A house price index defined in the potential outcomes framework*

Nicholas T. Longford
SNTL and UPF, Barcelona, Spain

Abstract

Current methods for constructing house price indices are based on comparisons of sale prices of residential properties sold two or more times and on regression of the sale prices on the attributes of the properties and of their locations. The two methods have well recognised deficiencies, selection bias and model assumptions, respectively. We introduce a new method based on propensity score matching. The average house prices for two periods are compared by selecting pairs of properties, one sold in each period, that are as similar on a set of available attributes (covariates) as is feasible to arrange. The uncertainty associated with such matching is addressed by multiple imputation, framing the problem as involving missing values. The method is applied to a register of transactions of residential properties in New Zealand and compared with the established alternatives.

Key phrases: *Hedonic regression; house prices; matching; potential outcomes; propensity scoring; repeat-sales method.*

JEL Classification: C1 (Econometric and Statistical Methods): C13 — Estimation; C15 — Simulation methods. C3 (Multiple or Simultaneous Equation Models): C31 — Treatment effect models. E3 (Prices, Business Fluctuations, and Cycles): E31 — Price level; inflation; deflation.

*N. T. Longford, SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain; email: NTL@sntl.co.uk

1 Introduction

The prices of residential properties have an important role in the economy of every developed country. They are a key factor in the decision of an individual or household to buy for the first time, to sell the property they currently own and buy another, to carry on renting, to purchase a property as an investment or as a source of rental income, and the like. The construction, real estate and finance industries, the government and all other services associated with housing closely monitor the movement of house prices, as do the authorities in charge of collecting taxes that are derived, directly or indirectly, from property transactions and ownership. The purpose of a house price index is to enable such monitoring by summarising the changes of the house prices over time. A house price index is defined as a table, one dimension of which is time, and the others, if any, may be geographical areas (regions or districts), property types (detached, semidetached, apartment, and the like) and circumstances of the transaction (first-time buyer, purchase of a newly built property, freehold, and the like). To avoid convoluted expressions, in motivation we refer to a two-way index compiled annually for three regions of the country. An example is given in Table 1 and, graphically, in Figure 1.

Each entry of the index (e.g., 110.2 in the first row) is interpreted as a comparison of the house prices in the year concerned (2002) with the reference year (2000) in a region (A). Thus, house prices in region A have risen by 10.2%, or a house sold in 2000 in region A for a given amount v can be expected to sell in 2002 for $1.102v$. These two statements are not equivalent because the first compares average prices in the two years, whereas the second concerns a single property, or the average over a given set of properties. We forego these and some other ambiguities by defining the key terms related to house price indices.

A typical house price index is associated with a geographical area (a country) and its *housing stock*. The housing stock is defined as the collection (set) of all dwellings that serve the purpose of the primary residence (home) of a single household. The definition has to be supplemented by small print that arbitrates whether the housing stock also contains dwellings that are still in the

Table 1: A house price index for a country and its three regions. Fictitious data.

	Year							
Region	2000	2001	2002	2003	2004	2005	2006	2007
A	100.0	103.7	110.2	119.6	134.6	141.1	158.4	167.8
B	100.0	106.2	109.1	122.5	138.0	149.5	161.3	172.3
C	100.0	103.1	107.9	118.2	130.7	144.4	155.8	163.0
Country	100.0	103.8	109.0	119.4	133.4	144.0	157.7	166.4

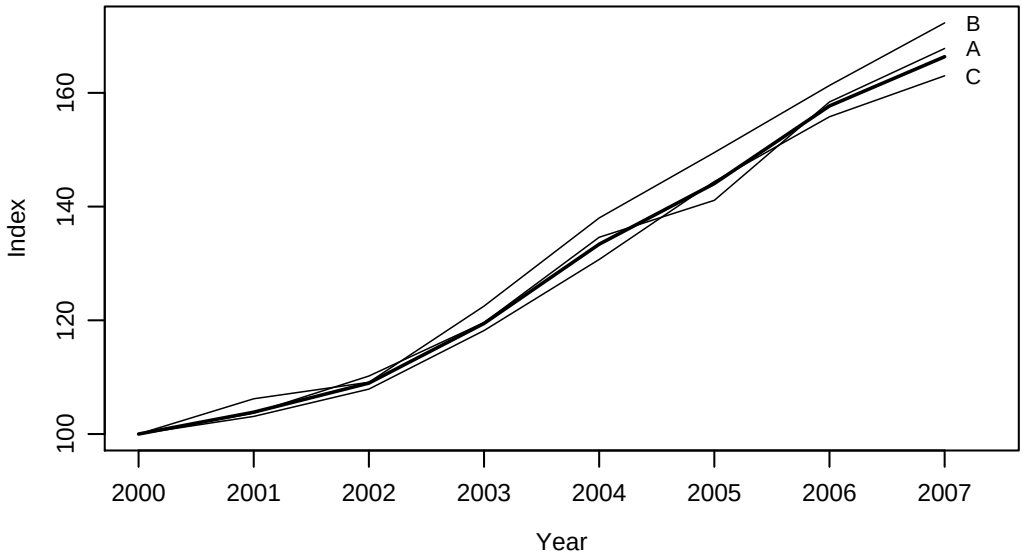

Figure 1: The house price index for the three regions of a country. Data in Table 1.

process of construction, but with parts already habitable, that are in a desolate state (and would fail a regulatory inspection) but are still occupied, that apart from serving as a home are also used for pursuing a business or trade, and the like. The definition of the term 'household' may require similar attention, although the attribute 'single-household' may be associated with a property, not with its occupants. In a country that has a comprehensive register of all residential properties, these definitions may be formulated administratively.

The purpose of a house price index also requires a careful specification. Each entry of an index is a comparison, of a given year and housing stock with a *reference year*. In Table 1, the reference year is 2000, implied by the entries equal to 100.0. A straightforward comparison of the mean sale prices in any two years is problematic because the properties that were sold in the two years may differ systematically. For example, smaller properties may have become relatively more sought after, because of an increase in the number of small (single-member) households and reduced availability of such properties.

A comparison of like with like is highly desirable, but what constitutes similarity of properties is difficult to specify. A residential property has a multitude of salient attributes, and many of them are not recorded in a typical database. Among them are not only the attributes of the structure (size, design and structural quality of the building, insulation and heating, state of disrepair and decoration), but also of the location (employment and leisure opportunities, quality of schools, the local rates of taxation and absence of crime and dilapidation), environment (level of pollution, including noise), surroundings and neighbourhood (garden and other land that belongs to the property, fencing, access

to the roads and local services) and the general ambience.

Some of these attributes are difficult to measure or assess, and their perceived importance, as well as the attributes themselves, may change over time. Every property is subject to wear and tear (including occasional catastrophic damage), but also maturation (development of good relationships among the neighbours; growth of the trees and hedges in the gardens and common spaces), and properties occasionally undergo renovation. Substantial structural changes to a property are subject to local authority approval, and so they are recorded, but not necessarily in the analysed database. Minor changes, such as the installation of a security alarm, new doors and double glazing, are not recorded. There are also fashions or fads, such as preference of some buyers for properties at the end of a cul-de-sac or with nice views, which tends to be stronger in some periods (and locations) than in others. Further, the dispositions of the buyer and seller and their skills in the process of negotiating a transaction have an impact on the sale price, as does the availability of similar properties for sale at the time in the neighbourhood. In brief, it is difficult to establish that two properties, as assets, are alike, even if they are adjacent, were constructed at the same time and have (or originally had) the same design.

Notwithstanding these problems, an index aims to compare the sale prices of a typical property, or a collection of properties, *if* they were sold in one year as well as in the other. We might argue that a comparison would be more appropriate if the second transaction were not informed by the outcome of the first, but such 'blinding' is impossible to arrange. The collection of properties for which the index is intended is called the *reference housing stock* (RHS). For example, the RHS for the index in Table 1 is the set of all residential properties sold (at least once) in year 2000. It may be practical to define a RHS that is much smaller, but it is essential that it be a good representation of the overall housing stock, akin to a simple random sample. The properties sold need not satisfy this criterion.

An index may have a universal (fixed) RHS or this housing stock may be updated from time to time. A fixed RHS has the advantage that the comparison for any two years is straightforward; for example, the comparison of 2006 with 2002 in Table 1 is 158.4/110.2=143.7, with reference to the housing stock in 2000. However, a RHS may become outdated after some years, faster in countries that undergo substantial economic and social changes. For instance, increased prosperity, urbanisation and the change in the distribution of household sizes may lead to substantial changes in the housing stock, making a long-established RHS unsuitable for any comparison spanning a few years. The introduction of a new reference breaks the continuity and makes any comparison for one year before and one after the change problematic. In the UK, the HBOS (formerly Halifax) House Price Index was established in the early 1980's, and its current RHS comprises the properties for which the then Halifax Building Society granted a mortgage in 1984.

The calculation of a house price index may be based on all the properties in the housing stock that were sold in the year concerned, a random (or *designed*) sample, or a convenient sample of such

properties. It is more appropriate to refer to *transactions*, because a property may be sold more than once in a period. Transactions have some relevant attributes that may qualify or disqualify them from consideration. For example, an index may be confined to first-time buyers or new properties, and it is reasonable to exclude any transactions that involve some constraints, such as a sitting tenant, duties to conduct certain activities, and the like. Also, some properties are purchased with the intention to renovate or redevelop them (to increase their value) and sell them at a higher price. Therefore, even the purpose of the purchase is a relevant attribute.

We draw a distinction between an index as a population quantity, and its estimator and estimate based on the (available) data. The former is a rather abstract quantity, because it refers to hypothetical sales, of *all* the properties in the RHS, in the year in question, *and* without one transaction influencing another in any way. Instead of a physical list (enumeration) of properties, a RHS may be given by the joint *distribution* of its property attributes.

A survey of transactions is difficult to design because no obvious sampling frame is available and a high nonresponse rate is likely when inquiring about sensitive financial information. Sampling from the housing stock is very ineffective because only a small fraction of properties is sold in any one year. Kiel and Zabel (1997) explore the feasibility of using the American Housing Survey for constructing a house price index. The HBOS and Nationwide House Price Indices (UK) are two examples of indices based on convenience sampling. They use the data from the mortgage approvals of the respective institutions. In the recent years, each institution granted about 20% of all the mortgages in the UK. Thus, these indices ignore properties that were paid for in full, and entail the assumption that their agents and branches cover the country uniformly and attract the business related to a representative sample of transactions.

A transaction is not an instant process, and so the date associated with a transaction has to be carefully defined. The date when the contracts are exchanged between the buyer and the seller might seem to be the most appropriate. However, the value of a house price index is greatly enhanced by its timely publication. More precise data may be collected from later stages of the transaction process, but an index based on it, being somewhat out of date, is less valuable. The UK Land Registry publishes a house price index for England and Wales, based on registration of all the transactions of residential properties. It lags behind the HBOS and Nationwide Indices because it captures data about transactions at a later stage, when the buyers register as the new owners.

Our discussion indicates that any house price index is a compromise. First, the target of its estimation is a comparison of two monetary amounts that cannot be realised in practice, because a property cannot be put up for sale without the owner's intent, and neither can its purchase and its circumstances be assigned (controlled) by design. Next, the value of a property as an asset is not a constant, because the property deteriorates with use, is improved by maintenance and renovation, and is, in effect, altered in concert with the attributes of its location and environment. The (temporal)

economic environment in the area, the country, or even internationally, is also an important factor. And finally, even if properties remained constant in all relevant aspects, their prices might change with time unevenly, because of changing trends in the preferences, postures and circumstances of the buyers and sellers and their representatives (agents). A clear statement of what an index is intended to capture and transparency in how its data are collected and processed is therefore essential for its meaningful interpretation.

The next section reviews the established methods for house price indices, and the following section introduces our proposal based on the potential outcomes framework. Section 4 discusses the assumptions of the framework as they relate to our proposal. An application is presented in Section 5, estimating the entries of a house price index in six city districts of New Zealand in 2006.

2 Methods for constructing house price indices

A simple index may be constructed from the within-period means or medians (summaries) of the house prices. Let $\hat{\mu}_{kh}$ be the summary of the transactions in year k and region h, and suppose year 0 is the reference. Then $\hat{\mu}_{kh}/\hat{\mu}_{0h}$ is the (estimated) entry in the index. When μ_{kh} are linear functions of the sale prices in the transactions, the entries in the margins, μ_k, are evaluated or estimated by combining the summaries μ_{kh} ($\hat{\mu}_{kh}$), $h = 1, \ldots, H$, regarding the regions h as (homogeneous) strata. Such stratification can be applied even when the index entries are not reported for the regions. The weights accorded to the regions may change from year to year, reflecting the changing population sizes of the strata. The Australian Bureau of Statistics publishes a house price index for the capital cities of the country (the state capitals and the national capital, Canberra) based on this approach. The entries of the UK Land Registry Index are based on the median house prices in the local authorities and London boroughs. Apart from these simple approaches there are two established methods, based on hedonic regression and repeated sales.

In hedonic regression, a model is posited for the sale price in a transaction in terms of a set of (available) covariates, including the date of the transaction, and the average sale price for a RHS is estimated by prediction. There is a concensus in the literature that the model should be formulated for the logarithm of the sale price, because the convenient assumptions of normality are then much more palatable and linear models more suitable. Hedonic regression was originally developed for the second-hand car market (Court, 1939). Early applications to house prices include Rosen (1974) and Goodman (1978). Harrison and Rubinfeld (1978) applied hedonic regression to the median sale prices of the residential properties in the census tracts of the Boston Metropolitan Area (U.S.A.) in 1974. The purpose of the analysis was to estimate the marginal value of the clean air in the transactions. The analysis was subjected to scrutiny by Belsley, Kuh and Welsch (1980) who showed that the

assumption of (log-) normality is not satisfied. A reanalysis by Longford (1993) pointed out the relative within-district homogeneity of the log-prices.

Hedonic regression methods demand data of high quality, with records of several important covariates. The models rely on their validity and on completeness of the list of covariates. The latter assumption is untestable and easily dismissed by a pessimist, because there are bound to be some covariates that are not recorded. Also, the values of some of the covariates may be subject to some distortion (measurement error or misclassification). Databases, usually registers of transactions, may contain some errors in the sale prices, some covariates may have inconsistent definitions, and some sale prices are vastly inflated or are unrealistically small for various undisclosed reasons, such as a means of covert transfer of funds or assets. We refer to such transactions as *not genuine*.

A problem of the hedonic regression that has received little attention is that units have unequal influences on the regression fit. Units with extreme (outlying) configurations of the values of $\mathbf{x}$ are more influential than units with more typical values. However, such units, outliers in the covariate space, are least relevant to the comparisons we wish to make, especially when they occur infrequently or in only one of the years.

The repeat-sales method is based on properties that were sold two or more times in a designated period. The foundations of the method were laid down by Bailey, Muth and Nourse (1963), and some improvements and extensions are due to Case and Shiller (1987). For a pair of consecutive sales of a property, we define the difference of the log-prices as the outcome variable, and associate it with the years Y1 and Y2 of the first and second sale (sales $s = 1, 2$). We consider only pairs of sales in different years; see the end of this section for further discussion. We define a covariate for every year except the reference year. Its value is equal to $(-1)^s$ if the sale s was in the year, and it is zero for all other years. For example, a pair of transactions that took place in 2002 and 2006 in the setting of Table 1 is associated with the covariate vector $\mathbf{x} = (0, -1, 0, 0, 0, 1, 0)$, for the respective years 2001, ..., 2007. The corresponding model contains no intercept. It yields the estimates of the entries for each year, on the log-scale. The model may be adapted to cater for variance heterogeneity, assuming that the log-ratios of the sale prices of properties sold further apart in time are associated with greater variance.

The repeat-sales method relies on the close likeness, with all its attributes, of the versions of the property at the time points of the two sales. An important practical advantage of the method is that it requires no covariates, although some auxiliary information can be used as additional covariates in the regression. Also, the pairs of transactions can be split into categories according to an (unchanging) attribute and separate regressions fitted for the categories. The estimates are then weighted to reflect the distribution of the categories in the RHS, as in stratification.

Only a small fraction of the properties is sold in any one year, and the fraction of properties that are sold twice in a period of, say, five years is also very small. Thus, a sizeable database may

contain only a modest number of properties with multiple transactions. The problem is only slightly alleviated by using data that precede the start of the index. For example, data about transactions from 1995–1999 are useful for estimating the index that started in 2000. Pairs of transactions, say, from 1995 and 2002 and from 1998 and 2001 provide some (indirect) information about the change in the prices between 2001 and 2002. Index entries for past years can be updated as new transactions are recorded.

The frequency of sales of properties in the housing stock has a very uneven distribution, and so the set of properties involved in more than one transaction is bound to be a highly selective sample from the housing stock (Gatzlaff and Haurin, 1997). For example, properties owned and sought after by more mobile households, and single-member households in urban areas in particular, tend to change owners more frequently. The problem of selection bias applies also in the hedonic regression, but for repeated sales it is much more acute, because the selection of the properties is so extreme. In hedonic regression, the covariates reduce the impact of the selection bias, although it is difficult to establish to what extent. The regression in the repeat-sales method can be supplemented with covariates that aim to reduce the selection bias; stratification (weighting) can be regarded as such a device.

Properties sold twice in quick succession (say, within a year) often involve transactions in very different circumstances. The first buyer may have purchased the property with a specific intent to renovate (redevelop) it and re-sell it at a profit or the buyer may have been dissatisfied with the purchase because of some attributes of the property (or its location) and sold it without being able to negotiate from a position similar to the previous seller's. It is therefore prudent to exclude such transactions from the analysis. Similarly, a property owned for more than ten years is likely to have undergone such a lot of changes, not only by major alterations but also by wear and tear, maintenance, and changes in its location and environment, that the two transactions cannot be regarded as sales of the same asset. Therefore, the repeat-sales method should be based on a judiciously chosen 'window' of ownership period, such as 1 – 10 years. A problem rarely discussed in the literature on house prices is that two consecutive sales of a property share an agent — the purchaser in the first transaction is the seller in the second — and so the two transactions should be regarded as correlated, contrary to the standard assumptions.

3 Propensity score matching

We borrow the context of an epidemiological or educational study in which *treatments* T = A and T = B are applied to units $i = 1, \ldots, N$ of a population $\mathcal{P}$. Each unit i is associated with a $1 \times p$ vector of covariates $\mathbf{x}_i$, with no values missing, and a pair of *potential outcomes* $Y_i^{(\mathrm{A})}$ and $Y_i^{(\mathrm{B})}$, but for each member i only one (or none) of the outcomes $Y_i^{(\mathrm{A})}$ or $Y_i^{(\mathrm{B})}$ is realised. In an observational study, we have no control over the *assignment process*, that is, over which unit receives which treatment, if any.

In the setting of transations (sales) of residential properties in two years (A and B), the year of the sale is the treatment and the sale price is the outcome. The covariates comprise the attributes of the property (structure), location (surroundings and environs) and the circumstances of the transaction. Although a property may be sold in both years, we formally regard such a property as two distinct properties; after all, some of its attributes may have changed between the two sales. If the transaction takes place in year A, the sale price is $Y_i^{(\mathrm{A})}$, and if in year B, it is $Y_i^{(\mathrm{B})}$. The difference between $Y_i^{(\mathrm{B})}$ and $Y_i^{(\mathrm{A})}$, on the appropriate (log) scale, can be attributed solely to the year of the sale. Note that the date (day and month) of the sale may be a covariate; then the comparison of $Y_i^{(\mathrm{B})}$ and $Y_i^{(\mathrm{A})}$ would refer to sales exactly B – A years apart.

With the year of the transaction as the treatment, we can consider a fixed RHS, such as a list of (existing) properties in year 0, and compare their would-be sale prices in year A with their would-be sale prices in year B. We cannot preserve a property (together with its location and environment) in a vacuum, and so none of the contrasts $\Delta_i^{(\mathrm{BA})} = Y_i^{(\mathrm{B})} - Y_i^{(\mathrm{A})}$ can be observed. For every property i in RHS we seek a transaction in year A and one in B that are as similar to i as can be arranged. If the properties involved in the transactions in A are the RHS, then matching properties have to be found for them only in year B. By the rules that we set for such matching, we may fail to find a suitable match for some of the properties in RHS. This is not a deficiency of the method, but an indication that in the configurations of their values of $\mathbf{x}$ there are (substantial) systematic differences between RHS and the properties sold in year B. For example, the housing stock in year B may be so different from RHS that some types of properties that are common in RHS are scarce and sought after in B. This would suggest that the RHS is out of date, because the comparison of the years A and B is mediated by a reference stock of questionable relevance.

In the Rubin's causal model (Holland, 1986; Rubin, 2005), the treatment effect for unit i, $\Delta_i^{(\mathrm{BA})}$, is defined as the difference $Y_i^{(\mathrm{B})} - Y_i^{(\mathrm{A})}$. The difference can be replaced by the ratio or by the difference on another scale, such as the difference of the logarithms, $\log(Y_i^{(\mathrm{B})}) - \log(Y_i^{(\mathrm{A})})$. The average treatment effect Δ^{BA} is defined as the population average of these differences (contrasts). The arithmetic mean $(\Delta_1^{(\mathrm{BA})} + \cdots + \Delta_N^{(\mathrm{BA})})/N$ can be replaced by another form of averaging, such as the median or a robust mean, in which the weight assigned to units with outlying values of $\Delta_i^{(\mathrm{BA})}$ is smaller than $1/N$. The fundamental difficulty in evaluating or estimating $\Delta^{(\mathrm{BA})}$ is that only one term (or none) in any one of the differences $\Delta_i^{(\mathrm{BA})}$ is observed. For some definitions of RHS, neither $Y_i^{(\mathrm{A})}$ nor $Y_i^{(\mathrm{B})}$ is observed.

This problem could, in principle, be resolved by randomized assignment of the properties in RHS to the years. It would arrange that the two treatment groups (transactions within the years to be contrasted) are comparable, and can be compared without any regard for the values of $\mathbf{x}$ of the properties in the two groups. Without randomization, even if the transactions in years A and B involve sets of properties with similar (joint) distributions of the recorded attributes, the assumption that the joint distributions of all attributes are identical, is not warranted. In practice, the process

of allocating units to treatments (the assignment process) is influenced by the values of the outcomes even after taking the values of covariates $\mathbf{X}$ into account. We address this problem by matching (pairing), finding for each property in the RHS a pair of transactions, one in year A and one in year B, that involve properties similar to it. We focus on comparing years $\mathrm{A}=0$ and B. The ideal solution is a perfect match, when every pair has the same vector $\mathbf{x}$. In most settings very few, if any, such matches would be found. Various methods that define the nearest neighbour for each unit offer a more practical solution. Their drawback is that some units may be very popular as matches (suitable as pairs for several units) and some matched pairs are very far apart from one another. On the one hand, when there are many variables (components) in $\mathbf{x}$, we have more background information, and therefore a greater potential to assess accurately what is a good match. On the other hand, the matching is (analytically or computationally) more complex.

A universal solution was put forward by Rosenbaum and Rubin (1983) who showed that the propensity score is the optimal (univariate) score on which to base matching. The propensity score of unit i is defined as the conditional probability of assignment to a treatment (say, to B), given the values of the covariates $\mathbf{x}_i$. If the propensity scores were available, matching would be relatively simple. A unit in RHS has a match treated by A and another treated by B that have the same (or sufficiently similar) propensity scores. A strictly monotone transformation of the propensity scores is equally suitable for such matching.

The propensity scores can only be estimated; this complication is resolved by treating the problem as involving missing values. For details of *multiple imputation*, the method applied, see Rubin (2002), Schafer (1997) and Longford (2005, part I). The propensity scores are estimated by logistic regression of the assignment (indicator of one of the treatments, say B) on the covariates $\mathbf{x}$. It yields a vector of estimates $\hat{\boldsymbol{\beta}}$ of the regression parameter vector $\boldsymbol{\beta}$, and the fitted linear predictions $\mathbf{x}_i\hat{\boldsymbol{\beta}}$. Instead of $\hat{\boldsymbol{\beta}}$, we use several plausible parameter vectors $\tilde{\boldsymbol{\beta}}^{(m)}$, $m=1,\ldots,M$, generated as random draws from the estimated sampling distribution of $\hat{\boldsymbol{\beta}}$, the multivariate normal with expectation $\hat{\boldsymbol{\beta}}$ and (estimated) variance matrix $\hat{\boldsymbol{\Sigma}}=\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}})$. They are the basis of the plausible propensity scores $\mathbf{x}_i\tilde{\boldsymbol{\beta}}^{(m)}$. Separate sets of matched pairs of units are formed based on each set of plausible propensity scores. For each set, the unit-level and average treatment effects are evaluated straightforwardly; they are associated with no variation. Denote them by $\tilde{\Delta}_i^{(\mathrm{BA}),m}$ and $\tilde{\Delta}^{(\mathrm{BA}),m}$. The sole source of sampling variation is the uncertainty about the match, and its impact is captured by the differences among the plausible values $\tilde{\Delta}^{(\mathrm{BA}),m}$, $m=1,\ldots,M$.

In summary, the method we propose comprises the following steps:

1. fit a logistic regression to the assignments (A *vs.* B) in terms of the covariates $\mathbf{x}$;
2. generate M replicate sets of plausible propensity scores $\mathbf{x}_i\tilde{\boldsymbol{\beta}}^{(m,\mathrm{A})}$ and $\mathbf{x}_i\tilde{\boldsymbol{\beta}}^{(m,\mathrm{B})}$ for units i in RHS;

3. within each replication, match each unit in RHS with a unit in the transactions in year A and in year B;

4. evaluate the plausible mean effects (contrasts) $\tilde{\Delta}^{(\mathrm{BA}),m}$;

5. evaluate the multiple-imputation (MI) estimate and estimated sampling variance:

$$\begin{aligned} \tilde{\Delta}_{\mathrm{MI}}^{(\mathrm{BA})} &= \frac{1}{M}\sum_{m=1}^{M}\tilde{\Delta}^{(\mathrm{BA}),m} \\ s_{\mathrm{MI}}^2 &= \frac{M(M+1)}{M-1}\sum_{m=1}^{M}\left(\tilde{\Delta}^{(\mathrm{BA}),m}-\tilde{\Delta}_{\mathrm{MI}}^{(\mathrm{BA})}\right)^2 . \end{aligned} \tag{1}$$

In the application in Section 5, we simplify this by setting RHS to the properties involved in the transactions in year A.

4 Assumptions, settings and related details

An important assumption associated with the potential outcomes framework introduced in the previous section is the so-called SUTVA: stable unit-treatment value. It states that the units have an *autonomy* — the assignment of a unit to a treatment exercises no influence on (does not interfere with) the outcome of any other unit. In other words, a transaction i could not have any other outcome than $Y_i^{(\mathrm{A})}$, when assigned to A, and $Y_i^{(\mathrm{B})}$, when assigned to B. In this framework, the outcomes are *fixed*, and the sole source of randomness is the assignment.

We can devise a variety of replication schemes relevant to the analysed transactions. In one realistic scheme, the outcomes are highly dependent and correlated, because the conclusion of one transaction, facilitated by a real-estate agent, informs how the transactions due in the near future will be conducted; in particular, how sale prices in a neighbourhood or town in the near future are proposed and negotiated. In a replication, even if the same set of properties is involved in transactions, the circumstances of the transactions may differ, for example, if different potential buyers express interest and the buyers and sellers adopt different postures. The general economic climate is shared by all the transactions in a period, but it may be appropriate to keep it fixed in replications. Other (local) economic and environmental attributes may vary across replications, although to a much lesser extent than they do across areas. The dependence in such replications contravenes SUTVA, but is difficult to incorporate in any model, and neither the hedonic and repeat-sales methods nor the propensity score method address this issue in any way. However, the problem can be clearly appreciated only within the potential outcomes framework, which helps us address some of the ambiguities in the conceptual definition of a house price index.

Finding a match for each property in RHS is impractical in very large datasets. Inverse proportional weighting (IPW) is an alternative to matching. In IPW, the transactions are classified by their

propensity scores into bands delimited by cutpoints. The within-band means are then weighted to compensate for the discrepancies in the distributions of the scores for transactions assigned to B and to A. The cut-points can be set to suitable quantiles of the propensity scores or a fixed distance apart (at $r, r+a, r+2a, \ldots$). It is more practical to work with the propensity scores on the logit scale. Uncertainty about the propensity scores is represented by applying IPW to several (replicate) sets of plausible propensity scores.

Propensity score matching has some commonality with the repeat-sales method, in which a straightforward 'matching' is applied, using consecutive sales of the same property. With propensity score matching we overcome the difficulties of matching on several variables and alleviate the threat of selection bias, to the extent permitted by the available information. Just as the hedonic regression method relies on a suitable model, propensity score matching also requires a rich set of covariates, but these are for the description of the assignment process. Unlike in the hedonic regression, no concerns arise about the distributional assumptions for the outcomes (log-normality). The key modelling step does not involve the outcomes at all. Hedonic regression and repeat-sales methods assume a constant effect $\Delta_i^{(\mathrm{BA})} \equiv \Delta^{(\mathrm{BA})}$, and that is difficult to sustain. With propensity score matching, the variation of the effects is taken for granted.

The strength of hedonic regression and repeat-sales methods is variance reduction, and possible bias is their weakness. In contrast, bias reduction is the strong suit of propensity score matching, so it is suitable especially in settings in which the sampling variance is small. When there is a covariate with a dominant predictive power the strengths of the two approaches can be combined by matching not only on the propensity scores but also on the values of this covariate. This way of matching is called *fine* by Rosenbaum, Ross and Silber (2007). In the application in the next section, it yields estimators with greatly reduced (estimated) sampling variances. In the application, we generate $M = 20$ sets of plausible propensity scores based on the logistic regression with a set of variables $\mathbf{x}$. To make the comparison of the methods meaningful, we use the same set of variables in the hedonic regression and the propensity score analyses, except for the year of the sale (the 'treatment'). All the methods we apply have their robust versions. Hedonic regression and repeat-sales are adapted by replacing the ordinary least squares (OLS) with minimisation of the iteratively reweighted least squares $\sum_i e_i^2 w_i(e_i)$, in which the weights w_i depend on the residuals e_i. We report results for the Tukey bi-weight (kernel) function w_i defined as

$$w_i(e) = \left(1 - \frac{e^2}{c^2 \hat{\sigma}^2}\right)^2$$

if $|e| < c\hat{\sigma}$, and $w_i(e) = 0$ otherwise (Maronna, Martin and Yohai, 2006). The tuning constant c is set to 4.6. With other kernel functions we have obtained only slightly different results. Our main empirical argument in favour of basing a house price index on matched-pairs comparisons is that the results we obtain differ very little across the details of matching and whether or not we apply robust

versions of the of the complete-data methods. In contrast, robust and ordinary versions of the hedonic regression and repeat-sales analyses differ substantially.

5 Application

We compare the three methods for constructing a house price index on the list of all transactions of residential properties in years 2005 and 2006 in the six largest urban districts of New Zealand. Year 2005 is the reference and the transactions in the districts in 2005 define the respective RHSs. A random sample from the entire housing stock of each district in 2005 would be a better RHS, but we do not the relevant data. We give details of the analyses for Christchurch City District on the South Island and discuss the results for Wellington, the capital of the country, and the four districts that comprise the Auckland Metropolitan Area. The population of the district that covers Christchurch is 316 000. Auckland City is larger (population 368 000), and the other three districts that comprise Auckland are smaller (Manukau — 283 000, North Shore —185 000, and Waitakere — 169 000). The population of Wellington is 164 000. The district next in the order of population size is Hamilton City, with 115 000 residents.

Every residential property in New Zealand is assessed once in three years, or more frequently, on dates specific to each district (usually 1st of September of a year). The outcome of the assessment of a property is the *capital value*, the estimated value of the property on the day of the assessment, in New Zealand dollars (NZD). The land that belongs to the property is assessed similarly, with reference to the same date. There are 74 districts in New Zealand, 49 of them on the North Island, 24 on the South Island, and Chatham Islands, several hundred miles east of either island, form the remaining district. Eleven districts on the North Island and four on the South Island are designated as city districts. For more background, see Longford, McCarthy and Dowse (2008).

The database of all transactions in Christchurch City District in the period July 2004 – June 2007 comprises 38 483 records. The transactions have unique identifiers, and the database contains a small number of duplicates which are removed. We focus on comparing the house prices for the years A = 2005 (the reference) and B = 2006. We exclude from the analysis transactions with land area exceeding two hectares, floor area smaller than 30m^2 or greater than 500m^2, (assessed) land value greater than 50% of the capital value, capital value greater than three times the sale price or smaller than one-third of the sale price, and transactions in excess of NZD 2 000 000. These constraints are intended to exclude transactions that are not genuine, may be more appropriately classified as sales of land or of properties that serve a purpose different than (or additional to) primary residence. They reduce the data to 26 485 transactions, 13 497 (51.0%) of them from 2005. The distribution of these transactions over the months of the two years is given in the first two rows of Table 2. The remainder of the table lists the monthly minima, means and maxima of the numbers of transactions among the

Table 2: The numbers of transactions of residential properties in Christchurch City District, New Zealand, 2005–2006.

	Jan	Febr	March	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
2005	1146	1307	1360	1130	1161	973	1067	1184	1121	1134	1091	823
2006	952	1014	1295	1142	1216	1044	948	892	940	1248	1271	1026
2006 *matches*												
Minimum	1022	1106	1354	1195	1260	1072	867	841	886	1156	1114	875
Mean	1085	1164	1430	1242	1304	1114	956	892	926	1226	1218	940
Maximum	1167	1222	1485	1308	1348	1184	1003	935	1001	1275	1296	1027

20 sets of matches; details of the matching scheme are given below. The matches are assigned without replacement, so that in a replicate set of matched pairs a transaction in 2006 can be a match for at most one transaction in 2005. The table shows that the monthly numbers of transactions among the matches tend to be close to the distribution in 2005 on average, but they vary substantially across the sets of matches. (The month of the transaction is represented in the logistic regression for propensity scoring only as a continuous variable.)

Table 3 displays the means of the covariates used in the propensity score analysis for the two years and the minima, means and maxima of the means for the 20 sets of matched transactions from 2006. The means of the sets of matches (plausible transactions) differ from the means for 2005, but the averages over the 20 sets of matches are for most variables much closer to the 2005 means than they are to the 2006 means. The definitions of the variables are given in the Appendix (Table 8).

The means of the matched transactions from 2006 are closer to the corresponding means for 2005 uniformly (for every matched set) for *Land area*, *Month*, *Condition* and *Market*. We emphasize that differences among transactions in the two years are not substantial, but they are likely to be non-ignorable. Partial confounding of the covariates causes the means of the covariates to differ so much from one set of matched transactions to another. We explored several other models for the propensity scoring, including several interactions, but did not obtain any better agreement of the means of the covariates in the two years.

The fitted propensity scores are summarised in Figure 2. The histograms of the sets of plausible propensity scores have very similar shapes. There is a distinct set of about 700 transactions (2.7%) that have very large propensities (for being realised in 2006). All of these properties were sold 'non-market', that is, not through estate agents. There are 730 such properties; their smallest fitted propensity score is 0.47, and 721 of them have fitted scores greater than 0.70. In fact, 557 of the 'non-market' transactions (76%) took place in 2006. In 2006, these transactions tended to take place later in the year (mean month 7.67 *vs.* 6.47 for the other properties in 2006). They tended to have greater capital value in both years, by about 0.10 on the log-scale, that is, by about 10%, greater

Table 3: Distribution of the covariates for the transactions in 2005 and 2006, and in the sets of matched transactions in 2006; residential properties in Christchurch City District, New Zealand.

Year	*Capital value*	*Land value*	*Floor area*	*Land area*	*Chattels*	*Month*
2005	12.359	11.475	4.863	−4.728	8.556	6.242
2006	12.370	11.476	4.871	−4.830	8.609	6.528
2006 *matches*						
Minimum	12.350	11.460	4.856	−4.791	8.556	6.218
Mean	12.364	11.478	4.866	−4.731	8.610	6.294
Maximum	12.375	11.491	4.877	−4.655	8.638	6.390
	Cat-2	*Cat-3*	*Cat-4*	*Zone*	*Condition*	*Market*
2005	0.796	54.574	0.917	0.660	0.255	0.987
2006	0.795	55.094	0.914	0.666	0.302	0.957
2006 *matches*						
Minimum	0.787	54.283	0.911	0.647	0.248	0.987
Mean	0.795	54.875	0.917	0.661	0.258	0.987
Maximum	0.801	55.724	0.925	0.669	0.270	0.988

Notes: The *values* and *areas* are log-transformed. See Appendix (Table 8) for the definitions of the variables.

average land value and floor area, also by about 10% each, but smaller average land area, especially in 2005.

The purpose of propensity scoring is to match (balance) the two groups not only on the means, but on the entire (multivariate) distributions of the covariates. This is difficult to check, even though several covariates are binary. The variances or the variance matrices of the covariates can be compared, for the continuous variables in particular, as can the crosstabulations for the binary covariates. As far as one can judge from these summaries, propensity score matching accomplishes the intended goal; details are omitted.

A plausible match for a transaction in 2005 is selected by a random draw from the same band of propensity scores. That is, $K-1$ cutpoints $c_1 < c_2 < \ldots < c_{K-1}$ are set, defining K bands (intervals). If a band contains more transactions from 2005 than from 2006, then each transaction in 2006 is matched, at random and without replacement, with a transaction in year 2005. In the converse case, when in a band there are more transactions from 2006 than from 2005, some transactions from 2006 are without matches, and are not used in the analysis. This process of matching is replicated for each set of plausible propensity scores. We set the bands to the quantiles $1/K, 2/K, \ldots, (K-1)/K$ of the fitted plausible scores $\mathbf{x}_i\hat{\boldsymbol{\beta}}$, so that they are constant across the replicates. The numbers of matched pairs within the bands set with $K = 20$ in the $M = 20$ replicates are in the range 11 988 – 12 245, accounting for 90.5 – 92.5% of the transactions in 2005 and 2006. The allocation of (some) transactions to bands differs across replications because the plausible propensity scores differ from

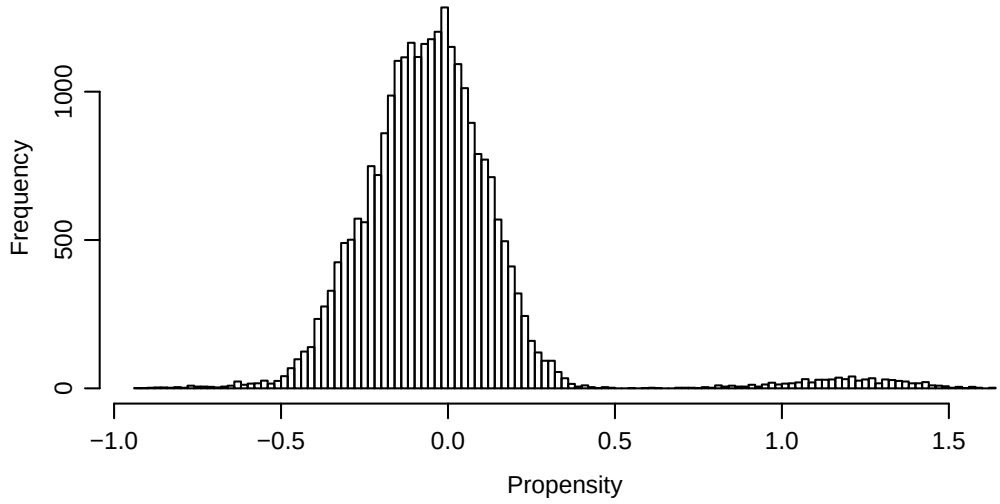

Figure 2: Histogram of the fitted propensity scores.

one set to another.

The mean of the $M = 20$ contrasts of the matched sets of transactions is 0.0877, and the associated standard error, estimated from (1), is 0.0055. If instead of the mean we apply the robust mean as the complete-data analysis, we obtain the estimate 0.0873, with nearly the same estimated standard error 0.0055. We apply IPW with the same $K = 20$ bands as for propensity score matching. The MI estimate of the mean difference on the log-scale is 0.0903 with estimated standard error 0.0049. This method is computationally undemanding and repeating it with different settings, to assess the sensitivity of the estimator, entails no difficulties. For 10, 15, 25, and 30 bands, we obtained estimates in the range 0.0900 – 0.0912, with estimated standard errors decreasing from 0.0051 (10 bands) to 0.0048 (30 bands). The method has a robust version, in which the robust mean is evaluated for every band and year. The within-cell totals of the robust weights are used instead of the number of observations. Its estimate for $K = 20$ bands is 0.0867 with estimated standard error 0.0046.

In most bands, the transactions have a wide range of sale prices and capital values. To take advantage of the capital value as a dominant predictor of the sale price, we apply fine (or *two-way*) matching on both propensity scores and capital values. We use ten bands for both, so that there are 100 matching categories. We have experimented with other ways of coarsening the values of the propensity score and the capital value, but obtained very similar results, until we defined so many categories that several of them contained very few (or even no) transactions in every replication. For smaller datasets, for the districts other than Auckland City, we applied 8×8 matching categories.

The MI estimate of the average difference between 2006 and 2005 is 0.0878, with standard error 0.0013, using $M = 20$ sets of plausible values. The robust version of this estimator is defined by

Table 4: Hedonic regression for City of Christchurch, fitted by ordinary least squares (top) and using Tukey bi-weight kernel (bottom); the respective estimated residual variances are 0.0223 and 0.00611.

OLS fit	*Intercept*	*Capital value*	*Land value*	*Floor area*	*Land area*	*Chattels*	*Month*
Estimate	2.10812	0.70764	0.09337	0.05514	0.00323	0.02548	0.00958
St. error	0.03826	0.00540	0.00354	0.00375	0.00044	0.00090	0.00027
	Cat2	*Cat3*	*Cat4*	*Zone*	*Condition*	*Market*	*Year*
Estimate	−0.00711	0.00002	−0.02605	−0.02514	0.01070	0.20045	*0.09188*
St. error	0.00333	0.00005	0.00353	0.00211	0.00240	0.00565	*0.00185*
Robust fit	*Intercept*	*Capital value*	*Land value*	*Floor area*	*Land area*	*Chattels*	*Month*
Estimate	3.28233	0.44472	0.05204	0.04363	0.00264	0.33758	0.00579
St. error	0.02360	0.00360	0.00205	0.00217	0.00025	0.00229	0.00015
	Cat2	*Cat3*	*Cat4*	*Zone*	*Condition*	*Market*	*Year*
Estimate	−0.00047	0.00013	−0.01923	−0.01050	0.00637	0.06067	*0.05584*
St. error	0.00189	0.00003	0.00206	0.00120	0.00136	0.00352	*0.00107*

replacing the mean with the robust mean in the complete-data algorithm. The robust estimate is 0.00866, with estimated standard error 0.0011. With IPW, we obtained similar estimates, 0.0894 (0.0011) and 0.0891 (0.0010). The standard errors are several times smaller than for univariate ('one-way') matching.

In hedonic regression analysis, the ordinary regression (OLS) and robust regression using the Tukey bi-weight kernel are fitted. The results are displayed in Table 4. The average effect of the year is estimated as the regression coefficient associated with the indicator of year 2006. With OLS, the estimate of the average effect is 0.0919 (standard error 0.0019) and with the robust regression it is 0.0558 (0.0011). Such a large difference in the results is disconcerting, more so that other regression models and kernels for robust estimation yield a wide range of (other) results. Details are omitted. The (estimated) standard errors in hedonic regression are of the same order of magnitude as obtained with fine (two-way) propensity score matching.

In the repeat-sales analysis, we include 3791 pairs of transactions. The pairs are in distinct years and at least half a year apart. No transaction is involved in more than one pair. The estimate of the average effect for 2006 *vs.* 2005 is 0.1188, with estimated standard error 0.0043. If we restrict the analysis to the 1252 properties sold once in 2005 and once in 2006, we obtain the estimate 0.1492, with estimated standard error 0.0042. Thus, the data from 2004 and 2007 appear not to be useful because the standard error is not reduced. The substantial change of the estimate indicates a serious problem with the method. The robust versions of the estimates are 0.1074 (0.0025) for the analysis with data

from all four years, and 0.1297 (0.0027) for the pairs of transactions in 2005 and 2006. The extent to which the price log-differences contain extreme values can be assessed informally by comparing the total of the weights in the concluding iteration of the model fit with the number of observations. For all the pairs of transactions, the total of the weights is 3217.1 (85% of $n = 3791$) and for the pairs of transactions in 2005 and 2006, 1049.2 (84% of $n = 1252$).

The results obtained by the hedonic regression and repeat-sales methods suggest that either the standard errors substantially under-represent the sampling variation or bias should be our principal concern in the analysis. In addition to the theory (Rosenbaum, 2002; Rubin, 2006), the stability of the results across a wide range of settings, including the application of robust methods, provides support for propensity score matching.

This conclusion is reinforced by the analyses of other large city districts in New Zealand, summarised in Table 5 for hedonic regression and repeat-sales methods, and in Table 6 for methods based on propensity scoring. In Table 5, the sample sizes involved are given in the columns with respective headings n (hedonic regression) and 'Pairs' (repeat sales). For completeness, the mean log-differences and their estimated standard errors are listed in the columns with headings 'Raw'. For repeat-sales, they are the mean log-differences in the sale prices for properties that have one sale in 2006 and one in 2005. They are based on about one-third of the pairs in the proper application of the method, yet the estimated standard errors are reduced much less than $\sqrt{3} \doteq 1.7$-fold. The large differences of the 'raw' and OLS estimates within and across the methods confirm the nonignorable nature of the analysed sales prices interpreted as (potential) outcomes subject to selection. The instability of the estimates for Auckland stands out.

The sets of eight estimates obtained by the methods based on propensity scoring are in a narrow range for every district. Two-way matching or IPW are associated with substantially smaller (estimated) standard errors, and IPW standard errors are smaller, except for Wellington. Note that the estimates in Table 6 are subject to random variation even conditionally on the data, unlike the estimates obtained by the established methods in Table 5. Rubin (2002) and Schafer (1997) imply that in most cases a small number of replicates in MI is sufficient to makes this uncertainty negligible and $M = 5$ suffices. That is the case for estimating the quantity of interest, but for reliable estimation of the standard errors more replications are usually required. This is not a serious constraint in most applications, including ours. The entire computation to produce the estimates displayed in Tables 5 and 6, with $M = 20$ replications, took about 200 seconds of CPU time, using `R` functions compiled specifically for this analysis. All the propensity score analyses are based on the same set replicates. The estimates in Tables 5 and 6 are in log-NZD. Subject to an approximation, they can be expressed as percentages by multiplying them by 100, since $\exp(x) \doteq 1 + x$ when $|\, x \,|$ is small. The corresponding standard errors are well approximated by the 100-multiples of the standard errors in the tables.

The stability of the estimators of the median difference can be checked similarly. Simply, the

Table 5: Estimates of the house price index entries for the metropolitan areas of New Zealand. Hedonic regression and repeat-sales methods. Estimated standard errors are given in parentheses underneath the corresponding estimates. The column 'Pairs' gives the number of pairs of transactions in the repeat-sales analysis, with the numbers of pairs that are from 2005 and 2006 underneath.

	Hedonic regression				*Repeat sales*			
District	*n*	*OLS*	*Robust*	*Raw*	*Pairs*	*OLS*	*Robust*	*Raw*
Christchurch	26 485	0.0919	0.0558	0.0978	3791	0.1188	0.1074	0.1492
		(0.0018)	(0.0011)	(0.0077)	1252	(0.0043)	(0.0025)	(0.0042)
Auckland	24 513	0.0083	–0.0019	0.0617	2920	0.1003	0.0831	0.1201
		(0.0027)	(0.0021)	(0.0094)	925	(0.0052)	(0.0031)	(0.0039)
Manukau	20 738	0.0998	0.0958	0.1060	3315	0.1220	0.1091	0.1511
		(0.0020)	(0.0014)	(0.0088)	1077	(0.0048)	(0.0027)	(0.0046)
North Shore	14 360	0.0735	0.0698	0.0920	2011	0.0906	0.0837	0.1118
		(0.0024)	(0.0016)	(0.0104)	679	(0.0054)	(0.0029)	(0.0042)
Waitakere	12 540	0.0727	0.0732	0.0742	1921	0.0967	0.0885	0.1232
		(0.0025)	(0.0017)	(0.0101)	597	(0.0055)	(0.0032)	(0.0050)
Wellington	10 372	0.0983	0.1061	0.1119	1329	0.1219	0.1197	0.1452
		(0.0030)	(0.0021)	(0.0126)	407	(0.0074)	(0.0043)	(0.0072)

Table 6: Estimates of the house price index entries for the metropolitan areas of New Zealand. Methods based on propensity scoring, matching (PS) and inverse proportional weighting (IPW). Estimated standard errors are given in parentheses underneath the corresponding estimates.

	Univariate matching/IPW				*Two-way matching/IPW*			
District	*PS*	*Robust PS*	*IPW*	*Robust IPW*	*PS*	*Robust PS*	*IPW*	*Robust IPW*
Christchurch	0.0877	0.0873	0.0903	0.0867	0.0878	0.0866	0.0894	0.0891
	(0.0055)	(0.0055)	(0.0049)	(0.0046)	(0.0013)	(0.0011)	(0.0011)	(0.0010)
Auckland	0.0058	0.0054	0.0069	0.0050	0.0047	0.0047	0.0040	0.0053
	(0.0073)	(0.0069)	(0.0068)	(0.0071)	(0.0023)	(0.0020)	(0.0011)	(0.0010)
Manukau	0.0974	0.0962	0.0973	0.0932	0.0969	0.0942	0.0983	0.0977
	(0.0035)	(0.0039)	(0.0030)	(0.0035)	(0.0012)	(0.0009)	(0.0007)	(0.0009)
North Shore	0.0748	0.0755	0.0733	0.0723	0.0726	0.0695	0.0729	0.0693
	(0.0051)	(0.0055)	(0.0051)	(0.0047)	(0.0021)	(0.0018)	(0.0014)	(0.0014)
Waitakere	0.0712	0.0720	0.0722	0.0735	0.0739	0.0754	0.0751	0.0775
	(0.0053)	(0.0048)	(0.0046)	(0.0046)	(0.0013)	(0.0015)	(0.0010)	(0.0010)
Wellington	0.0910	0.0906	0.0906	0.0902	0.0964	0.1003	0.0998	0.1074
	(0.0111)	(0.0118)	(0.0108)	(0.0107)	(0.0027)	(0.0029)	(0.0031)	(0.0031)

Table 7: MI estimates of the differences of the median house prices in 2005 and 2006 using propensity score matching.

	One-way matching		*Two-way matching*	
	Estimate	*Standard error*	*Estimate*	*Standard error*
Christchurch	0.0846	(0.0064)	0.0855	(0.0017)
Auckland	0.0076	(0.0077)	0.0056	(0.0031)
Manukau	0.0938	(0.0052)	0.0924	(0.0013)
North Shore	0.0712	(0.0076)	0.0681	(0.0022)
Waitakere	0.0741	(0.0070)	0.0760	(0.0022)
Wellington	0.0918	(0.0117)	0.0944	(0.0038)

complete-data method (the difference of the means) is replaced by the difference of the medians; the same sets of matched pairs can be used. Note that the difference of the annual medians differs from the median of the differences of the potential outcomes; the operations of difference and median do not commute. The MI estimates and estimated standard errors are listed in Table 7. They differ insubstantially from their counterparts in Table 6, because the distribution of the (plausible) log-differences is close to symmetry. The standard errors are inflated because, for distributions similar to the normal, the sample median is less efficient than the sample mean as an estimator of both the population mean and median. There is no natural adaptation of IPW for the median, because the sample median is not a linear function of the data. The counterparts in the established (regression) methods can be defined by minimising the respective totals of the absolute residuals, but they inherit the problems of the ordinary and robust regression.

6 Conclusion

We have formulated the problem of constructing a house price index in terms of estimating the effect of the date of transaction (sale) on the sale price. The causal interpretation of the problem is natural, desiring to distil the effect of time from the myriad of factors that influence the housing market. The implementation of the potential outcomes framework is not without flaws, but these are shared with the established methods. Foremost among them are the incompleteness of the set of covariates and our imperfect understanding of the processes of sale and purchase of properties. However, we dispense with any distributional assumptions related to the outcomes (sale prices). We use the scale of log-prices because we regard it as a more nature scale. Robust methods are relevant in our case because the database we analyse is contaminated by transactions that are not genuine and properties that may be used for purposes other than solely residence.

More detailed knowledge of the housing market can be translated into a better house price index by adjusting the matching procedure. For example, matching may be constrained within (sets of) geographical areas. We applied two-way matching, on the coarsened versions of the propensity score and the capital value, a *prima facie* good predictor of the sale price. This way of matching yields estimates with much smaller (estimated) standard errors. Using another covariate instead, such as the floor and land area, yielded much more modest gains, or was even slightly detrimental.

Our approach involves modelling the propensity scores and associated diagnostics to check whether a balance of the covariates has been attained. When indices for several areas (or subpopulations) and years are estimated, consistency in the specification of the models has some administrative advantage, even when it is contradicted by model selection criteria. Our experience suggests that it is preferable to err on the side of including more covariates in the propensity model. Propensity modelling does not involve the outcome variable, and so the usual concerns about about model selection bias are allayed. The very weak dependence of the estimates on the details of how the method is implemented are established by sensitivity analysis.

Acknowledgements

Research described in this manuscript was partly supported by the Grant SEJ2006–13537 from the Spanish Ministry of Science and Technology. The database analysed in this paper was obtained from Headway Systems Ltd, Christchurch, New Zealand.

References

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

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Colinearity.* Wiley and Sons, New York.

Case, K. E., and Shiller R. J. (1987). Price of single family homes since 1970: new indexes for four cities. *New England Economic Review*, Sept./Oct. 1987, 45–56.

Court, A. (1939). Hedonic price indexes with automotive examples. In American Statistical Association (Ed.) *The Dynamics of Automotive Demand*, pp. 99–117. General Motors Corporation, New York.

Gatzlaff, D. H., and Haurin, D. R. (1997). Sample selection bias and repeat-sales index estimates. *Journal of Real Estate Finance and Economics* **14**, 33–50.

Goodman, A. C. (1978). Hedonic prices, price indices, and housing markets. *Journal of Urban Economics* **5**, 471–484.

Harrison, D., and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* **5**, 81–102.

Kiel, K., and Zabel, J. (1997). Evaluating the usefulness of the American Housing Survey for creating house price indices. *Journal of Real Estimate Finance and Economics* **14**, 189–202.

Longford, N. T. (1993). *Random Coefficient Models.* Oxford University Press, Oxford, UK.

Longford, N. T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician.* Springer-Verlag, New York.

Longford, N. T., McCarthy, I., and Dowse, G. (2006). Patterns of house price inflation in New Zealand, 1996–2002. Chapter 17 in K. van Montfort, J. Oud and A. Satorra (Eds.) *Longitudinal Analysis in Behavioral and Related Sciences.* L. Erlbaum and Assoc., Mahwah, NJ.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics — Theory and Methods.* Wiley, New York.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* **82**, 34–55.

Rosenbaum, P. R. (2002). *Observational Studies.* 2nd ed. Springer-Verlag, New York.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* **102**, 75–83.

Rubin, D. B. (2002). *Multiple Imputation for Nonresponse in Surveys.* Wiley and Sons, New York.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modelling, decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association* **100**, 322–331.

Rubin, D. B. (2006) *Matched Sampling for Causal Effects.* Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

Appendix

Table 8 lists the covariates used in the hedonic regression and propensity score analyses.

Table 8: Definitions of the variables used in the hedonic regression and propensity score analysis. The outcome variable in the regression is the Sale price and the the treatment variable in the propensity score analysis is the Year. The values, or their ranges, listed in the right-most column apply for the Christchurch City District. Values that occur in the data frequently are printed in boldface.

Name	*Definition*	*(Range of) Values*
Year	The year in which the transaction took place	2004—2007
Cat2	Category D of the structure	*B*, *C*, **D**, **F**, *H*, *M*, *R*, *V*
Cat3	Age of the property (in years)	1910, 1920, . . . , 2000
Cat4	Category B of the structure	*1A*, *1B*, **A**, **B**, C, *XA*, *XB*, *XC*
Month	Month of the sale	1, 2, . . . , 12
Capital Value	The logarithm of the capital value (log-NZD) at the latest assessment	11.13 – 14.60
Land Value	The logarithm of the land value (log-NZD) at the latest assessment	9.40 – 14.05
Floor Area	The logarithm of the floor area (log-m^2) of the dwelling	3.43 – 6.20
Land Area	The logarithm of the land area† (section) in (log-hectares)	-9.21 – -0.19
Zone	Local zoning 9A	0 – 9 and A – Z
Condition	AA condition of the dwelling	AA – XX (**GG** and **AA** are the most frequent categories)
Chattels	The logarithm of the (additional) payment‡ for the contents of the property (log-NZD)	0 – 10.71
Market	Whether sold through an estate agency (1) or privately (0)	
Sale Price	The logarithm of the sale price (NZD)	10.70 – 14.50

Notes:
† — 0.0001 is added to the land area to avoid the expression log(0).
‡ — 1.0 is added to the chattels to avoid the expression log(0).