Journal of Property Research

ISSN: 0959-9916 (Print) 1466-4453 (Online) Journal homepage: http://www.tandfonline.com/loi/rjpr20

A robust house price index using sparse and frugal data

Phil Maguire, Robert Miller, Philippe Moser & Rebecca Maguire

To cite this article: Phil Maguire, Robert Miller, Philippe Moser & Rebecca Maguire (2016) A robust house price index using sparse and frugal data, Journal of Property Research, 33:4, 293-308, DOI: 10.1080/09599916.2016.1258718

To link to this article: https://doi.org/10.1080/09599916.2016.1258718

Published online: 09 Dec 2016.

JOURNAL OF PROPERTY RESEARCH, 2016
VOL. 33, NO. 4, 293–308
http://dx.doi.org/10.1080/09599916.2016.1258718

Routledge
Taylor & Francis Group

A robust house price index using sparse and frugal data

Phil Maguire[a], Robert Miller[a], Philippe Moser[a] and Rebecca Maguire[b]

[a]Department of Computer Science, National University of Ireland, Maynooth, Ireland; [b]School of Business, National College of Ireland, Dublin, Ireland

ABSTRACT

In this article, we describe a house price index algorithm which requires only sparse and frugal data, namely house location, date of sale and sale price, as input data. We aim to show that our algorithm is as effective for predicting price changes as more complex models which require detailed or extensive data. Although various methods are employed for determining house price indexes, such as hedonic regression, mix-adjusted median or repeat sales, there is no consensus on how to determine the robustness of an index, and hence no agreement on which method is the best to use. We formalise an objective criterion for what a house price index should achieve, namely consistency between time periods. Using this criterion, we investigate whether it is possible to achieve strong robustness using frugal data covering only 66 months of transactions on the Irish property market. We develop a simple multi-stage algorithm and show that it is more robust than the complex hedonic regression model currently employed by the Irish Central Statistics Office.

ARTICLE HISTORY
Received 28 July 2016
Accepted 6 November 2016

KEYWORDS
House price index; sparse data mining; frugal heuristics; index robustness; central price tendency model

1. Introduction

House price indexes play a critical role in top-level decision-making, and have impacts on investment decisions by both the private and public sectors (Plakandaras, Gupta, Gogas, & Papadimitriou, 2015). House owners, bankers and policy-makers all pay close attention to relative price levels and the magnitude and direction of price changes in both regional and local markets (Costello & Watkins, 2002; Leishman, 2009; Munro, 1987). This information can be useful in forecasting inflation, economic output and real GDP growth (Case, Quigley, & Shiller, 2005; Forni, Hallin, Lippi, & Reichlin, 2003; Gupta & Hartley, 2013; Gupta & Kabundi, 2010; Stock & Watson, 2004)

House price indexes are also important for academic research aimed at understanding the dynamics of the market, and for investigating issues of societal relevance, such as housing affordability and price bubbles (Bourassa, Hoesli, & Sun, 2006). The study of index robustness is particularly relevant in the contemporary financial environment, given the recent price volatility in international housing markets and the prominence of housing market debt instruments as a primary cause of the global financial crisis (Goh, Costello, & Schwann, 2012). The possibility of hedging against housing risk (e.g. Englund, Hwang, & Quigley, 2002; Shiller, 2003) depends on access to extremely accurate price indexes.

CONTACT Phil Maguire pmaguire@cs.nuim.ie

Most house prices indexes require either extensive data which stretch back decades or else detailed data, which describe numerous features of each home. Our aim in this article is to develop an algorithm which requires only a few months of transactions (sparse data) and the barest of details (frugal data). We hope to show that such algorithms can match the robustness of more complex data-intensive methods. If feasible, such techniques would have numerous advantages over the systems currently in use. For a start, they would be less labour-intensive, relying on information scraped automatically from webpages, with no need for the input of expert statisticians. Second, they would be more responsive, giving consistent up-to-date information about house price changes. At the moment, statistics offices, such as the Irish Central Statistics Office (CSO), release information only once a month, with nearly a month of delay. An automatic algorithm could recompute the changes every few minutes, using not only sale prices, but also asking prices gleaned from online property websites, thus capturing immediate shifts in market sentiment.

2. House price index approaches

We begin by providing an overview of existing strategies for determining house price indexes. Given the importance of house price movements and the voluminous associated literature (see Hansen, 2009, for an overview), it is perhaps surprising that no consensus exists on how an index should be constructed. When comparing house price index models, researchers are faced with numerous data and methodological issues which stand in the way of constructing an accurate index (Goh et al., 2012). First of all, housing markets are highly illiquid. Due to substantial search, transaction and relocation costs, only a fraction of the total housing stock is sold each year. The 2008 financial crisis and subsequent Irish property crash led to a much lower number of transactions than usual for this period (see Lyons, 2015). For example, according to the stamp duty returns maintained by the Irish Property Services Regulatory Authority (PSRA), in the period January 2010–July 2015 less than 150,000 properties were transacted, out of a total of 1.65 million (9.1%), implying that the average house is transacted once every 60 years.

Another problem is that the properties being sold have varying characteristics which are affected by geographical and temporal factors, introducing potential bias into the sample selection. Houses are also subject to quality change over time, which can also vary by area.

To overcome the problem of small sample size, data are often pooled arbitrarily into broad representations of time and geography. The assumption here is that the pooled sample will produce price indexes that are statistically equivalent to those that would have been obtained from the smaller constituent subsamples. This must be done carefully, as excessive pooling of data for house price index construction can lead to biased price index estimates (Englund et al., 2002; Goh et al., 2012). Developing and maintaining an unbiased index according to best international practice is a complex and demanding process (see de Haan & Diewert, 2011). In the following sections, we describe the three main techniques used for deriving house price indices, namely hedonic regression, repeated sales and adjusted-mix median.

2.1. Hedonic regression

The hedonic modelling method is used to construct house price indexes in Ireland and in the UK. The central idea, originally introduced by Kain and Quigley (1970), is that of determining the quality of a given house by decomposing it into its constituent characteristics, then estimating the contributory value of each characteristic (e.g. number of bedrooms, distance to city centre, plot size, etc.). The results of the regression indicate the changes in property values for a unit change in each characteristic, assuming that all the other characteristics are held constant.

The advantage of the hedonic approach is that physical attributes such as location, age and size are introduced into the regression model, and their net contribution to the market price is estimated (Bourassa, Hoesli, & Sun, 2006). Although hedonic regression is found in the literature to provide a good fit with the data (e.g. Goh et al., 2012; Shimizu, Nishimura, & Watanabe, 2010; Wallace & Meese, 1997), the disadvantage is that it requires a lot of data, which are not always available, or can be impractical to obtain. Many of the attributes that can be expected to influence the price of a property, particularly neighbourhood and location variables, are often not available, and other relevant attributes may go undetected (Case, Pollakowski, & Wachter, 1991). Hedonic models are relatively complex to interpret, and require a high level of statistical knowledge and expertise (Bourassa et al., 2006). The fact that there are many free parameters available to be tuned also increases the risk of overfitting (see Heene, Coyne, Francis, Maguire, & Maguire, 2014).

2.2. Repeat-sales

The repeat-sales method is another popular house price index technique that controls for the heterogeneity of properties. The method, originally developed by Bailey, Muth, and Nourse (1963), and further enhanced by Case & Shiller (1987), holds house quality constant by measuring the same asset in two different periods. As a result, there is no need to include the property attributes in the model; transaction prices and property address are sufficient. This index methodology has evolved into the most widely used and reported US house price index.

One drawback of this approach is that, because repeat-sales models consider only dwellings with multiple transactions, they require large amounts of data stretching back in time (de Vries, de Haan, van der Wal, & Marin, 2009). Only a fraction of transactions at any given time period will have matching historical sales, and this sample may not be representative of the market as a whole, leading to aggregation biases (Dombrow, Knight, & Sirmans, 1997).

For example, frequently transacted houses may have some idiosyncratic characteristics that make the owner eager to sell (Sommervoll, 2006). In contrast, frequent transactions might equally indicate that a property has some characteristics that make it easy to resell. Further complicating matters, an analysis carried by Case, Pollakowski, and Wachter (1997) suggests that frequently resold houses tend to appreciate more than those that are less transacted. Short holding periods may indicate significant renovation activity has occurred between sales, therefore violating the assumption of constant quality. Costello (2000) demonstrates that the accuracy of repeat-sales indexes improves significantly when long holding periods (more than one year) are used in estimation of repeat-sales indexes.

There are a number of other weaknesses associated with repeat-sales indexes. One of the most serious is revision, which means that past values of the index are perturbed and revised by present-day information (Baroni, Barthélémy, & Mokrane, 2005). Additional sales reverberate on the index values because new sales pairings provide information on price movements which go beyond the information originally available.

2.3. Central-price tendency methods

The idea of central-price tendency models is that, by aggregating large amounts of data, random noise will naturally tend to cancel out following the law of large numbers, leaving a reliable signal. This approach is far less data-intensive than either hedonic regression or repeat-sales, requiring neither detailed information about properties, nor extensive historical data-sets. One feature that central-price tendency methods do assume is that the data being aggregated are drawn from the same distribution, and cannot be subdivided into different distributions which might be differentially affected over time.

In the US, the index published by the National Association of Realtors is based on median prices (Bourassa et al., 2006). Although such indexes are simple to construct, there is little control for robustness (Case & Shiller, 1987).

Central price tendency models are often criticised as they do not control for the attributes of houses sold either directly in estimation, or indirectly by sample selection (Goh et al., 2012). This can result in inaccurate indexes, susceptible to variations in the mix of houses sold from period to period in a particular region.

Richards and Prasad (2008) argue that stratifying the full sample by suburb, and then taking the simple average of the median sale prices across each suburb, yields price index estimates that are not significantly different from hedonic regression. Given the effectiveness of this strategy for stratification, Richards and Prasad (2008) suggest that the marginal benefits of the more complex and data-intensive methods, such as hedonic regression and repeat-sales, are not justified.

2.4. Comparison of approaches

Goh et al. (2012) directly compared these three different strategies and concluded that hedonic regression models give the best performance. Two variants of the hedonic model were used, namely the standard explicitly intertemporal model and the 'imputed' cross-sectional model. The latter was found to outperform all other index models, matching the findings of previous studies (e.g. Diewert & Hendriks, 2011). Schwann (1998) observed that price indexes constructed using standard hedonic regression are the most robust to finer levels of temporal and geographic disaggregation. He also proposed a time series model employing a stochastic structure for hedonic parameter evolution, which achieves further stabilisation in sparse markets.

The mix-adjusted median was the next best performer in Goh et al.'s (2012) study, with repeat-sales faring the worst, which, given its prominence in the evaluation of the US housing market, is surprising. Shimizu et al. (2010) found that the repeat-sales approach measures market turning points later than the hedonic approach, the former being more than two years delayed in the case of the Tokyo housing market. Wallace and Meese (1997)

also concluded that repeat-sales and other hybrid methods produce less reliable estimates of price movements than the hedonic approach.

Goh et al.'s (2012) results reject the null hypothesis of equality between mean hedonic characteristics for the samples of single-sale and repeat-sale dwellings, revealing that repeat-sales are not representative of the market in general. Houses sold more than once are significantly smaller, have fewer amenities and are of poorer quality, supporting the observation that repeat-sale dwellings are generally sold at a discount to non-repeat-sales.

Goh et al. (2012) reported that, although the performance of the mix-adjusted median was merely 'modest', the method deserves some credit because of its simplicity and transparency. Because it assumes that all houses in a given location stratum are drawn from the same distribution of hedonic quantities, there is no need to identify hedonic attributes of individual houses, collect large amounts of data or carry out any esoteric statistical procedures. Goh et al.'s (2012) findings support Richards and Prasad's (2008) claim that, in absence of information on hedonic attributes, the mix-adjusted median is likely to be the best alternative. Our aim is to investigate whether the central tendency approach can be enhanced to the point where it can compete with, or even outperform, hedonic regression, as applied to the Irish housing market.

3. Case study: the Irish residential property price index

The Irish property market is an example of a relatively sparse data-set. For the period 2010–2015, there were only, on average, 2,200 transactions per month nationwide, motivating the development of techniques for achieving high levels of robustness from small amounts of data.

Currently, property price changes in Ireland are reported only monthly, more than three weeks into the new month, and only broken down for two subregions, Dublin, and outside Dublin, for apartments and for all properties. The Residential Property Price Index (RPPI) is compiled by the CSO, using a hedonic regression 12-month rolling time dummy model (O'Hanlon, 2011). In addition, the monthly results that are released to the public are based on a rolling average of the previous three months, thus enhancing the smoothness of the time series. However, the disadvantage of such artificial smoothing is that the RPPI loses responsiveness to changing market conditions, and can appear misleadingly precise to observers who are not aware of the use of rolling average.

Currently, there are two significant sources of data available for compiling a house price index in Ireland. The first is mortgage returns, which are filed by all lending agencies for properties whose purchase was partly funded by a mortgage. Irish mortgage lenders are required, under Section 13 of the Housing Act 2002, to submit monthly mortgage returns to the Department of Environment, Heritage and Local Government containing data on both mortgage approvals (occurring where a formal letter of mortgage offer has issued) and mortgage drawdowns (O'Hanlon, 2011).

The advantage of this information source is that it carries detailed information about the property, such as the number of bedrooms, the floor area, year of build, plot size, etc. The disadvantage is that not all properties are purchased with a mortgage, hence the sample is unrepresentative. As property prices rise, more people are in a position to trade down to cheaper properties without a mortgage. In addition, lending restrictions following the property crash have led to an increase in cash transactions: from 2010 to 2014, mortgages

on house purchases fell from 88% to only 50% (Dalton & Moore, 2014). Furthermore, 68% of mortgage returns contain errors, such as a missing year of construction, missing number of rooms or missing plot size. Missing, erroneous and implausible values are imputed by the CSO (O'Hanlon, 2011).

The fact that half of transactions are missing from the mortgage records is not necessarily a problem. If 50% of data is randomly removed from a data-set, it has at most a mild effect on the robustness of any index computed from it, amplifying the standard error by $\sqrt{2}$. What matters more is when the missing data are not a random sample, but have some relationship with the rest of the data, which is not taken into account by the model.

For the mortgage data, it is likely that the missing 50% is not a representative sample. Cheaper investment properties are more likely to be transacted in cash, without a mortgage. By contrast, the purchase of larger family homes is more likely to require a mortgage. For this reason, even if the CSO's hedonic regression achieves high goodness-of-fit statistics, the performance is potentially taking place within a biased sample, meaning that goodness-of-fit is not a reliable measure of robustness.

A second source of available information is stamp duty returns, maintained by the PSRA. This publicly available online database reports the date of sale, sale price and address of every property sold in Ireland since 1 January 2010, with a typical latency of around 10 days. The disadvantage of this information source is that it includes no information on the property. Even the addresses can be unreliable, as Ireland has only recently introduced a postal code system, which is yet to be adopted by the PSRA. Although the large majority of returns are lodged immediately, some are delayed by up to 3 months before being submitted to the National Stamp Duty Office. A final disadvantage is that as well as including market sale transactions, the records also include a small proportion of non-sale transactions (e.g. properties that are inherited), which could potentially bias a house price index because the values involved are much lower.

In the case of stamp duty returns which are delayed, it seems reasonable that the subset of records which get delayed is a random selection: the type and location of property purchased should have no predictable relationship with the issue of whether the associated stamp duty is lodged promptly or not. As regards the non-market transactions, if these occur randomly through time periods and geographic locations, then this noise should tend to cancel out for large data-sets using a central-tendency approach.

According to O'Hanlon (2011), the failure of stamp duty returns to collect details on the characteristics of properties rules out the possibility of carrying out an appropriate level of mix-adjustment. He concludes that the Property Price Register can only be of benefit to users with detailed knowledge of the characteristics of specific properties (such as local inhabitants, local estate agents).

In this article, we investigate the hypothesis that a frugal data-set recording only address, date of sale and sale price is sufficient for deriving an index of equivalent robustness to the RPPI currently produced by the CSO. Addresses can, with high reliability, be converted to GPS locations through freely available mapping systems, such as Google Maps. This geographic positioning should permit mix-adjustment and stratification using appropriate central-tendency strategies.

Heene et al. (2014) have argued that simple models with fewer parameters are better suited to modelling complex phenomena, because they minimise the risk of arbitrary overfitting. If an automated frugal data model can match the performance of the CSO's

hedonic regression it would have many advantages, requiring no labour or expense to maintain, and being available with 10 days latency, rather than 3 or 4 weeks.

But first we need to set the rules by which the competition will be decided: we must define index robustness.

4. Measuring robustness

Despite being of critical importance for research in this area, the issue of robustness has received little attention (Goh et al., 2012).

What does a good index look like? According to Chandler and Disney (2014), it is surprisingly hard to identify what exactly house price indexes are intended to measure. Even the language used by the organisations compiling the indexes is vague. For example, the UK's Office for National Statistics (ONS) states that 'the aim of the ONS House Price Index is to measure the change in the average house price for owner-occupied properties in the UK'. But what does 'average' mean? This ambiguity creates difficulties in assessing the relative accuracy and robustness of different index models.

The 'true' house price trend is unobservable, since identifying 'true' house prices would require measurement of the total stock of housing in the local market (Goh et al., 2012). Wallace and Meese (1997) addressed the problem by assuming that the 'true' index can be proxied by the median house price, though Goh et al. (2012) argue that this is contrary to a large body of literature which argues against the application of the median (e.g. Case & Shiller, 1987; Hansen, 2009).

Case and Szymanoski (1995) and Richards and Prasad (2008) developed methods for comparing various models by directly comparing goodness-of-fit statistics. However, Sommervoll (2006) argues that, due to the risk of overfitting, goodness-of-fit statistics can be misleading, especially where indexes are estimated at high levels of disaggregation or for sparse data. Serious mis-measurements may occur, even in cases where the statistical diagnostic tools like R^2, t-values and standard deviations indicate good explanatory power.

The underlying problem with goodness-of-fit is that it fails to account for complexity: models should somehow be penalised for the number of degrees of freedom they exploit to achieve a certain level of fit. In the light of this, model performance is better evaluated through *forecast error*. One way to test forecast error is to randomly divide a data-set of property transactions into two halves: if the index is robust, both halves should yield the same index value.

Following this idea, Goh et al. (2012) adopt a within-sample cross-validation strategy. They randomly select a 75% subset of transactions and evaluate how well the index computed on this selection predicts the sale price of properties in the other 25% subset. The closer the match, the more robust the index.

A problem with Goh et al.'s (2012) test for robustness is that a single iteration of cross-validation is not reliable. For example, two random halves might by chance produce close agreement, where nearly every other partition would have resulted in diverging values. Specifically, the values returned from a single implementation of the cross-validation technique are themselves drawn from a distribution, with an associated mean and standard deviation. The process must be repeated many times to identify a reliable sampled mean. As the number of partitions n increases, the reliability of the robustness value increases with order $\sqrt{n}$.

While Goh et al.'s (2012) test can provide a weak heuristic for assessing robustness, it cannot provide the basis for a definition, since it is easy to construct an index which is not robust, yet does well at the test. For example, we could hardcode an algorithm that outputs 100 if the number of transactions in the sample is even, and 99.999 if the number is odd. The agreement will always be very high for any random split, and this agreement can be boosted to any arbitrary level by adjusting the hardcoded values. And yet the index is uninformative.

We propose a minor refinement of Goh et al.'s (2012) test, which can serve as a definition for robustness. Given two competing indexes, the more robust index is the one which, when run repeatedly on two random partitions of a given data-set, produces a pair of values which, on average, are closer to each other than those produced by the other index. We also restrict the set of functions to those which vary monotonically with the change in any sale price in the set (i.e. if any sale price is altered, the function output must either stay the same or move in the same direction). To formalise this mathematically, a valid index function is a computable function $i : R^n \rightarrow R$ that is monotonic, i.e. for all $\epsilon > 0$ and for all $x \in R^n$, it holds that $i(x_1 + \epsilon, x_2 + \epsilon, \ldots, x_n + \epsilon) > i(x_1, \ldots, x_n)$. This is close to Goh et al.'s idea of cross-validation prediction, except that it knocks out the pathological examples, as highlighted above, where an index ignores the input, and always produces the same hardcoded output.

In practice, the most robust index is the smoothest index. Our argument is as follows: given an index, some component of the monthly price fluctuation is due to random sampling error, and the remaining component is due to genuine shifts in market sentiment. We want to eliminate as much of the background noise as possible, thus allowing us to tune in to the signal of the market itself. Comparing discrepancies between successive months is similar to Goh et al.'s (2012) idea of comparing different samples drawn from the same month: the goal is to develop an index with the smallest discrepancies.

Changes in market sentiment have a lower frequency than that of background noise: for example, we expect the market to move in cycles, with prices drifting consistently upwards for months, then drifting consistently downwards during a recession (see Agnello & Schuknecht, 2011). In contrast, sampling error stemming from the construction of the index will jump randomly from month to month. While changes in house prices have momentum, sampling error does not (thus explaining why the CSO chooses to publish three-month rolling averages). Because of this differential in frequencies, smoothness acts as an indicator of noise filtering. The smoother the trending of the index (i.e. the greater the extent to which changes in successive months agree with each other), the smaller the noise component, and the higher the reliability of the remaining signal. Accordingly, we will evaluate index robustness in terms of the average absolute monthly change in market momentum; a steadily rising index would have an average change of zero.

According to Wang and Zorn (1997), an index should be defined by its use in practice, rather than by the more complex, higher level concerns of statistics and models. They find that much of the debate over index methodology can be distilled to implicit and largely unrecognised disagreement as to the intended application.

Taking Wang and Zorn's recommendation into account, we can express the above mathematical definition in terms of a clear practical application: the most robust index is the portfolio that investors would naturally seek to hold if house price indexes were openly traded in a prediction market (as recommended by Englund et al., 2002 and Shiller, 2003).

Investors seek to hold a portfolio which is as diversified as possible, thus minimising risk, while maintaining return (see Maguire et al., 2014). For example, a diversified index, such as the S&P 500 index for the US stock market, should have a better risk to reward profile than any of its constituents, or indeed, any subset of its constituents (c.f. Maguire et al., in press). This is why investors seek to hold the S&P 500 index, and why it provides the gold standard for financial models.

Choueifaty, Froidure, and Reynier (2013) propose that portfolio diversification is related to volatility, and can be evaluated by the extent to which independent sources of information combine to smooth the overall volatility of a portfolio. For example, if numerous house price indexes published by different organisations were freely available to trade, investors would naturally hold the portfolio which minimises overall volatility, thus in effect creating a more diversified super-index with a better risk to reward profile than any of its individual constituents. In sum, the optimal house price index, the one that would be most traded and hence most quoted in the media, is the smoothest house price index.

In the following section, we describe an algorithm developed to meet this objective standard for index robustness, which functions on sparse (no long-term historical records) and frugal data (only location, date and price).

5. Algorithm

Our algorithm involves several stages of processing the online data provided by the PRSA. First, we collected all the available data, stretching back from January 2010 to the end of June 2015. Google Maps API provided the best option for geocoding the addresses into GPS co-ordinates. The service has a rate limit of 2,500 requests per day, so the process was carried out automatically over a period of 2 months.

Approximately 90% of addresses were successfully converted, giving us the GPS co-ordinates, date of sale and sale prices for 147,635 unique transactions. These transactions were analysed in monthly sets, with January 2010 providing the base index of 100. There were considerable differences in the number of transactions per month, from a low of 677 in January 2011, to a high of 3,894 in December 2014.

The first index we calculated was based simply on the raw average property price for each month. The time series of price changes for this index had an average monthly change in momentum of 12.40%. Subsequently, we computed the raw monthly median. The monthly shift in momentum of this index was lower than that of the raw average, at 8.42%.

5.1. Stage 1: Filtering

The stamp duty return data show that whole housing estates and blocks of apartments are sold in bulk at the same time, greatly distorting the average price in a given month. The next stage of the algorithm was to remove these distortions.

Making use of the geographic co-ordinates, we eliminated any property transaction for which there was another transaction within 100 m in the same period of 48 h. This eliminated all bulk sales, with the number of valid transactions being reduced by 14.4% to 126,444. The average monthly change in momentum of the median of this filtered subset of transactions was lower again, at 6.37%.

5.2. Stage 2: Mix-adjustment through proximity voting

A potential problem with median-based approaches is that fluctuations in the relative number of properties sold in unrepresentative regions can have a dramatic effect on the median, even when there is no increase in price. For example, if twice as many properties in Dublin are sold as usual, a region in which the value of most properties is higher than the national median, these sales will act to drag the median upwards, despite no actual change in price.

Richards and Prasad (2008), for example, found that, based on a database of 3.5 million transactions in the six largest Australian cities, compositional shifts between higher and lower priced parts of cities led to much volatility in unadjusted median prices. Similarly, realtors in the US report that median house prices rise in the summer: most families with children, who typically buy more expensive homes, time their purchase based on school year considerations (Richards & Prasad, 2008).

To enhance robustness, it is important to control for the mix of properties which are sold in any particular month. What we want is to identify a subset of the given sample which is more representative of the houses in the market as a whole. Specifically, we want the analysis set to be as spatially autocorrelated as possible with the set of historical records, featuring the same relative distributions of transactions in different regions of the country, and the same types of properties within those regions. Spatial autocorrelation arises in housing data due to the proximity of units that are the same or among contiguous units (Hamid, 2001). In general, properties in close proximity tend to have similar structural characteristics, such as size of living area, dwelling age and design features (Ismail, 2006). The similar quality of proximate properties is a natural consequence of the fact that they tend to be developed at the same time (Gillen,Thibodeau, & Wachter, 2001). Residents in the same neighbourhood may also follow similar commuting patterns, and share the same neighbourhood amenities such as public schools and shopping centres (Ismail, 2006).

In light of this, we developed a system for enhancing autocorrelation based on geographical proximity to a historical target set of transactions. Specifically, we eliminate the 10% least representative properties from the sample, using a single transferrable voting system. The system operates as follows.

Let N be the entire set of filtered property transactions from Stage 1. Let n be the set of properties transacted in the current month. Each property in N votes for the nearest property to it in the set n.

If any property in n exceeds the threshold for election of $\frac{|N|}{0.9|n|}$ votes, then it is elected from the set; any excess votes are redistributed to its nearest neighbour. Subsequently, the property with the least number of votes is eliminated and its votes are redistributed in a similar manner. The process continues until all properties in n have either been eliminated or elected. In the end, 90% of the properties in n will be elected. This algorithm ensures there will be roughly the same number of properties included from each geographic location. In addition, because the same kinds of houses tend to be located beside each other (e.g. detached bungalows, three-bed semi-detached, apartments), the algorithm should also ensure a representative quantity of each type of dwelling.

The average monthly change in momentum of the adjusted-filtered median index was lower again, at 4.47%.

5.3. Stage 3: Localised stratification

Mix-adjustment alone is not sufficient for maximising stability between months. The reliance on a median ignores all information about the distribution below and above the median value, effectively ignoring the shape of the distribution. If this shape varies between months, such information is overlooked, thus passing up on an opportunity to enhance stability. This kind of situation arises when different regions have different medians, and the prices in these regions are diverging.

For example, the national mix-adjusted median house price at the start of 2015 was €180,000. However, because capital cities are more expensive, a large proportion of homes in Dublin were sold for more than this value (85%). The issue that hence arises is that any fluctuations in house prices that are unique to Dublin will have little impact on the overall national median.

In contrast, regions whose median price is closest to the national median will have a disproportionate effect on influencing the national median price. These areas are contributing too much information, while other areas are contributing too little. This reduces robustness, and increases volatility between months.

Accordingly, Goh et al. (2012) take the view that disaggregation of data along geographic lines is extremely important when constructing house price indexes. Studies from the Australian housing market, for example, reveal the existence of marked geographical differences in the behaviour of house prices across metropolitan areas (see Costello, Fraser, & Groenewold, 2011; Hatzvi & Otto, 2008).

One way to address this issue is through stratification. Richards and Prasad (2008) proposed a novel stratification method and tested it on an Australian data-set. They grouped together suburbs according to the long-term average price level of dwellings in those regions, taking the equally weighted average of the medians for each stratum. This measure of price growth was found to improve substantially upon an unstratified median, and was very highly correlated with regression-based measures (see also McDonald & Smith, 2009).

One limitation with Richards and Prasad's (2008) stratification technique is that it imposes arbitrary strata. The point at which a property shifts from being in one stratum to another is completely arbitrary. There is no guarantee that the strata Prasad and Richards selected reflect the most pertinent or delineated divisions in the market. Over time, these strata might shift, with more houses being built in one region than another, or a particular area being improved due to redevelopment projects. The possibility of changing relationships between the strata is not accommodated by Richards and Prasad's (2008) approach.

Our simple solution is not to impose any arbitrary stratifications, but to derive a different local base for every single property. The algorithm proceeds as follows: Two months of transaction records are selected, a stratification-base and the current month to be evaluated. We divide each sale price in the current month by that of the closest property in the stratification-base, giving a set of ratios. We then take the median of this set. This is the stratified-adjusted-filtered median.

For example, consider a house that is sold in Donegal for €105 K in February 2015. The closest house to it sold in January 2015 for €120 K. So we turn €105 K into $.875 - 1 = -12.5\%$. Alternatively, consider a house that is sold in Dublin for €420 K in February

2015. The closest house to it sold in January 2015 for €360 K. So we turn €420 K into $1.167 - 1 = +16.7\%$. Now take the median of all the percentage change values.

Under this system, all areas contribute equally to the index, thus reducing volatility. Choosing the stratification-base by default as the previous month, the average monthly change in momentum comes out at 3.76%.

5.4. Stage 4: Multiple base-month calibration

We are not limited to using only a single month as a stratification-base: we can run the same algorithm using different historical bases. For example, we can derive the index value for January 2015 using December 2014 as the stratification-base, November 2014, October 2014 and so forth.

As an example, Table 1 displays the stratified-adjusted-filtered median price change for January 2015 using the six preceding months as stratification-bases.

We recomputed the index by calculating monthly change using every available historical stratification base and averaging them. Using multiple base-month calibration, the average absolute monthly change in momentum of the stratified-adjusted-filtered median index was lower again, at 2.83%.

Table 2 displays descriptive statistics for the time series of price changes from January 2010 to July 2015 that results following the various stages of the algorithm, with the RPPI for comparison. Note that the mean and median are based on absolute monthly change, while 'smooth' refers to the average absolute monthly change in momentum.

5.5. Comparison with CSO index

Our frugal index achieved a 'smoothness' of 2.83%, which is a slightly lower level of volatility than the CSO's RPPI index, which had a 'smoothness' of 3.35% for the same period. For example, the maximum monthly change between any consecutive months for our frugal index was –6.7% for December 2012–January 2013, while the largest jump for the RPPI was a jump of +8.1% between November and December 2012. The correlation between the monthly changes of the two indexes was only $r = .43$, suggesting that they contribute slightly different sources of information.

Quigley (1995) found that hybrid models which combine information from repeat-sales and hedonic regression can be even more robust than either method in isolation. Our findings support this idea: when the two indexes are optimally weighted to minimise the smoothness value of the resulting composite index (56.1% for the frugal index, 43.9% for the RPPI), the resulting monthly change in momentum drops to only 2.51%. For the sake of comparison, the average monthly change in momentum of the RPPI's 3-month rolling average is .76%, while that of the 12-month rolling average is .25%.

Figure 1 plots the two time series, frugal and RPPI, plus their minimised volatility composition. Although our frugal index is more robust than the RPPI produced by the

Table 1. Price change between Dec 2014 and Jan 2015 using different stratification-base months.

	Jul 14	Aug 14	Sep 14	Oct 14	Nov 14	Dec 14
Change Jan 15	+5.7%	+3.9%	+3.8%	+2.7%	+2.2%	+6.0%

Table 2. Descriptive statistics for price change index produced at various stages.

	Mean (%)	Median (%)	Max(%)	Min (%)	StDev (%)	Smooth (%)
Raw average	7.01	5.74	+31.1	−17.9	9.23	12.40
Raw median	5.06	4.07	+23.8	−15.2	6.79	8.42
Stage 1	3.85	3.23	+11.1	−15.6	4.81	6.37
Stage 2	2.58	2.70	+9.19	−7.38	3.31	4.47
Stage 3	2.72	2.22	+7.33	−8.38	3.38	3.76
Stage 4	2.05	1.64	+5.41	−6.67	2.55	2.83
RPPI	2.16	1.61	+8.06	−5.50	2.73	3.35

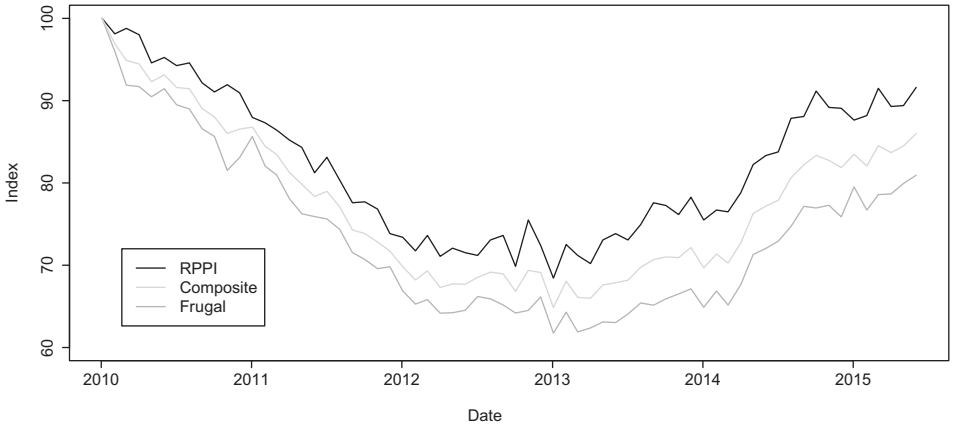

Figure 1. RPPI, frugal and composite indexes from January 2010 to June 2015.

CSO, the composite index is the most robust of all, and is what investors would choose to hold if both indexes were available to trade in an open market. By splitting their investment 56–44, investors would maximise the risk to return profile of their portfolio, and create a more robust index in the process.

6. Conclusion

We have shown that, contrary to the assertions of O'Hanlon (2011), the frugal data available from stamp duty returns, namely sale price, date of sale and address, are sufficient for developing an index that matches and exceeds the robustness of the CSO's RPPI, which relies on recording a multitude of characteristics for each property.

Admittedly, our frugal index doesn't improve greatly on the existing RPPI (though further refinements may lead to enhanced performance). The main advantage of our novel algorithm is the ease and flexibility with which it can be implemented. The code can be run on any database containing property prices and locations. It automatically controls for outliers, noise and data-set bias. As soon as new data become available, the index can be recomputed instantly with no overhead. It can also be applied to houses that have not been sold yet, using their asking prices to anticipate future changes in sale price. Because the algorithm is completely automated, it also allows users to analyse changes for any subset of records (e.g. by province, county or any arbitrarily selected geographical area).

The speedy measurement of changes in house prices is of great importance to policy-makers and investors, and is also crucial to understanding the operation of the housing market. Empirical evidence suggests that mobilising real estate derivative markets brings about significant economic benefits in the form of rapid adjustments towards supply–demand equilibriums in housing markets, lower rents on real estate and reduced amplitude of speculative house price movements (Englund et al., 2002; Lacoviello & Ortalo-Magné, 2003; Quigley, 1999). Our algorithm could be used to support derivative markets by providing an objective means of deciding a target outcome to be speculated on, one which can be recomputed hour by hour.

Critics of our frugal approach may argue that, over the period of decades, carefully calibrated statistical techniques provide a clearer picture of gradual changes in the market. This may well be the case. However, it can also be argued that an important goal of a house price index is to communicate immediate changes in market sentiment. According to Wang and Zorn (1997), there is little value in pursuing a goal of statistical or modelling accuracy if it does not lead to improved decision-making and better economic outcomes. Short- and medium-term price fluctuations can have significant impact on government and market participants, as reflected by frequent media headlines (e.g. are prices currently rising? has the market bottomed out? is this the right time to buy?) We have provided a proof of concept that algorithms using sparse and frugal data can fill this niche, providing market participants with reliable up-to-date information on house price fluctuations.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Phil Maguire is a lecturer in Computer Science in the National University of Ireland, Maynooth. His research explores portfolio optimisation, predictive modelling and the foundations of measurement.

Robert Miller is an undergraduate student at National University of Ireland, Maynooth taking the degree in Computational Thinking. He has been awarded the Alan Turing and Stephen Cook prizes in Computational Thinking and the Delort and McMahon prizes in mathematics.

Philippe Moser is a lecturer in Computer Science in the National University of Ireland, Maynooth. His research interests include algorithmic information theory, randomness, computability, complexity theory and computational finance.

Rebecca Maguire is a lecturer in Psychology in the National College of Ireland. Her area of expertise is the cognitive modelling of uncertainty.

References

Agnello, A., & Schuknecht, L. (2011). Booms and busts in housing markets: Determinants and implications. *Journal of Housing Economics, 20*, 171–190.

Bailey, M., Muth, R., & Nourse, H. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association, 58*, 933–942.

Baroni, M., Barthélémy, F., & Mokrane, M. (2005). Real estate prices: A Paris repeat sales residential index. *Journal of Real Estate Literature, 13*, 303–322.

Bourassa, S., Hoesli, M., & Sun, J. (2006). A simple alternative house price index method. *Journal of Housing Economics, 15*, 80–97.
Case, B., Pollakowski, H., & Wachter, S. (1991). On choosing among house price index methodologies. *Real Estate Economics, 19*, 286–307.
Case, B., Pollakowski, H., & Wachter, S. (1997). Frequency of transaction and house price modeling. *The Journal of Real Estate Finance and Economics, 14*, 173–187.
Case, K., Quigley, J., & Shiller, R. (2005). Comparing wealth effects: The stock market vs. the housing market. *Advances in Macroeconomics, 5*. Retrieved from https://www.degruyter.com/view/j/bejm.2005.5.1/bejm.2005.5.1.1235/bejm.2005.5.1.1235.xml
Case, K. E., & Shiller, R. J. (1987, September/October). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review,* 45–56.
Case, B., & Szymanoski, E. J. (1995). Precision in house price indices: Findings of a comparative study of house price index methods. *Journal of Housing Research, 6*, 483–496.
Chandler, D., & Disney, R. (2014). *Measuring house prices: A comparison of difference indices*. Institute for Fiscal Studies, Briefing Note BN146. Retrieved from http://www.ifs.org.uk/bns/bn146.pdf
Choueifaty, Y., Froidure, T., & Reynier, J. (2013). Properties of the most diversified portfolio. *Journal of Investment Strategies, 2*, 49–70.
Costello, G. (2000). Pricing size effects in housing markets. *Journal of Property Research, 17*, 203–219.
Costello, G., Fraser, P., & Groenewold, N. (2011). House prices, non-fundamental components and interstate spillovers: The Australian experience. *Journal of Banking & Finance, 35*, 653–669.
Costello, G., & Watkins, C. (2002). Towards a system of local house price indices. *Housing Studies, 17*, 857–873.
Dalton, P., & Moore, K. (2014). *How to quickly adapt to new policy needs? The experience of the central statistics office, Ireland in developing house price indicators*. Cork: Central Statistics Office.
de Haan, J., & Diewert, W. E. (2011). *Handbook on residential property price indexes*. Luxembourg: Eurostat.
de Vries, P., de Haan, J., van der Wal, E., & Marin, G. (2009). A house price index based on the SPAR method. *Journal of Housing Economics, 18*, 214–223.
Diewert, W. E., & Hendriks, R. (2011). The decomposition of a house price index into land and structures components: A hedonic regression approach. *The Valuation Journal, 6*, 58–105.
Dombrow, J., Knight, J., & Sirmans, C. (1997). Aggregation bias in repeat-sales indices. *The Journal of Real Estate Finance and Economics, 14*, 75–88.
Englund, P., Hwang, M., & Quigley, J. (2002). Hedging housing risk. *The Journal of Real Estate Finance and Economics, 24*, 167–200.
Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics, 50*, 1243–1255.
Gillen, K., Thibodeau, T. G., & Wachter, S. (2001). Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics, 23*, 5–30.
Goh, Y., Costello, G., & Schwann, G. (2012). Accuracy and robustness of house price index methods. *Housing Studies, 27*, 643–666.
Gupta, R., & Hartley, F. (2013). The role of asset prices in forecasting inflation and output in South Africa. *Journal of Emerging Market Finance, 12*, 239–291.
Gupta, R., & Kabundi, A. (2010). Forecasting macroeconomic variables in a small open economy: A comparison between small- and large-scale models. *Journal of Forecasting, 29*, 168–185.
Hamid, A. M. (2001). *Incorporating a geographic information system in hedonic modelling of farm property values* (PhD ed.). Lincoln University, Christchurch.
Hansen, J. (2009). Australian house prices: A comparison of hedonic and repeat-sales measures. *Economic Record, 85*, 132–145.
Hatzvi, E., & Otto, G. (2008). Prices, rents and rational speculative bubbles in the Sydney housing market. *Economic Record, 84*, 405–420.
Heene, M., Coyne, J., Francis, G., Maguire, P., & Maguire, R. (2014). *Crisis in cognitive science? Rise of the undead theories*. Proceedings of the 36th Annual Meeting of the Cognitive Science Society.

Ismail, S. (2006). Spatial autocorrelation and real estate studies: A literature review. *Malaysian Journal of Real Estate, 1*, 1–13.
Kain, J., & Quigley, J. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association, 65*, 532–548.
Lacoviello, M., & Ortalo-Magné, F. (2003). Hedging housing risk in London. *The Journal of Real Estate Finance and Economics, 27*, 191–209.
Leishman, C. (2009). Spatial change and the structure of urban housing sub-markets. *Housing Studies, 24*, 563–585.
Lyons, R. C. (2015). East, West, boom and bust: The spread of house prices and rents in Ireland 2007–2012. *Journal of Property Research, 32*, 77–101.
Maguire, P., Kelly, S., Moser, P., & Maguire, R. (in press). Further evidence in support of the low volatility anomaly: Optimizing buy-and-hold portfolios using historical volatility. *Journal of Asset Management*.
Maguire, P., Moser, P., O'Reilly, K., McMenamin, C., Kelly, R., & Maguire, R. (2014). *Maximizing positive portfolio diversification*. IEEE Computational Intelligence for financial Engineering & Economics (CIFEr) Conference, London, 174–181.
McDonald, C., & Smith, M. (2009). *Developing stratified housing price measures for New Zealand*. Wellington: Reserve Bank of New Zealand, DP2009/07.
Munro, M. (1987). Intra-urban changes in housing prices: Glasgow 1972–1983. *Housing Studies, 2*, 65–81.
O'Hanlon, N. (2011). Constructing a national house price index for Ireland. *Statistical and Social Inquiry Society of Ireland, 40*, 167–196.
Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling, 45*, 259–267.
Quigley, J. M. (1995). A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics, 4*, 1–12.
Quigley, J. M. (1999). Real estate price and economic cycles. *International Real Estate Review, 2*, 1–20.
Richards, A., & Prasad, N. (2008). Improving median housing price indexes through stratification. *Journal of Real Estate Research, 30*, 45–75.
Schwann, G. M. (1998). A real estate price index for thin markets. *The Journal of Real Estate Finance and Economics, 16*, 269–287.
Shiller, R. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives, 17*, 83–104.
Shimizu, C., Nishimura, K. G., & Watanabe, T. (2010). Housing prices in Tokyo: A comparison of hedonic and repeat sales measures. *Jahrbücher für Nationalökonomie und Statistik* [Journal of Economics and Statistics]*, 230*, 792–813.
Sommervoll, D. (2006). Temporal aggregation in repeated sales models. *The Journal of Real Estate Finance and Economics., 33*, 151–165.
Stock, J., & Watson, M. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23*, 405–430.
Wallace, N., & Meese, R. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics, 14*, 51–73.
Wang, F. T., & Zorn, P. M. (1997). Estimating house price growth with repeat sales data: What's the aim of the game? *Journal of Housing Economics, 6*, 93–118.