Reserve Bank of New Zealand
Analytical Notes

Evaluating alternative monthly house price measures for New Zealand

AN2017/02

Jed Armstrong, Ashley Dunstan, and Tobias Irrcher

March 2017

Reserve Bank of New Zealand Analytical Note Series
ISSN 2230-5505

Reserve Bank of New Zealand
PO Box 2498
Wellington
NEW ZEALAND

www.rbnz.govt.nz

RESERVE BANK
OF NEW ZEALAND
TE PŪTEA MATUA

NON-TECHNICAL SUMMARY[1]

This paper outlines the production of three monthly house price indices (HPIs) for New Zealand produced using data from the Real Estate Institute of New Zealand (REINZ) using three alternative methodologies. REINZ approached the Reserve Bank of New Zealand at the end of 2015 for technical guidance on possible improvements to their house price index methodology, in light of significant improvements to their dataset in recent years. The paper documents the guidance, providing an overview of the alternative methodologies and an empirical evaluation of the resulting indices.

The database provided by REINZ is a rich unit-record sales dataset with information on price, location, valuation, and property characteristics (such as the number of bedrooms and the floor area). We use the database to produce HPIs based on three well-established and widely adopted methodologies: 1) sales-price to appraisal ratio (SPAR); 2) hedonic regression; and 3) repeat sales. All three methods are found to produce credible-looking indices, which match the turning points and well-established cyclical properties of New Zealand's existing house price statistics.

As a benchmarking exercise, the three candidate indices are evaluated alongside a simple median and a stratified median index (similar to the methodology currently used by REINZ). Applying a range of criteria to assess index performance, we find that all three alternative candidate methodologies out-perform the simple median and the stratified median methodologies.

The SPAR method is found to perform the best, due to lower month-to-month noise (especially for more disaggregated regional indices), greater stability as more data are added, robustness to sample changes, and higher accuracy in predicting sales prices.

[1] We thank Paul Johnstone and Reuben Billings from REINZ for data assistance throughout this project. We also thank Arthur Grimes, Frances Krsinich, Michelle Lewis, Chris McDonald, Christie Smith, Martin Wong, and Fang Yao for comments on the paper. Seminar participants at the Household Statistics User Group and Housing Forum also provided useful feedback.

1. Introduction

Accurate measurement of house prices is of significant interest to policy makers and to the wider public. A house is typically the largest single investment that individuals make in their lifetime, so fluctuations in house prices have a very large influence on household wealth. Moreover, house prices along with incomes drive housing affordability, which is of significant interest to the public and government. The Reserve Bank of New Zealand closely monitors house price indices (HPIs) as they can have a significant bearing on both macroeconomic conditions and the build-up of risks to the financial system.

In order to provide an accurate read on underlying price movements, HPIs need to be carefully designed to account for the *quality-mix* problem of real-estate sales. The quality-mix problem refers to the fact that the composition of houses sold will differ from period to period, making it difficult to discern whether observed price changes reflect genuine movements in underlying house prices or simply changes in the composition of houses sold. For example, prices may increase from one month to the next simply because of an increase in the average size of houses sold. Larger homes tend to sell for higher prices, so it's not clear whether the observed increases in prices represent genuine market movements or simply changes in sales composition. This quality-mix problem is of particular concern in the property market since housing quality varies significantly along multiple dimensions.

In order to mitigate this issue, best-practice house price indices should be designed a way that minimises the influence of sales composition. In New Zealand, two agencies produce nationwide indices: the Real Estate Institute of New Zealand (REINZ) and CoreLogic. The CoreLogic index is based on the well-established sales-price to appraisal ratio (SPAR) methodology. The REINZ index is currently produced using a stratified median methodology, developed in conjunction with the Reserve Bank of New Zealand (McDonald and Smith, 2009). The stratification approach goes some way to overcoming the quality-mix problem, especially compared to simpler median house price indices previously used. The REINZ index also has a timeliness advantage of two to three months over CoreLogic, reflecting the timing of when sales data is received. [2]

REINZ has invested in significant improvements to their data-capture system in recent years, resulting in a much more comprehensive dataset than was available in 2009 when the stratified median index was developed. Given this improved dataset, REINZ approached the Reserve Bank in 2015 seeking technical guidance on possible improvements to their house price index methodology. This *Analytical Note* details the approach undertaken and documents the technical guidance given. In particular, we produced and evaluated three well-established and widely adopted methodologies: sales-price to appraisal ratio (SPAR), hedonic regression, and repeat sales.

The rest of this *Note* is organised as follows. Section 2 discusses the methodologies we use to produce the indices. Section 3 discusses the key features of the unit-record dataset provided

[2] As noted in section 3, REINZ receives sales price data once the sale goes unconditional. This provides a timeliness advantage compared to CoreLogic, where we understand that sales price data is received only once it has been processed by the relevant council.

by REINZ, and section 4 outlines the empirical approach used and presents the resulting indices. Section 5 evaluates the three candidate indices alongside relevant benchmarks using a range of criteria, and section 6 concludes.

2. Review of methodologies for constructing house price indices

There are a wide range of methods that a practitioner can use to construct HPIs. These approaches vary widely in their data requirements and their ability to effectively control for the quality-mix problem inherent in real estate sales. This section summarises the three candidate methodologies that we use to produce alternative house price measure for New Zealand. All of the methodologies are well regarded internationally and are used or endorsed by a number of statistical agencies and academics.[3]

The discussion below describes how each methodology addresses the quality-mix problem, and outlines their main advantages and disadvantages. We also provide a brief outline of the simple median and stratified median indices, which are used as benchmarks in our evaluation of the three candidate methodologies.

Hedonic regression

The hedonic regression methodology assumes that the value of a house can be derived from a bundle of observable characteristics (such as the land size, the number of rooms, and the quality of the structure). For example, each additional bedroom will add some value to the price of the house and each additional square metre of land will add some value to the price of the house. A hedonic regression allows us to determine the *unobserved value* of these observed characteristics. Then a mix-corrected index can be produced by removing the component of the price change between two periods that corresponds to changes in housing characteristics. For example, if the average size of houses sold increases from one month to the next, the hedonic methodology allows us to remove the associated price impact.

The hedonic regression methodology is the most data-intensive of the three methodologies we use in this paper. Indeed, there is theoretically no limit to the amount of data that can be used for a hedonic regression house price model. This means that there is a multitude of modelling choices to construct a hedonic house price index.

In practice, the literature suggests that most applied indices use fairly parsimonious models with a linear specification (Eurostat, 2013). In this paper we use a time dummy variable method (DVM), with a small set of property attributes to construct our hedonic HPI. The DVM is a log-linear hedonic house price model pooled across all time periods with time dummies:

$$\log p_{nt} = \beta_0 + \sum_{t=1}^{T} \delta_t D_t + \sum_{k=1}^{K} \beta_k Z_{(n,t)k} + \varepsilon_{nt} \qquad (1)$$

where: p_{nt} denotes the sale price of house n in time period t; D_t are the time dummies (one for

[3] Eurostat (2013) provides an overview of methodologies used in various jurisdictions.

each period, $D_t = 1$ for time period t and 0 otherwise); $Z_{(n,t)k}$ is the quantity of property characteristic k that house n has in period t (in our case $k = 5$ and the property characteristics are: number of bedrooms, land area, floor area, structure age, and suburb); β_k is the hedonic coefficient on characteristic k;[4] and ε_{nt} is an error term. The hedonic house price index is derived from (1) by exponentiating the time-dummy coefficients (δ_t). Some alternative regression and index construction methods were tested (for example including more characteristics and allowing the characteristic coefficients to vary over time); the resulting indices were very similar to the basic DVM approach.

Of the three candidate HPI methodologies, the hedonic regression approach probably provides the best theoretical control for the quality mix problem since it is able to control directly for changes in property quality along a number of dimensions. However, the hedonic approach has very extensive data requirements. High-quality data on multiple property characteristics is not commonly available. Nevertheless, a number of European jurisdictions have produced hedonic indices (Eurostat, 2013).

The large number of modelling and index construction options in hedonic modelling can be viewed as a disadvantage as it reduces index comparability. Revisions to historical index values, while possible under a hedonic methodology, can be minimised with index construction methods (Eurostat, 2013).

Repeat sales

The repeat sales approach to constructing a house price index involves tracking the prices of houses that sell multiple times (Bailey et al, 1963). Assuming that quality does not change over time (for example that quality lost to depreciation is approximately equal to quality gained through renovations and improvements) any change in house prices between sales must reflect a pure price effect since house quality is fixed by definition. A key requirement for the repeat sales approach is that there must be a sufficient number of houses that appear multiple times in the dataset (i.e. a sufficient number of 'repeat sales') to construct a meaningful index.

The most basic repeat sales house price index, which we use in this paper, is generated by regressing a vector of log price ratios for each adjacent matched sales pair on a series of time dummy variables. To be precise, if house n sells in periods s and t ($s < t$) for prices p_{ns} and p_{nt} then the methodology involves regressing the log ratio of prices (p_{nt}/p_{ns}) onto a series of dummy variables that take the value -1 in period s, 1 in period t, and 0 otherwise. Doing this for all properties with repeated sales gives:

$$\log\left(\frac{p_{nt}}{p_{ns}}\right) = \sum_{t=1}^{T} \delta_t D_t + \varepsilon_{n(s,t)} \qquad (2)$$

where: p_{nt} denotes the sale price of house n in time period t; D_t are the time dummies, and $\varepsilon_{n(s,t)}$ is an error term. Each *adjacent* matched pair of sales enters the regression as an

[4] That is, the marginal value of one more unit of that characteristic – for example the value of one more bedroom.

observation. For instance, if a house sells in periods r, s, and t (with $r < s < t$), the pair of sales r and s will enter the regression and the pair of sales s and t, but not the pair of sales r and t. The antilog of the coefficient on the time dummy at period t (δ_t) from (2) is the repeat sales index value for period t. As new repeat sales enter the dataset, it is possible that historical values of the repeat sales index will be revised.

Given that the methodology discards sales of houses that have not sold at least twice, the approach is most commonly applied in regions that have large populations (and thus sufficient data to produce an index even after discarding some sales). For example, the most well-known repeat sales index, the Case-Shiller Home Price Index (Case & Shiller, 1987; Case & Shiller, 1989) measures the value of U.S. residential real estate sales both nationally and in the 20 largest U.S. cities.[5] In the New Zealand context, Grimes and Young (2010) use four repeat-sales methodologies to produce quarterly house price indices for Waitakere City, which has a population of about 200,000 persons.

Sales-Price to Appraisal Ratio (SPAR)

The SPAR methodology for producing a house price index is based on the idea that the appraised value of a house contains useful property-specific information. This means that the appraised value can be used to control for quality-mix changes. In particular, if a large number of high-quality properties sell in one period and a large number of low-quality properties sell the next period, the appraisal values can be used to strip out the quality changes. Changes in the composition of sales should not distort prices, so long as assessed values are able to accurately differentiate properties of different quality.

Formally, the SPAR is an arithmetic repeat index, which uses the appraisal value as the first measure in each pair, and the sales price as the second measure in each pair (Bourassa et al, 2004).

The advantage of using official appraised values as the first measure in the pair is that appraisals for a geographic location are typically conducted on the same date, meaning that properties will have a common base period for comparison purposes. This simplifies the calculation of the index because it removes the need for an estimation technique (Bourassa et al, 2004). As such, the SPAR index is calculated directly as the mean of the ratio of sale price to appraised value in each geographic location at each point in time.

In New Zealand, councils rely on an assessed value called the Rateable Value (RV) for local government taxation. The database of RVs has long been used by CoreLogic (formerly Quotable Value) to compute a SPAR index for New Zealand. The SPAR methodology is also becoming increasingly popular overseas, and is used in Denmark, the Netherlands, and Sweden (Eurostat, 2013).

Formally, a SPAR index is calculated as the mean of the SPAR ratios for each property that sold in a given period:

[5] All of which have populations greater than 680,000 persons.

$$SPAR_t = \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{p_{nt}}{RV_{ns}} \quad (3)$$

where: N_t is the total number of sales in period t; p_{nt} is the sale price of house n in time t, and RV_{ns} denotes the appraised value of property n at time s (typically the appraised value taken as near to time t as possible is used).

The main advantages of the SPAR methodology are that it is straightforward to compute, makes efficient use of available data, and can be produced at a very granular level in a way that avoids historical revisions. The main disadvantages are that the SPAR methodology is heavily reliant on the quality and consistency of the property valuations process. For example, appraised valuations may not accurately reflect major repairs or depreciation in a property over time.

Raw Median / Stratified Median

We present here an overview of the raw median and stratified median house price methodologies, which are used as a benchmark against which our candidate methodologies are compared.

The most basic approach to constructing a house price index is to take the simple median[6] of the sales prices in a given region in a given time period. This approach provides no solution to the quality-mix problem – the house price index may increase from one period to the next due only to an increase in the quality of houses sold.

A more technical approach is to stratify the houses before taking a median. The stratification approach to constructing a house price index involves comparing groups (or strata) of houses that are sufficiently similar to each other. The ability to effectively control for the quality-mix problem inherent in house sales through stratification depends on how well similar groups of houses can be identified. While it is theoretically desirable to use multiple property attributes (such as price, geography, or specific property attributes) together to identify more comparable groups of houses there is an important trade-off to consider. Using many attributes to identify strata increases the number of comparison groups and each group will contain fewer houses, which could ultimately lead to increased index volatility.

It is therefore common to use a simple stratification approach, for example using measures of sales price to sort small geographical areas (suburbs) into groups (see McDonald and Smith (2009) among others). The rationale is that suburbs with similar price behaviour will tend to have similar characteristics and comparisons will better reflect underlying prices rather than differences in quality.

The stratified median index we use for benchmarking purposes is similar in concept to the current REINZ Stratified Index and is constructed as follows. First, suburbs are sorted into deciles based on median sales price over the relevant period of time. These deciles are the groupings (or strata) that are used as the basis for comparison. Then, for each stratum the

[6] Alternative measures of central tendency such as mean could be used, but the median is robust to the observed skewness of sale prices.

median price is computed from all sales in the strata, which is interpreted as a volume-weighted median. The stratified median index is then calculated as the average of the 10 weighted medians (one for each stratum). This method can be used to generate national and regional indices.

The main advantages of the stratified median approach are that it is simple to compute and does not require a lot of detailed information on property sales if a simple stratification approach is used. One disadvantage is that stratification does not deal directly with changes in property quality (i.e. depreciation and renovations) unless indicators of quality are used in the stratification approach. Another drawback is that there is a trade-off between reducing the composition bias by using a more granular stratification scheme and increasing index volatility. Finally, for simple stratification approaches like the one used here for benchmarking purposes, maintenance is required to ensure that the comparison groups remain relevant over time.

Summary of advantages and disadvantages

Table 1 provides a high-level summary of the main advantages and disadvantages for each of the three candidate methodologies plus the stratified median.

Table 1: Advantages and disadvantages of house price index methodologies

	Candidate indices			Benchmark
	SPAR	Hedonic	Repeat Sales	(Stratified) median
Advantages	Light on data needs. Easy to interpret and construct. Can produce at very granular level.	Excellent theoretic sales composition control.	Light on data needs. Easy to interpret and construct.	Light on data needs. Easy to interpret and construct.
Disadvantages	Depends heavily on quality/consistency of council valuation. Requires method to handle valuation updates. Appraised value may not accurately reflect major repairs or depreciation.	Heavy data needs. Challenge to communicate. Revisions, but these can be minimised.	Difficult to apply at more granular levels. Revisions as new sales-repeats occur. Possible selection problem for frequent sales. No direct control for quality changes (i.e. major repairs, depreciation).	No direct control for property quality changes. Trade-off between addressing quality mix issue and increasing volatility. On-going maintenance needed.

Our view is that there is no clear cut 'best' HPI approach based on solely methodological grounds (other than the fact that our three candidate indices represent a theoretical improvement over the median / stratified median indices). Ultimately the relative performance of the approaches in the New Zealand data context is an empirical question. For example, the performance of a SPAR index critically depends on the quality of appraised values, a hedonic index depends on the quantity and quality of data on property characteristics, a repeat sales index is highly dependent on the availability of repeat sales without major renovations and a stratified median index depends on identifying meaningful strata (i.e. groupings) of sufficiently similar houses that are stable over time.

3. Data

The unit-record data for this project were provided by REINZ, a membership organisation for registered real estate agents, representing 14,000 agents (REINZ, 2016). REINZ receives data on sales from their members once the sale becomes unconditional. The data used in this paper are records for every sale processed by REINZ's member-agents between January 1992 and December 2015. During this period, over 2 million sales were recorded and a large portion of these are used in the analysis.[7]

The dataset contains a rich range of property attributes. This is due to significant investments by REINZ to streamline the collection and processing of sales data received from members, and to link these sales records to other property databases. Importantly, the dataset contains the necessary attributes to construct all three candidate HPI methodologies and the benchmark stratified median index.

Unique property identification

A key attribute in the dataset is a unique identifier for each property in New Zealand. This allows us to construct a repeat sales index by identifying all sales in the database that correspond to a given property. In the REINZ database, around half of the dwellings appeared only once, and were thus not usable for a repeat sales index. Around 250,000 properties appear twice, 130,000 appear three times, 60,000 appear 4 times, and 30,000 appear 5 or more times (figure 2).

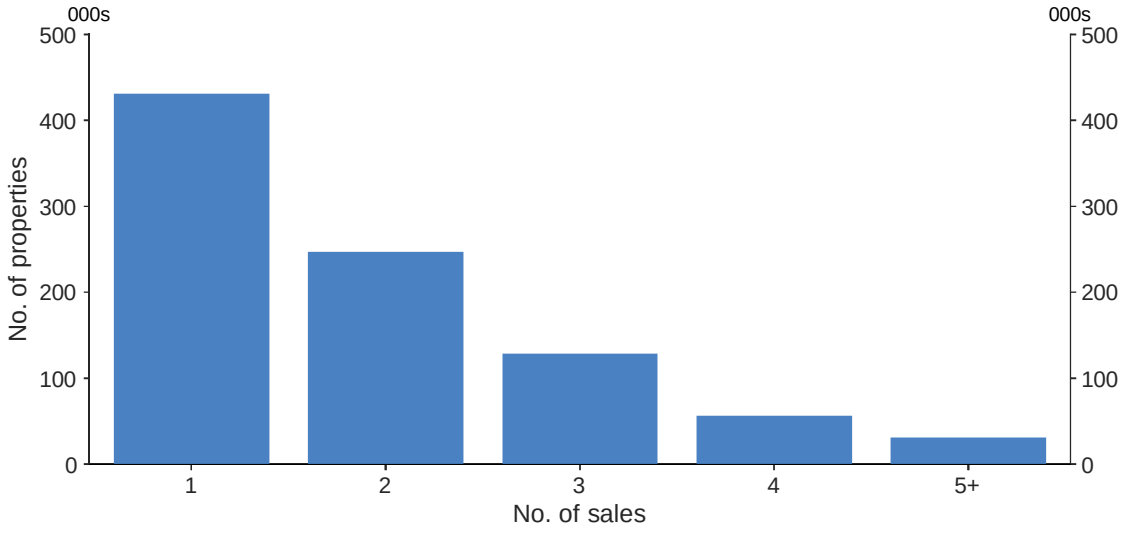
Appraised valuation

As noted above, councils rely on an assessed value called the Rateable Value (RV) for local government taxation. As they incorporate the judgement of expert valuers, RVs likely capture a lot of property-specific information that may not be observed in the data (such as views, proximity to amenities, etc.). Councils have an incentive to produce accurate appraisals since RVs form the basis for taxation, and property owners will likely contest excessive valuations. The RV of a property should be updated to reflect building work requiring a building consent. However, minor improvements or depreciation in a property over time are a possible source of error. In general, there is strong positive correlation ($r = 0.62$) between sale price and

[7] For comparison, at the 2013 census there were 1,561,956 private dwellings in New Zealand.

valuation, which suggests that the valuations tend to be reasonably accurate.[8]

Figure 2: Number of property sale repeats



The version of the REINZ dataset we use contains only the latest set of valuations, rather than the full history of valuations for the sample period (1992-2015).[9] This means the SPAR approach will provide less-accurate adjustment for quality earlier in the sample period. RVs are reviewed on a three-yearly cycle and the timing of these reviews differs in each Territorial Local Authority (TLA). For this reason, the SPAR index must be constructed at the TLA level and then aggregated (see below for a description of the weighting approach).

Updates to RVs present somewhat of a challenge to the SPAR index – changing the RV changes the base value of the SPAR index (and the base period), which can introduce revisions to historical data. There are ways of ameliorating this issue, such as chaining the updated-RV indices and the old-RV indices together to maintain a constant historical series.

Property characteristics

The database contains a range of property characteristics that are input by REINZ's member-agents. These characteristics are cross-checked and, in cases where they are missing, replaced by data from external sources such as realestate.co.nz, Land Information New Zealand (LINZ), and property records from councils. Our hedonic regression model focuses on a small set of property attributes where the quality is acceptable and records are reasonably complete: the number of bedrooms, floor area of the dwelling, property land area, and age of the structure. These property attributes (occasionally with the addition of the number of bathrooms) are generally considered to be necessary and sufficient to construct a well-specified hedonic regression model.[10] We also include a set of suburb-level locational dummy variables to control for neighbourhood effects on property values.

[8] The correlation is this high despite the fact that we are correlating current (end-of-period) valuations with real-time sales prices. If we had real-time valuations, this correlation would likely be even higher.
[9] Operating a SPAR index in real-time will require some procedures to handle periodic RV updates.
[10] Of course, some authors suggest more complex models.

4. Constructing the candidate indices

To compare the performance of the indices, we construct each index at a nationwide level and for 12 regional areas. There are 76 TLA classifications, which we aggregate into 12 regional areas. The 12 regional areas are designed to be geographically meaningful while also containing enough sales to produce reliable and stable indices. The regional classification we use also matches the regional classification at which many tier-one macroeconomics statistics (such as labour market statistics) are produced.[11] The hedonic and repeat sales indices were calculated at the 12 region level and then weighted together to generate national level indices. As noted above, the SPAR index instead needs to be constructed for each TLA, and then aggregated up to the 12 regions and to the national level.

There are a number of possible ways to weight regional indices to the national level and the choice will depend on the purpose of the index. For the analysis that follows, we have weighted each region according to its estimated share of the total value of the housing stock in New Zealand. This approach means that the HPIs measure the *total value* of the housing stock, which aligns best with the Reserve Bank's interest in monitoring macroeconomic conditions and financial stability. Other weighting options include using a measure of the *volume* of the housing stock (i.e. number of houses in existence) or the number of sales. In this case, the resulting HPIs would measure *the value of a typical house*.

Table 3 shows how alternative regional weighting schemes compare in New Zealand. The value weights, which are used in the remainder of this paper, place more weight on regions with higher house prices, and less weight on regions with lower house prices. They are derived by summing up all of the latest RV's in each regional area from the rating roll, and then estimating the total valuation at the required base date using the relevant regional price index. Stock volume weights are derived from Census data on the number of private residential dwellings, and sales volume weights are derived from REINZ sales data.

In constructing the indices, we remove outliers from the data based on a number of rules, on the basis that they are likely to be either coding errors or non-standard properties that are not reflective of general house prices. The removal process varied across the different approaches. For example, for the purpose of computing a hedonic house price index, we remove all records with implausible values for characteristics.[12] For the SPAR index we remove all observations that were significantly above or below the median SPAR for the TLA over the past three months. For repeat sales, we remove repeat sales pairs that occurred within 3 months on the basis that these sales likely demonstrate some selection bias.

Figure 3 shows the three candidate indices at a nationwide level. All three candidate house price methodologies produce credible-looking indices, which correlate well with existing New Zealand house price indices, and fit with well-established stylised facts of the cyclical properties of house prices.

[11] New Zealand has 16 formal regional council areas – we aggregate some of the smaller ones to produce our 12 regions.
[12] Generally we remove the largest 1% of observations for each property attribute, for example removing any observation with more than 10 bedrooms.

Table 2: Regional weights (as at 2013)

	Weighting approach			
	Sales volume (REINZ sales data)	Stock volume (2013 Census data)	Sales value (REINZ sales data)	Stock value* (RVs)
Northland	2.7	3.8	1.8	1.9
Auckland	32.5	30.2	50.8	46.5
Waikato	9.2	9.7	6.9	6.9
Bay of Plenty	6.5	6.6	5.4	5.2
Gisborne/Hawkes Bay	4.1	4.7	2.4	2.6
Taranaki	2.6	2.8	1.6	1.8
Manawatu/Wanganui	5.2	5.6	2.4	2.8
Wellington	11.2	11.3	10.1	11.2
Nel./Tas./Marl./W.C.	4.1	4.4	2.8	3.2
Canterbury	14.4	13.2	11.2	12.8
Otago	4.9	5.1	3.5	3.9
Southland	2.7	2.4	1.0	1.1

*This is the weighting approach used in the analysis.

Figure 3: Nationwide candidate indices (annual percent change, s.a.)

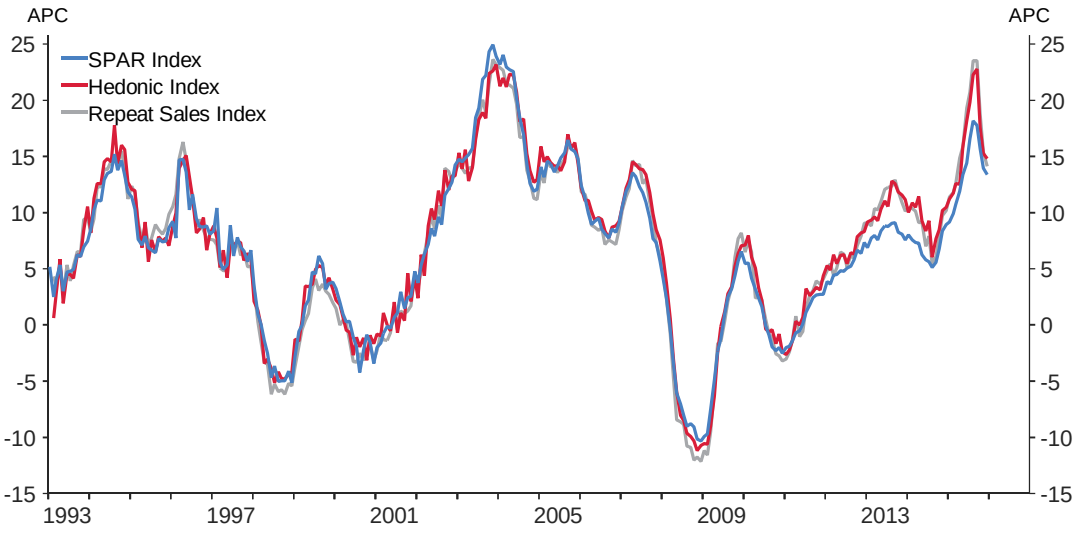
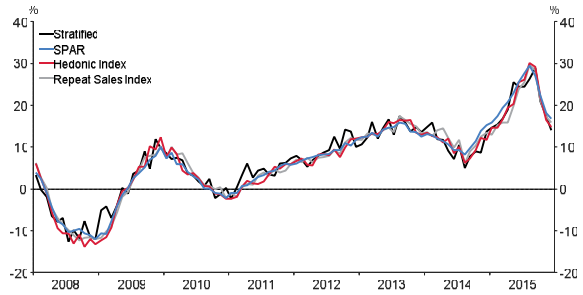


Figure 4 provides a more detailed overview of the index results since 2008 for Auckland, Wellington, Canterbury and Waikato. Consistent with figure 3, the regional indices show similar cyclical patterns and correlate well with existing regional indices in their low frequency

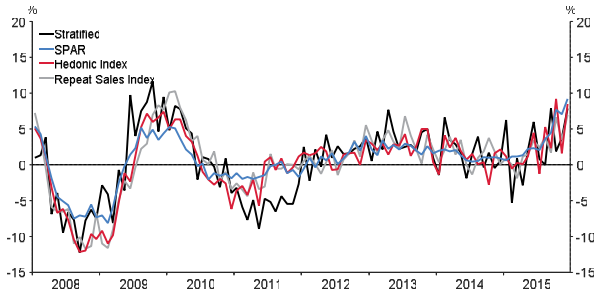
(business cycle) movements.[13] However, there are some notable differences in the month-to-month volatility. First, the three alternative approaches appear less volatile than the current REINZ stratified approach, particularly for the Canterbury index. Second, the SPAR index is noticeably less volatile than the other alternative approaches, especially outside of Auckland. While visual inspection provides a good starting point to compare index behaviour, section 5 presents an empirical evaluation that provides a more formal basis on which to evaluate the relative performance of alternative HPIs.

Figure 4: Candidate indices for selected regions (annual percent change)

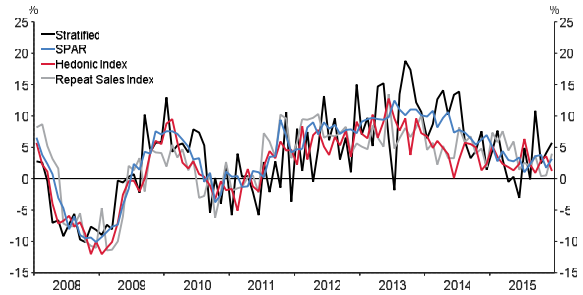
Auckland



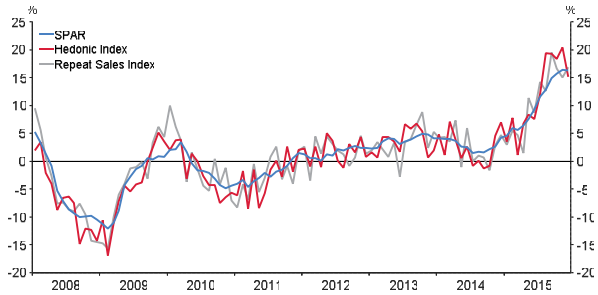
Wellington



Canterbury



Waikato



5. Index evaluation

In this section we present some metrics which, in addition to visual inspection, provide a basis for evaluating the indices against each other. There is no standard or widely accepted process with which to evaluate house price indices since there is no 'true' house price measure against which to compare candidate indices. Our approach is to use a variety of metrics to assess the relative performance of our suite of indices:

- **Volatility statistics** – an ideal house price index should not be unnecessarily noisy. An analyst should be able to detect the underlying signal of an index without significant noise. This metric becomes more important at higher levels of disaggregation (e.g. at regional or TLA level), where volatility naturally increases as sample size decreases. We use two volatility measures: the number of turning points in the index and the

[13] The indices broadly match the cyclical patterns of the existing stratified indices for Auckland, Wellington and Christchurch. They also exhibit similar patterns to the CoreLogic regional indices.

average deviation from a statistical trend.

- **Real-time revisions** – an ideal house price should not be revised much as additional data are added. Where applicable, we estimate the indices in real-time to see how much the index values were revised. This metric can only be computed for the hedonic and repeat sales indices.
- **Robustness measures –** we use a sampling technique to estimate how much the indices are impacted by the exclusion of some properties. An ideal house price index should not change much if you exclude a small proportion of the observations.
- **Accuracy in predicting sales prices** – fundamentally a house price model should be able to accurately detect underlying changes in house prices. One way to test this is to compare predicted house prices from each model to actual house prices. The methodology that most accurately predicts actual house prices can be seen as the best methodology.

Where applicable, we also present an evaluation of two benchmark indices – a raw median and a stratified median.[14] Table 5 provides a summary of the evaluation results and shows which tests were applied to each of the five indices subjected to our evaluation exercise.

a. Volatility measures

The first metric for evaluating the indices involves determining the smoothness of the series. Although it is not necessarily the case that house price growth is a very smooth process, in order to provide an accurate reading of turning points in the housing market and improve the signal-to-noise ratio the indices should not be unnecessarily volatile. In particular, all of the candidate indices have similar turning points, and thus similar signals. Our volatility statistics show which index has the lowest amount of noise associated with that signal.

We present two candidate metrics of volatility which are standard for time-series evaluation: the number of turning points in the series and the deviation of the series from its short-term trend (as in McDonald and Smith, 2009). Volatility results for the nationwide indices, including our calculated stratified median and raw median, are shown in table 2.[15] In each case, the volatility statistics are calculated for the annual percent change (APC) of the seasonally-adjusted series.

Under the metric of the number of turning points, all three of the candidate indices out-perform the benchmark stratified index on the criterion of minimising volatility. Over 12-month cycles, the three candidate indices saw 12 fewer turning points than the benchmark stratified median, and 29 fewer than the raw median. Under the metric of root mean square deviation from trend, the three candidate indices out-perform the stratified median and raw median for the 3-month trend and the SPAR out-performs both benchmark indices for the 6-month trend as well.

[14] These benchmark indices are constructed as outlined in Section 2.
[15] These statistics for each of the 12 regions are shown in Appendix 1.

Table 3: Volatility statistics for the house price methodologies (nationwide)

	Period	SPAR	Hedonic	Repeat Sales	Stratified median	Raw median
Turning points (fewer is better)	1	100	145	126	150	171
	3	35	57	42	75	105
	12	18	18	18	30	47
Trend deviation	3	1.1	1.2	1.2	1.4	1.4
(lower is better)	6	2.2	2.4	2.5	2.5	2.6

Note: Period refers to the time over which the statistic is calculated. For example, turning points over one period counts the number of times that the APC is increasing (decreasing) in period *t* and decreasing (increasing) in period *t+1* while turning points over 12 periods counts the number of times that the APC is increasing (decreasing) in period *t* and decreasing (increasing) in period *t+12*. The deviation from trend is calculated as the root mean square error relative to a 3-month and 6-month moving average trend.

Together, these results suggest that the stratified median represents an improvement over the raw median, and that all three candidate indices represent an improvement over the stratified median. Therefore, movements in house prices generated by the new methods are likely to be more representative of actual change (i.e. unlikely to be simply reversed the next period) than the stratified or simple median indices.

Among the three candidate indices, the SPAR index appears to perform marginally better than the other two indices at a nationwide level. Consistent with figure 4, this dominance of the SPAR becomes more pronounced at higher levels of disaggregation (Appendix 1).

b. Robustness to real-time estimation

Another measure of index success is how robust the historical values are to adding new data points. This test is only relevant for the two methodologies which involve model estimation – the hedonic index and the repeat sales index. In these cases, the estimated coefficients may change as new data are added, which influences historical values of the index.[16] On the other hand, the SPAR does not involve regression estimation and so does not change as additional data are added. The benchmark stratified and raw median indices are also unchanged as more data are added.

We examine the real-time performance of the hedonic and repeat sales indices by calculating each index in pseudo-real time and comparing the real-time values to the *ex-post* values (as in Silverstein, 2014). The estimation is done over an expanding window, starting at 108 months (1992M1 to 2000M12) and adding one additional year of data (12 data points) at a time until we look at the full sample (288 months between 1992M1 and 2015M12). We find that the real-time and ex-post estimates are almost identical, which suggests that the hedonic and repeat sales index are robust to estimation period.

[16] The hedonic index can be constructed in such a way to avoid historical revisions when new time periods of data become available but we believe the exercise is still a useful indicator of index behaviour.

c. *Robustness to changes in sample*

Another robustness test commonly used to evaluate the stability of an unobservable index is the 'bootstrapping' procedure (see, for example, Gospodinov and Ng, 2013). This technique involves taking a random sample of the observations, estimating the index, and repeating hundreds of times to produce a large number of estimated indices. The dispersion of these bootstrapped indices gives an indication of the stability of the index – a stable index should be similar irrespective of which sample was used, while an unstable index will tend to be quite different given different samples of the data.

To evaluate the three candidate indices on this dimension, we construct a common dataset which contains all those observations which have a unique property ID, a valuation, and all hedonic characteristics, and appear at least twice in the dataset (so that the same dataset can be used to calculate all three indices). In order to simplify computation, we focus on three TLAs (Auckland, Wellington, and Christchurch) in turn.

We then take a sample of 80 percent of these observations, estimate all three models plus the stratified median and raw median using that sample, and retain the index values. We repeat this process to get 300 bootstrap replications. We draw the 2.5th and 97.5th percentile of the 300 replications of each index to form an approximation of a two-standard deviation band around the median,[17] in order to quantify the dispersion. The index with the lower variation around the median is the better index under this metric. To formalise this, we calculate the mean percentage point difference between the +2 standard deviation band and the -2 standard deviation band over the sample. Here a lower value indicates more stable (and hence better) index.

The results are summarised in table 4, and Appendix 2 shows the charts of the bootstrapped indices for Auckland.

Table 4: Mean absolute percentage point difference (lower is better)

	SPAR	Hedonic	Repeat Sales	Stratified median	Raw median
Auckland	1.10	1.36	1.24	3.13	2.81
Christchurch	1.53	2.06	1.66	4.74	3.99
Wellington	1.76	2.68	2.03	5.78	5.01

In terms of this robustness exercise, the three candidate indices outperform the median and stratified median benchmarks. This suggests that, relative to the simpler benchmarks, the candidate methodologies more accurately account for changes in the sample of houses sold between periods. Thus, the three candidate methodologies produce indices that are more likely to accurately capture underlying trends in the housing market.

[17] The bootstrapped indices are approximately normally distributed, so this calculation is valid.

Under this metric, the SPAR performs best, with lower volatility as the sample of sales changes than the hedonic and repeat sales approaches. This suggests that an index based on the SPAR methodology should provide an accurate assessment of reality, even if the sample of houses sold varies considerably from one month to the next.

d. Out-of-sample prediction exercise

The final approach we take to evaluating the indices is an out-of-sample prediction exercise in the spirit of Nagaraja, Brown, and Wachter (2010). This method involves choosing a random sample (80 percent) of observations as a *training set*, estimating the model on this set, and then predicting the sales prices of the remaining observations (20 percent) which make up the *test set* using the estimated model.

For example, for the SPAR methodology we estimate a SPAR index based on the training set, then rate the CVs in the test set up or down using the SPAR HPI to predict the sales price. For the repeat sales index we estimate the time period dummies, and then, based on the first sales price of properties in the test set and the repeat-sale HPI, estimate subsequent sales prices. For the hedonic index, we estimate the hedonic regression coefficients (including on each time dummy) using the training set, and then predict the sales price using characteristic data and sales period of the test set. We are not able to perform a comparable out-of-sample exercise on the median or stratified median indices since their unit of construction is groups of houses rather than individual sales records as is the case for the three candidate indices.

Given that the actual sales prices are known for the test set, we can make a comparison between the predicted and actual values. The metric for success in this test is the correlation between the predicted and actual sales values for the test set, with a higher correlation indicating better model performance. The results of this test at a nationwide level are shown in table 5, with charts in Appendix 3. Statistics at a regional level are shown in Appendix 4.

Table 5: Correlation between actual and predicted house prices

	SPAR	Hedonic	Repeat Sales
Correlation	0.942	0.831	0.916

As with other evaluation measures, the SPAR index performs best in this measure, with a 94.2 percent correlation between the predicted and actual sales price at a nationwide level. The repeat sales index performs nearly as well, with the hedonic index performing less well.

Summary of evaluation results

Table 6 provides a summary of the index evaluation results presented in this section. The rankings in Table 6 are relative and subjective since we have not performed any statistical significance tests.

Table 6 supports the finding that while all three candidate index methods produce credible

results and represent an improvement over the median and stratified median benchmarks, the SPAR index performs best on a number of measures. In particular, the SPAR index has the least volatility, is not subject to revisions through time, is robust to changes in sample, and provides very good predictions of underlying house prices.

Table 5: Summary of index evaluation results (relative rating*)

	SPAR	Hedonic	Repeat Sales	Stratified median	Raw median
Volatility statistics	I	III	II	IV	V
Real-time revisions	n/a	I	I	n/a	n/a
Robustness to changes in sample	I	II	II	III	III
Out-of-sample predictions	I	III	II	n/a	n/a

* Rankings are relative (a rank of 'I' is best) and somewhat subjective since we have not explicitly tested for the statistical significance. Where 'n/a' is noted, it was not possible to apply the particular evaluation test.

6. Conclusions

This paper has presented three alternative methodological approaches to producing house price indices for New Zealand, using data provided by REINZ. The three methodologies are all theoretically sound, and are supported by various academics and practitioners. We find that each methodology produces credible results and that all three methodologies are able to outperform the benchmark median and stratified median indices.

Evaluation of the three candidate indices suggests that the SPAR index performs best empirically along a range of dimensions. It has the lowest month-to-month volatility of the three indices (especially at low levels of disaggregation), it is not subject to revisions through time, and it provides very good predictions of underlying house prices. As such, our conclusion is that the SPAR methodology produces the best indices from an empirical perspective.

However, the ultimate usability of a house price index involves more than just empirical rigour. Practitioners must consider a range of subjective features such as ease of explanation to a wide audience, ease of calculation and maintenance, and ability to provide information at the desired regional level. Providing answers to these questions is beyond the scope of this paper.

References

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*, 58(304): 933-942.

Bourassa, S. C., M. Hoesli, and J. Sun (2004). A Simple Alternative House Price Index Method. *Journal of Housing Economics*, 15(1): 80-97

Case, K. E. and R. J. Shiller (1987). Prices of single-family homes since 1970: new indexes for four cities. *New England Economic Rev*iew (September/October): 45-56.

Case, K. E. and R. J. Shiller (1989). The efficiency of the market for single-family homes. *American Economic Review,* 79(1): 125-137.

Eurostat (2013). *Handbook on Residential Property Prices Indices (RPPIs)*, Luxembourg: Publications Office of the European Union.

Gospodinov, N. and S. Ng (2013). Commodity Prices, Convenience Yields, and Inflation. *The Review of Economics and Statistics*, 95(1): 206–219.

Grimes, A. and C. Young (2010). A Simple Repeat Sales House Price Index: Comparative Properties Under Alternative Data Generation Processes, *Motu Working Paper* 10-10.

McDonald, C. and M. Smith (2009). Developing stratified housing price measures for New Zealand, Reserve Bank of New Zealand *Discussion Paper*, DP2009/07.

Nagaraja, C. H., L. D. Brown, and S. M. Wachter (2010). *House Price Index Methodology*, Weiss Centre Working Papers 2010-06.

Real Estate Institute of New Zealand (REINZ) (2016). *About REINZ*, retrieved from https://www.reinz.co.nz/public-home-page.

Silverstein, J. M. (2014). *House Price Indexes: Methodology and Revisions*, Research Rap Special Report, Federal Reserve Bank of Philadelphia.

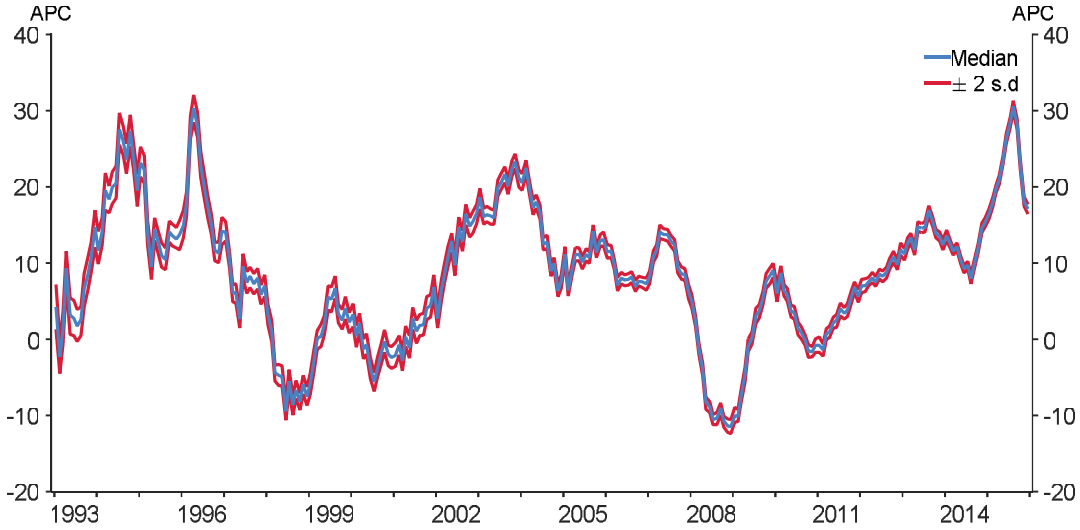
Appendix 1: Regional volatility statistics

	Number of turning points								
Method	SPAR			Hedonic			Repeat Sales		
Lag	1	3	12	1	3	12	1	3	12
Northland	187	111	62	177	121	97	187	136	94
Auckland	122	37	18	139	67	26	131	51	22
Waikato	163	87	38	182	114	68	173	113	66
Bay of Plenty	179	99	48	172	114	80	175	89	70
Gisborne/Hawkes Bay	168	84	56	175	121	82	171	111	76
Taranaki	182	128	80	163	121	78	186	122	75
Manawatu/Wanganui	155	99	57	163	111	76	161	102	73
Wellington	144	87	36	161	95	62	173	103	59
Nel./Tas./Marl./W.C.	169	105	76	177	128	116	188	112	72
Canterbury	154	76	34	188	120	68	158	109	52
Otago	169	95	57	184	122	85	172	111	62
Southland	160	106	60	176	96	69	172	102	90

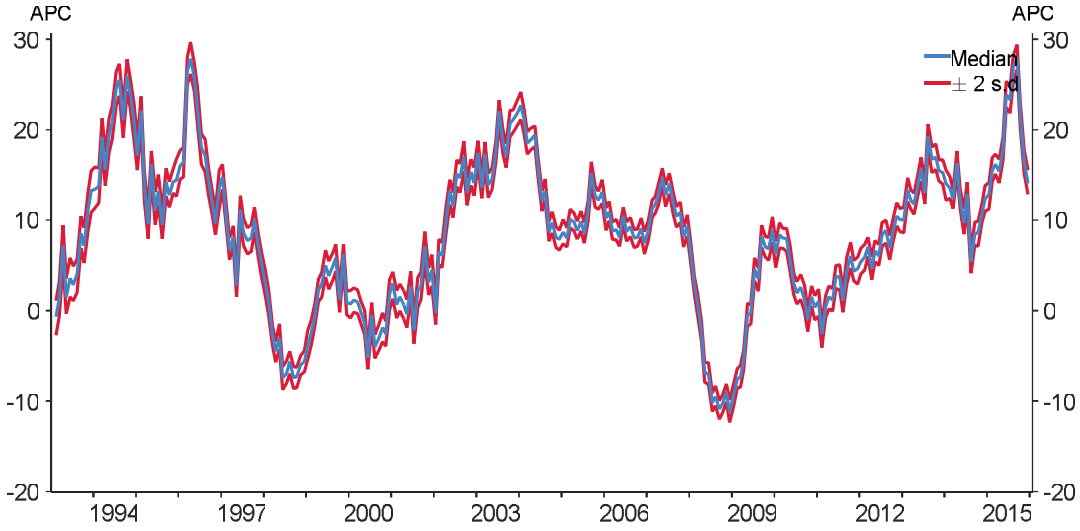
	Mean square deviation from trend					
Method	SPAR		Hedonic		Repeat Sales	
Period	3	6	3	6	3	6
Northland	2.7	3.6	4.1	4.8	4.3	5.4
Auckland	1.6	3.1	1.8	3.2	1.7	3.3
Waikato	1.4	2.4	2.3	3.1	2.1	2.9
Bay of Plenty	1.6	2.6	2.5	3.1	2.0	2.8
Gisborne/Hawkes Bay	1.9	2.8	2.9	3.7	2.3	3.0
Taranaki	2.3	3.1	3.3	4.2	3.4	4.2
Manawatu/Wanganui	1.5	2.3	2.5	3.3	2.5	3.2
Wellington	1.2	2.0	1.7	2.5	1.6	2.5
Nel./Tas./Marl./W.C.	2.3	3.4	3.5	4.1	2.5	3.4
Canterbury	1.4	2.3	2.9	3.9	1.6	2.6
Otago	2.4	3.3	3.0	4.1	2.8	3.8
Southland	3.4	4.3	3.6	4.8	4.0	5.2

Appendix 2: Bootstrapping charts for Auckland

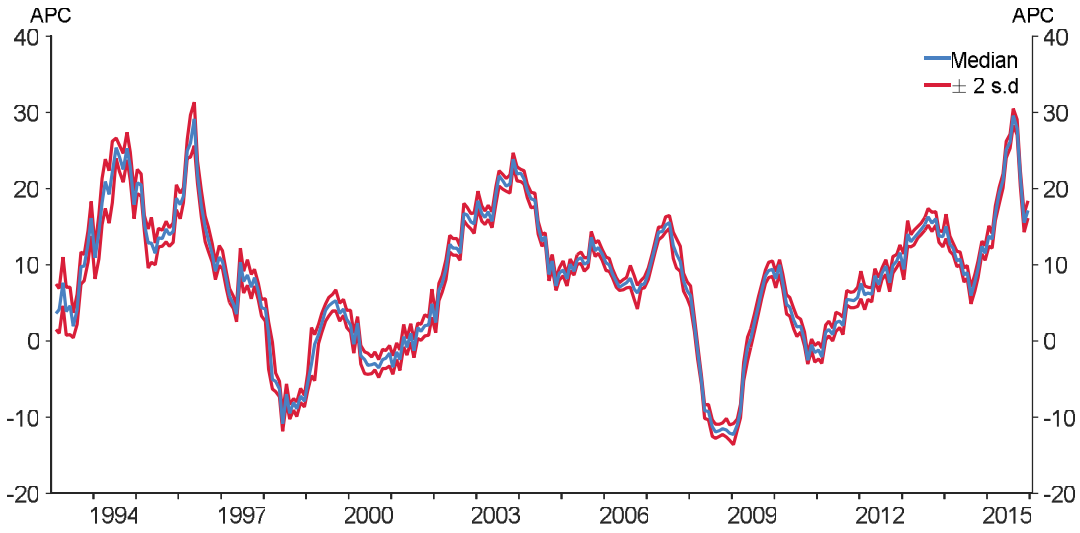
A2.1 – SPAR methodology



A2.2 – Hedonic regression methodology

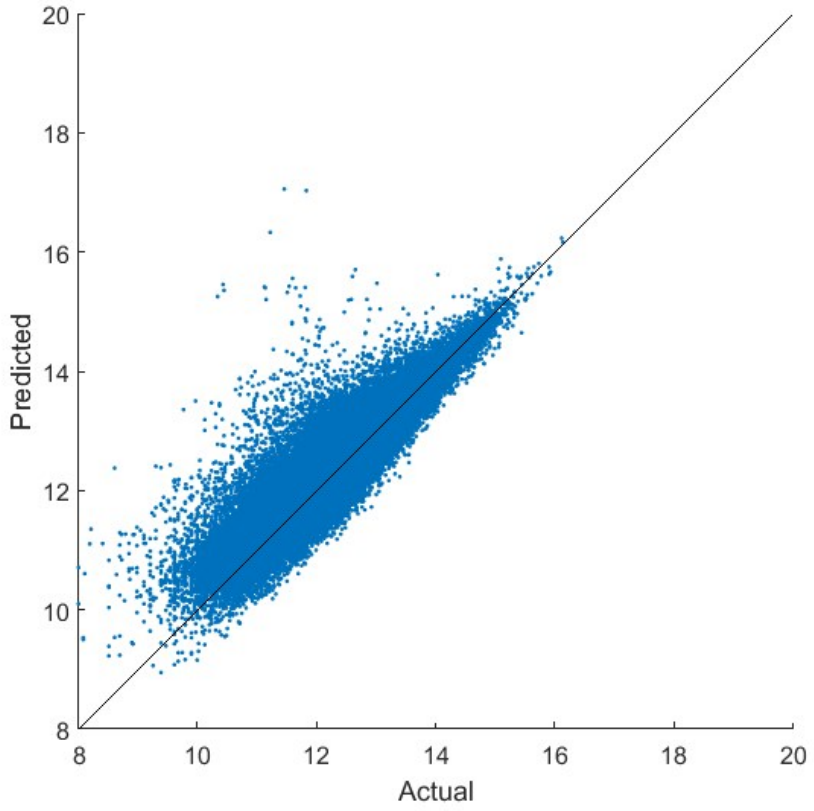


A2.3 – Repeat sales methodology



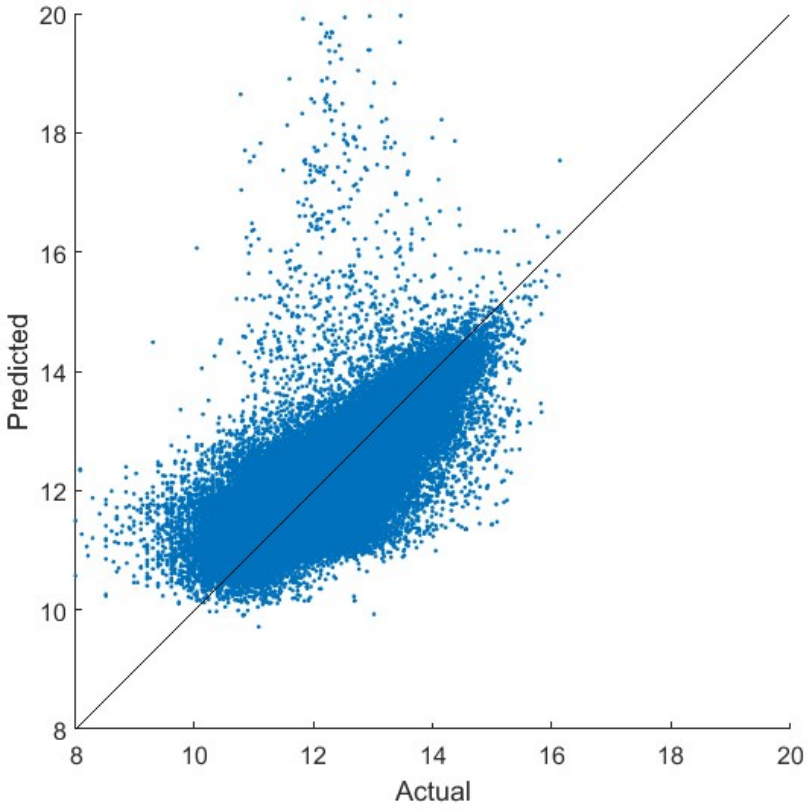
Appendix 3: Train and test charts (nationwide)[18]

A3.1 – SPAR methodology

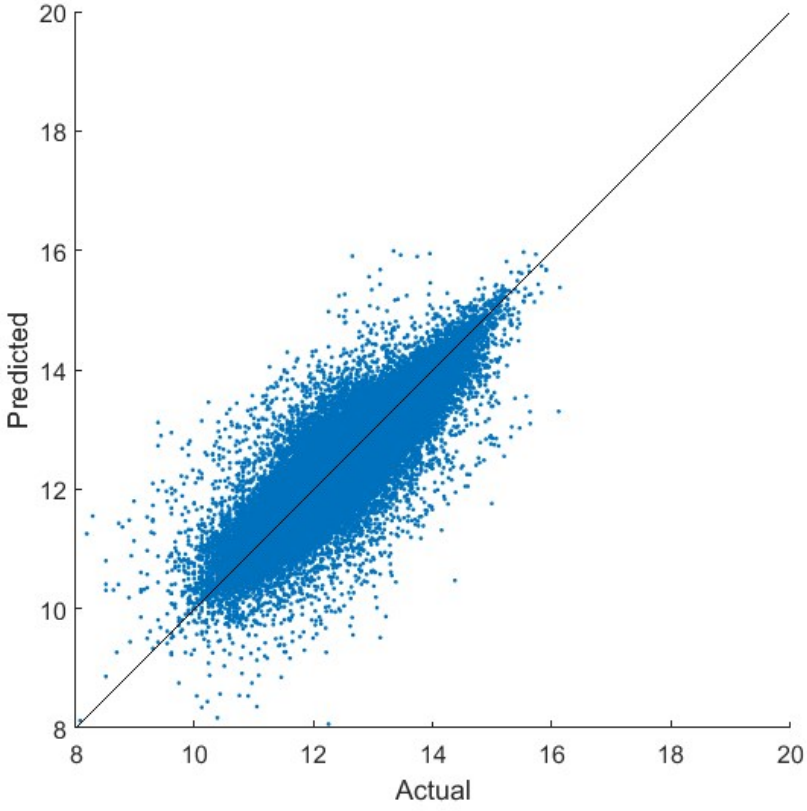


[18] These charts show the predicted (log) house prices for each methodology against the actual (log) house price. The black line shows the 45 degree line of equality – if the methodology perfectly predicted house prices then all the points would lie along this line.

A3.2 – Hedonic regression methodology



A3.3 – Repeat sales methodology



Appendix 4: Train and test correlations at a regional level

	Correlation		
	SPAR	Hedonic	Repeat Sales
Northland	0.9368	0.8679	0.9079
Auckland	0.947	0.8339	0.9231
Waikato	0.9287	0.8436	0.9282
Bay of Plenty	0.8405	0.6347	0.7666
Gisborne/Hawkes Bay	0.9411	0.8338	0.9217
Taranaki	0.9420	0.8700	0.9224
Manawatu/Wanganui	0.9237	0.8096	0.8915
Wellington	0.9331	0.8038	0.9022
Nel./Tas./Marl./W.C.	0.8942	0.7602	0.8878
Canterbury	0.9097	0.7666	0.8881
Otago	0.9118	0.7767	0.8886
Southland	0.9233	0.8243	0.9123