Repeat-sales Measures", *Research Discussion Paper 2006-03*, Reserve Bank of Australia.

McCarthy, P.; Branson, M. and King, M. (2005) *Information Paper: Renovating the Established House Price Index*, cat. no. 6417.0, Australian Bureau of Statistics, Canberra.

Milligan, G. And Cooper, M. (1985) "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, 50(2), pp. 159–179.

Olczyk, A. and Neideck, G. (2007) *Measuring House Price Movements: Methods, Issues and Some Recent Experience in the Australian Context*, Real Estate and Housing Indices Session of the 10th Meeting of the UN International Working Group on Price Indices (Ottawa Group), 9–12 October, Ottawa.

Prasad, N. and Richards, A. (2006) "Measuring Housing Price Growth: Using Stratification to Improve Median-based Measures", *Research Discussion Paper 2006-04*, Reserve Bank of Australia.

APPENDIX – THEORY OF CLUSTER ANALYSIS

Everitt (1993) gives the following definition of cluster analysis:

> "Given a collection of n objects [transaction prices]... each of which is described by a set of p characteristics or variables [suburb level attributes], derive a useful division into a number of classes [or clusters]. Both the number of classes and the properties of the classes are to be determined."

The goal of cluster analysis is to partition a data set of N objects into subgroups such that those in each particular group are more similar to each other than those of other groups. If we define the encoder function $c(i)$, which maps each object i to a particular group G_l $(1 \leq l \leq L)$, as

$$c(i) = l \Rightarrow i \in G_l$$

we can formalise the goal of cluster analysis as being the identification of the optimal encoder $c^*(i)$ which minimises a criterion $Q(c)$ that measures the degree to which the goal is not being met, i.e.

$$c^* = \arg \min_{c} \{Q(c)\}$$

To do this, attributes of the objects must first be specified in order to quantify $Q(c)$. Suppose each object i has n attributes

$$\underset{\sim}{X}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$$

then we can specify our criterion as a function of each object's (N in total) vector of attributes:

$$Q(c) = F(X_1, X_2, ..., X_N)$$

Many different clustering techniques exists to achieve this aim, where each differs according to how the criterion $Q(c)$ is defined. Ward's minimum variance technique was used in our analysis on the basis of its favourable recommendation in the literature as a robust method and for its conceptual simplicity. At each iteration, the algorithm considers the union of every possible pair of clusters and combines the two clusters whose fusion results in the minimum increase in the information loss ($Q(c)$).

The loss function used is the error sum of squares, defined as the sum of the squared differences between each object's attribute and the within-group mean for that attribute:

$$Q(c) = \sum_{l=1}^{L} \sum_{j=1}^{n} \sum_{i=1}^{N_l} (x_{ijl} - \bar{x}_{jl})^2$$

where

$\bar{x}_{jl}$ is the mean value of attribute j in group l;

x_{ijl} is the value of attribute j for object i;

L is the predetermined total number of groups;

N_l is the number of objects in group l;

n is the number of attributes used for the clustering;

i refers to individual object; and

j refers to one of the n attributes.

The clustering algorithm groups objects together into a pre-specified number of groups, using the attributes provided to minimise the error sum of squares in the sample.

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au