Using Revisions as a Measure of Price Index Quality in Repeat-Sales Models

Alex van de Minne\textsuperscript{1} · Marc Francke\textsuperscript{2,3} · David Geltner\textsuperscript{1} · Robert White\textsuperscript{4}

Published online: 12 January 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Repeat-sales indexes are the most widely used type of transaction based property price indexes. However, such indexes are particularly prone to revision. When a new period of transaction data becomes available and is used to update the repeat-sales model, all past index values can potentially be revised. These revisions are especially problematical for commercial real estate (as compared to housing), due to the relative scarcity of transaction data and the heterogeneity of the underlying properties. From a methodological perspective, the magnitude of the revisions is a particularly useful measure of the index quality, as it directly reflects both the precision of the index and its practical usefulness in economic and business applications. This paper focuses on index revisions in thin, commercial property markets, the type of market that is most challenging. We present multiple specifications of the repeat-sales model (both existing and new), seeking to reduce revisions. We are able to reduce overall index revisions by more than 50\%, compared to more traditional repeat-sales models.

Keywords Commercial real estate · Markov chained Monte Carlo · Property price indexes · State space models

Introduction

Many economic and financial time series are subject to periodic revision of their history (Clements and Galvão 2017). Revisions could occur when either the methodology or the underlying source data is updated. For example, the official report of the Gross Domestic Product (GDP) is revised because not all data sources needed for the
compilation process are available at the time of the initial report (Shrestha and Marini 2013; Clements and Galvão 2017). Therefore, statistical agencies publish a preliminary estimate first, which is subsequently revised. The S&P 500 is revised when the composition of the companies changes (Hegde and McDermott 2003).

In general, revisions improve the precision of the estimated indexes and should therefore not be ignored (Shiller 1993; Clapp and Giaccotto 1999). However, substantial and in particular systematic revisions of economic indicators, including real estate indexes, are not desirable. Economic models of mortgage prepayment and default, measures of (national) wealth, the total value of collateral behind a portfolio of mortgage loans, real estate derivatives, among many other applications, are all informed by real estate indexes subject to substantial (downward) revisions.

The most widely used methodology for empirical commercial property asset price indexing based on transaction prices is the repeat-sales model, based on price changes within properties that have sold at least twice (Bailey et al. 1963; Case and Shiller 1987, 1989; Goetzmann 1992, among others). The repeat-sales approach is often preferable, or necessary, because the other major approach – a hedonic price model – suffers from the heterogeneity of commercial properties combined with the lack of good and sufficient hedonic data on the properties: Commercial properties are not subject to the kinds of official data gathering that is applied to housing by the US Census and other agencies. However, a repeat-sales index is particularly susceptible to revisions. When new transaction data becomes available, new repeat-sale ‘pairs’ are formed in the estimation sample as we move forward in time. Not only do we receive the new prices at the resale date (second sales in the pairs), but the prices at the buy date also enters the data (first sales in the pairs). The entire index could thus be affected by adding new pairs. We distinguish three types of index revision: Systematic, random, and data generating process (DGP) driven revision.

Past studies on repeat-sales price indexes provide evidence for systematic downward revisions (Abraham and Schauman 1991; Shiller 1993; Clapp and Giaccotto 1999; Clapham et al. 2006; Bourassa et al. 2013). These studies have focused on residential property price indexes based on relatively large samples of transactions. The systematic downward revisions are due to sample selection bias. Properties with longer (shorter) holding periods are by construction underrepresented (overrepresented) in repeat-sales samples, in specific when the sample period is short. Moreover, homeowners and investors tend to sell ‘winners’ – properties experiencing greater than average price appreciation – more readily than ‘losers’, i.e. winners tend to have shorter holding periods, see Genesove and Mayer (2001) and Bokhari and Geltner (2011). Francke (2017) has recently proposed a simple methodological solution to the problem of systematic revisions. Other authors also tried to address systematic revisions. For example, Clapp and Giaccotto (1999) omit properties with a short holding period, and Bourassa et al. (2013) estimate a robust repeat-sales model. These systematic revisions are not the focus of this paper.

Random revision is fundamentally caused by the dispersion in transaction prices around the unobservable market value, sometimes referred to as noise. Consummated transactions in private property markets are essentially trades between two parties for unique, whole assets. Neither party can know for sure what the property market value is. Indeed, evidence suggests that transaction noise reflects a standard deviation
of at least ten to fifteen percent of the property market value (Geltner et al. 2014, p. 265). Granular indexes of commercial property – indexes of individual markets or segments, such as a particular type of property in a particular metropolitan location – are particularly prone to random revisions, because only a handful of repeat-sales transaction observations per period is available, resulting in noisy estimates of the price index. Yet granular indexes are particularly useful for many important practical applications. Of course, the precision of the index estimate increases when more transactions becomes available over time, see Shiller (1993) and Clapp and Giaccotto (1999). In other words, the index is revised over time. Random price dispersion and resulting index revisions are the underlying motivation for the present paper.

Though random price dispersion is the fundamental ‘culprit’, the specific nature of random revisions in repeat-sale indexes is determined by what may be viewed as a third type of index revision. We refer to this as data generating process (DGP) driven revision. Mechanically or mathematically, index revisions are quantitatively determined by the specification of the data generating process, and in particular the specification of the price index process in the index estimation methodology. In a standard repeat-sales model the log price index is specified by time fixed effects (dummy variables), so a priori no structure on the price index is assumed. However, in thin markets these time fixed effects need to be replaced by a more parsimonious structure in order to reduce the impact of noise on the index. In this paper we consider several stochastic specifications for the price index: The DGP could be specified as a random walk for the log index level, a first-order autoregressive model for the log index return, and so on. The specification of the DGP determines the filtering of the individual repeat-sales returns into a trend (index) and a noise component (error term). Different specifications lead in practice to differences in estimated price indexes, and to differences in the size of revisions, although the model fit for the various models – as measured by likelihood based information criteria – could be quite similar. The index revision driven by the specification of the data generating process is the main focus of this paper.

Acknowledging the importance of revisions for property price indexes in practical applications, we argue that the size of revisions should be used as an important additional selection criterion for price index models. Guo et al. (2014) have pointed out that traditional measures of price index quality, such as standard errors of the estimated index levels, goodness of fit measures, and signal-to-noise ratios are not always very appropriate for measuring real estate price index quality. The magnitude of revisions directly reflects the precision of the empirical price index estimate, and thus can provide a direct and particularly useful and practical metric of the index quality. Moreover, the revision statistics allow us to compare ‘fit’ between different markets. This is difficult to achieve by using the different information criteria. And finally, not least, the magnitude of revisions directly impacts the utility of price indexes for important purposes, such as reporting economic statistics, and supporting derivatives trading. For many index users, parameters such as volatility or WAIC metrics may be rather abstract, but having to revise their reports and conclusions, or contractual payment obligations, is a very real ‘headache’.

The goal of the paper is therefore to measure the quality of price indexes for different specifications of the repeat-sales model using several revision statistics. We
also compare our findings with more traditional model fit measures, like the standard error of regression (‘noise’) and the Watanabe Akaike Information Criterion.

At first, we only look at existing real estate price index models. More specifically, we estimate the ‘standard’ repeat-sales model (Bailey et al. 1963) and a robust repeat-sales model (Bourassa et al. 2013). We also use a random walk plus drift (Goetzmann 1992) and a local linear trend (Francke 2010) specification for the log price index level. On average, the revisions go down 50% for the structural time series models, compared to the ‘standard’ repeat-sales model. In most cases, the classical goodness of fit measures also indicate a better fit for the structural time series models over the ‘standard’ repeat-sales models.

Secondly, we also introduce alternative specifications for the stochastic trend component. We specify a first order autoregressive model for the periodic log index returns. Both theoretical reasoning and empirical evidence supports the proposition that in private real estate markets (log) returns are positively serially correlated (Case and Shiller 1989; Barkham and Geltner 1995; Geltner and de Neufville 2017). Another innovation is that we introduce the periodic returns from an aggregate index as an explanatory variable in a granular index model. The aggregate index represents a broader scale – a larger geography, or a broader range of property types – and hence larger sample than the (target) granular index being estimated. Especially introducing these aggregate returns to inform the periodic log returns of the granular index reduces the revisions even further. Another interesting finding, is that the classical goodness of fit measures do not report an improvement of our newly proposed models, over the existing (structural time series) models, even though the revision statistics improve considerably.

We apply the various repeat-sales index models on two different data sets: an ‘outer Los Angeles’ sample consisting of office property transactions in the combined Inland Empire and Orange County submarkets of the Los Angeles metro area; and an economically different but similarly small sample market consisting of apartment property transactions in Oakland CA, a submarket within the San Francisco metro area. We construct quarterly commercial property price indexes over the period 1997Q1 to 2016Q1. On average the number of observations per quarter is less than nine in both samples. Consequently, our analysis provides a good illustration of the efficacy of our index methodology for small-sample, granular indexes, while allowing us to demonstrate the index quality based on revision statistics.

Even though the focus of our paper is on repeat-sales models for commercial real estate, our findings are not limited to this specific model and application. They are also relevant for other economic and financial time series and for other models including the hedonic price model (for real estate applications, see Schwann 1998; Francke and De Vos 2000).

The setup of this paper is as follows. In “Empirical Strategy and Revisions Statistics” we discuss the basic repeat-sales model and how revision statistics can be used as a measure of index quality. Section “Data and Descriptive Statistics” describes the repeat-sales data. Section “Revisions in Existing Models” gives the specifications, estimation results and revision statistics for some existing models, including the basic and a robust repeat-sales model, and a local linear trend repeat-sales model.
Section “Revisions in Newly Proposed Models” provides specifications, estimation results, and revision statistics for newly proposed models. As a robustness we apply our models to multiple other markets as well in “Robustness”. Finally, “Conclusion” concludes.

**Empirical Strategy and Revisions Statistics**

**Model Specification**

The repeat-sales model is a widely popular method to produce property price indexes. For example, both the House Price Index of the Federal Housing Finance Agency and the Moody’s/RCA price index for commercial real estate are based on the repeat-sales methodology.\(^1\)

In the academic literature the major alternative method to produce transaction based indexes is the hedonic price model. However, a big advantage of the repeat-sales model is that it does not require data on property characteristics. This is especially valuable in the commercial real estate market as the properties are very heterogeneous and detailed property characteristics are mostly absent. Another advantage of the repeat-sales method, again especially for commercial property, is that the repeat-sale method transparently and directly tracks the price change experiences of investors (who must sell the same properties that they have bought). And there are fewer issues in repeat-sales models regarding the specification of the correct functional form of the model (Bailey et al. 1963; Deng and Quigley 2008). The repeat-sales model is therefore much less vulnerable to specification errors and omitted variable bias compared to the hedonic price model.

By using only properties that sell more than once, the repeat-sales model effectively includes property-level fixed effects. (Formally, this is immediately clear by specifying the model in levels rather than in differences.) This clarifies how the model minimizes the omitted variables problem that plagues hedonic indexes of commercial property. As pointed out in Guo et al. (2014) and Geltner et al. (2017), for the purpose of price indexing, the repeat-sale model can be derived and justified directly without recourse to the hedonic price model. As noted, the repeat-sales index directly reflects investors’ experiences. The downside of the model is that single sales – properties with only one transaction price – do not contribute to the estimation of the price index, reducing the effective sample size and possibly contributing to sample selection bias.

The repeat-sales model can be expressed as

\[
\ln P_{it} = \mu_t + \beta_i + \epsilon_{it},
\]

\(^1\)Both use a slightly different setup than the classical (Bailey et al. 1963) method. The HPI uses a WLS approach to estimate the repeat-sales model as proposed by Case and Shiller (1987). The Moody’s/RCA CPPI uses frequency conversion as proposed by Bokhari and Geltner (2012). Since September 2017 the RCA CPPIs employ a methodology based on the current paper.
where $P_{it}$ is the transaction price at time $t$ of property $i$. The pair fixed effect is denoted $\beta$, and $\epsilon$ is the error term which is assumed to be normally distributed with zero mean and variance $\sigma^2$. $\mu_t$ is the log price index at time $t$, and it is traditionally specified by time fixed effects. $\mu_0$ is initialized at 0.

The ‘price’ reflected in the dependent variable is in fact a ‘value’ that from a hedonic perspective represents the sum-product of property characteristic shadow prices times the characteristics quantities. The model is thus a representation of the longitudinal market component of value ($\mu_t$) plus the cross-sectional property-specific component of value ($\beta$). The index reflects only the former. This is comparable to stock market price indexes, where the corporations whose stock is traded certainly do not remain static over time, that is, stock market ‘price’ indexes are actually ‘value’ indexes in the sense of reflecting the combined effect of pure price change and pure quantity change.

Model 1 can be estimated using the least squares dummy variable regression method (for the pair fixed effects and time dummies). Computation becomes even more tractable when it is specified in ‘differences’, canceling out the property fixed effect $\beta$. This is the traditional way the repeat-sales model is specified (fundamentally equivalent to Eq. 1):

$$r_{ist} = \mu_t - \mu_s + \epsilon_{ist} = \delta_{t+1} + \cdots + \delta_t + \epsilon_{ist},$$

(2)

where $r_{ist} = \ln P_{it} - \ln P_{is}$ is the log price return realized between sales, with $s$ being the time of buy, and $t$ the time of sale (Bailey et al. 1963; Case and Shiller 1987, 1989, among others). The index return at period $t$ is denoted by $\delta_t = \mu_t - \mu_{t-1}$. The error term $\epsilon_{ist}$ is defined as $\epsilon_{it} - \epsilon_{is}$ with $\epsilon_{ist} \sim N(0, 2\sigma^2)$, for $t = 1, \ldots, T, s = 0, \ldots, t - 1$. We use Eq. 2 in the remainder of this paper.

### Revision Statistics

Repeat-sales price index estimates are subject to revisions, for the reason noted previously, as new ‘pairs’ are formed as we move forward in time, bringing new information relevant to the entire span of the index history.

Let $\hat{\mu}_{T|\tau}$ denote the estimated log price index at time $T$ based on all repeat sales up to time $\tau \geq T$. Suppose we have 2 estimates of the log price index level at time $T$, based on transaction prices up to $\tau_0$ and $\tau_1$, denoted by $\hat{\mu}_{T|\tau_0}$ and $\hat{\mu}_{T|\tau_1}$ respectively, where $\tau_1 > \tau_0$ and $\tau_0 \geq T$. We define revision of the log index level at time $T$ based on all repeat sales up to $\tau_0$ compared to all repeat sales up to $\tau_1$ as $\hat{\mu}_{T|\tau_1} - \hat{\mu}_{T|\tau_0}$. Likewise we define revision of the log index return as $\hat{\delta}_{T|\tau_1} - \hat{\delta}_{T|\tau_0}$. An algebraic example of revisions in a standard repeat-sales model is given in Appendix A.1.

Examples of revision paths for a specific quarter using the ‘standard’ (Bailey et al. 1963) repeat-sales index are visualized in Fig. 1. We use actual data, described in “Data and Descriptive Statistics”, on offices in the Los Angeles ‘rest’ area. These

---

2Note that $\beta_i$ is specified as a pair fixed effect in this paper, as this the most widely accepted form in both academic literature and in industry. However, it could also be specified as a property fixed effect. So if a property is sold 3 times, we ‘break it up’ into 2 pairs, instead of seeing it as 1 property. See Francke (2010) for the (small) difference in specification.
figures demonstrate the systematic remeasurement of parameter estimates from their ‘initial’ to ‘current’ values. For example, Fig. 1a reports that the ‘initial’ estimate of the price level in 2010Q3 was approximately 76 (with 2005Q1 being 100, see “Data and Descriptive Statistics”). Subsequently, as sales occurred in the following period, these data and their associated paired-sales were added to the data set, and prices were revised. At the end of the sample (which is 2016Q3 in our case) that same 2010Q3 index level sits at 68, which is over 10% less than the original estimate. (Also note that you can clearly see the aforementioned systematic downward revisions as well in Fig. 1). Figure 1b gives the revision path for the returns. Here we observe that the ‘initial’ estimate of the quarterly index return was 16%. At the end of the sample, the estimated 2010Q3 return is closer to 6%, which is a difference of over 60%.
In total we propose four different sets of revisions statistics, as indicators of the index quality. The first set concerns statistics on the revisions in the index levels, \( \hat{\mu}_{T|\tau_1} - \hat{\mu}_{T|\tau_0} \), over all possible combinations of \( T, \tau_0, \) and \( \tau_1 \). The second set consists of the same set of statistics on the revisions in the returns. The statistics of interest are: The absolute mean, the standard deviation, and the biggest (absolute) revision, and the amount of times the revision is larger than a predefined threshold value. We also employ a two-tailed Wilcoxon Signed-Rank test over the revised indexes.

The final two sets are based on the index levels and the returns, respectively, but exclusively considering the revision in the last period, that is, the period just prior to the current update of the index. In other words, for the index update at time \( T \), we consider the revision in the period \( T-1 \). We are particularly interested in the revision in the final (most recent) period of any given index. For an index that is published regularly, providing ‘news’ about the market, revisions in the final period may be the most problematical in practice for decision makers and analysts in the real world who might be making real-time conclusions based on what had recently been the ‘news’ of the previous period’s update. As may be expected and as we shall see empirically in our applications, revisions metrics based on index levels essentially tell the same story as those based on returns.

We also track some other statistics. A simple statistic for model fit is provided by the variance of the error term \( e_{ist} \) in Eq. 2. A more advanced goodness of fit, and model selection measure is the likelihood based Watanabe Akaike Information Criterion (or WAIC, see also Watanabe 2010; Vehtari et al. 2016).

Moreover, we compute the first-order autocorrelation and volatility of the log index returns as indications of index quality. High index return volatility and/or negative autocorrelation might be indications of noisy estimates. It is well known that real estate returns are characterized by positive autocorrelation. This ‘inertia’ is inherent to the price formation process in real estate and does not necessarily imply unexploited feasible arbitrage opportunities. More specifically: (1) Unique, whole assets are traded about which participants have incomplete information on the effect of news on the value of any one specific asset; (2) some period of costly search must be incurred by both buyers and sellers, due to the heterogeneity of real estate; (3) trades are decentralized, i.e. market prices are the outcome of pairwise negotiations; and (4) transaction costs are high relative to asset values (Case and Shiller 1989; Quan and Quigley 1991).

Data and Descriptive Statistics

Real Capital Analytics Inc. (RCA) provided us with repeat-sales data for commercial real estate in two US metro areas, starting in 1997 and ending in the first quarter of 2016. This gives us the opportunity to analyze the crisis and subsequent recovery. More specifically, RCA provided us transaction data on offices, retail, industrial and apartments for the Los Angeles metro area and San Francisco metro area.\(^3\) Our analysis is based on subsets of those markets.

\(^3\)The metro areas are defined by Real Capital Analytics.
The first subset consists of offices in the combined region of Inland Empire and Orange County in the Los Angeles metro area (LAI). In total we observe approximately 330 repeat-sale pairs for this market, over the entire 1997-2016 period. The second subset consists of apartments in Oakland in the San Francisco metro area (SFO). As in the previous market we only observe little over 330 pairs for the entire period. This is equivalent to approximately 4.5 pairs per quarter, the index frequency we use in our analysis. Even though the counts are roughly similar, the markets have been behaving very differently. Table 1 gives some descriptive statistics for both markets. The table provides statistics on the price levels, the log price returns (log price difference between buy and sell), and the quarterly log returns (i.e. log price return divided by the holding period in quarters). On average offices in Inland Empire and Orange County sold at approximately 50% higher prices than apartments in Oakland. In contrast, Oakland apartments had higher periodic returns.

The number of transactions (note that every pair consists of two transactions) per quarter is given in Fig. 2a and b for offices and apartments in Inland Empire / Orange County and Oakland, respectively. As expected we observe a big drop in the number of transactions just after the crisis, whereas the peak is just prior to the crisis.

Figure 2c and d give the number of pairs in our data as of any given date of index history after 2005. Note that transactions are ‘back filled’, meaning that with every new transaction, a pair is formed with a sale in the past. Thus, more transactions now, will also result in more transactions in the past. Second, to add a layer of realism

| Table 1 | Price and log return statistics, 1997–2016 |
|---------|------------------------------------------|
|         | LAI                         | SFO                         |
| **Price \times 1,000** |                             |                             |
| 1st Qu. | $4,600                      | $3,200                      |
| Median  | $9,038                      | $5,080                      |
| Mean    | $19,860                     | $12,460                     |
| 3rd Qu. | $21,560                     | $12,000                     |
| **Log Returns** |                             |                             |
| 1st Qu. | −0.212                      | 0.115                       |
| Median  | 0.207                       | 0.264                       |
| Mean    | 0.083                       | 0.254                       |
| 3rd Qu. | 0.391                       | 0.426                       |
| **By Holding Period** |                             |                             |
| 1st Qu. | −0.007                      | 0.005                       |
| Median  | 0.010                       | 0.013                       |
| Mean    | 0.012                       | 0.016                       |
| 3rd Qu. | 0.032                       | 0.024                       |
| **Other** |                             |                             |
| # of Pairs | 338                        | 339                        |
| # of Transactions | 676                      | 678                        |
| Property type | Office                   | Apartment                   |

LAI = Inland Empire / Orange County in the Los Angeles metro area.
SFO = Oakland in the San Francisco metro area.
The returns by holding period are expressed in quarters.

# Springer
to our analysis, we use the date in which observations were added to the database, instead of the transaction date (in case they are different). It takes time to find and add transactions to the database. Also, new data providers can be found which results in many ‘new’ past transactions coming into the database at once. So it might be possible that we leave out a 2010Q1 transaction even though our revision ‘window’ is past that date, simply because this transaction was only added to the database at a later point in time (in which we obviously do add this observation to the sample). The number of pairs in Fig. 2c and d takes this into account.

In our analysis we go back in time and estimate the index as if we were in the third quarter of 2007, indicated by the red lines in Fig. 2c and d, using only the transaction pairs available as of that point in time. At that time, there were only slightly more than 70 pairs for each market. We then update the indexes one quarter at a time, adding the newly available data each quarter and re-estimate the models moving forward in time, and we record how the indexes change using the metrics discussed in “Empirical Strategy and Revisions Statistics”. We record all such revisions through

Fig. 2 Data description offices and apartments in Inland Empire / Orange County (LAI) and Oakland (SFO). The red line represents the starting point of the monitoring of revisions.
the first quarter of 2016. This results in 34 (revised) indexes per model per region. In total we estimate in this paper close to a 1,000 indexes.4

Revisions in Existing Models

In this Section we compare the severity of revisions in existing repeat-sales models for our two markets: Offices in LAI and apartments in SFO. First, in “Existing Models” we specify and describe the models. The actual results of our revision analysis is subsequently given in “Results”.

Existing Models

We analyze the index revisions of five different existing repeat-sales models. The first method is the ‘standard’ (Bailey et al. 1963) repeat-sales model, that was discussed in “Empirical Strategy and Revisions Statistics”, see Eq. 2. This model is denoted RS.

The second model is a robust repeat-sales model. Different specifications are discussed in Bourassa et al. (2013). It is expected that the standard RS model is sensitive to outliers. As such, a robust repeat-sales model might result in less ‘noisy’ indexes and consequently less revisions. More specifically, we estimate median returns, using a quantile regression, see also McMillen and Thorsnes (2006). In order to get median returns we assume that the error term in Eq. 2 has a double exponential distribution with scale 2 (Yu and Moyeed 2001). We denote this model rRS.

Goetzmann (1992) introduced into the real estate literature what is perhaps the major approach to date for addressing small-sample problems in property price indexes, namely, the use of Bayesian inference within repeat-sales regression models. Goetzmann (1992) specifies the prior distribution for the periodic returns in Eq. 2 to be normally distributed, implying a random walk with drift for the log price index. Thus, the trend $\mu$ in Eq. 2 is specified as

$$\mu_t \sim N(\mu_{t-1} + \kappa, \sigma^2_\eta).$$

We denote this model - a random walk with drift - RWd.

Francke (2010) generalized the ‘Goetzmann’ model by assuming that the trend component follows a local linear trend, given by

$$\mu_t \sim N(\mu_{t-1} + \kappa_{t-1}, \sigma^2_\eta),$$
$$\kappa_t \sim N(\kappa_{t-1}, \sigma^2_\zeta).$$

Note that the local linear trend is very flexible and includes different specifications, like random walk ($\sigma^2_\zeta = 0$ and $\kappa_1 = 0$), random walk with drift ($\sigma^2_\zeta = 0$) and smoothed trend ($\sigma^2_\eta = 0$). It is denoted LLT.

The final model in this Section is the local linear trend model where we assume $t$-distributed errors in the measurement Eq. 2: $\epsilon_{ist} \sim T(0, 2\sigma^2_\varepsilon, \nu)$, where $\nu$ is the

---

4In fact, we also ran some additional indexes as a robustness check in “Robustness”. As such, the total number of indexes we estimate is closer to 5,000.
degrees of freedom to be estimated from the data. The $t$-distribution has fatter tails compared to the normal distribution, and can better deal with outliers. This model is denoted LLT, and is closely related to the research done by (Francke and van de Minne 2017). An overview of the models is given in Table 2.

The random walk (with drift) and the local linear trend repeat-sales model are examples of structural time series models in which the trend, error terms, plus other relevant components, are modeled explicitly. An illustration of revisions for a specific structural time series model, the random walk repeat sales model, is provided in Appendix A.2 and A.3. The appendices provide an algebraic expression and numerical examples of index revisions for different values of the signal-to-noise ratio $q = \sigma^2_\eta/(2\sigma^2_\epsilon)$.

For all specified models we apply full Bayesian inference to derive the posterior marginal distributions for our parameters of interest, the index levels. We specify largely non-informative priors for the hyperparameters, the variances and the degrees of freedom. More specifically, all the variance parameters have a uniform distribution with lower bound 0 and upper bound 1. The prior distribution for the degrees of freedom parameter $\nu$ is Exp(0.3). In order to derive the marginal distributions of the parameters of interest we use Markov Chain Monte Carlo (MCMC) techniques. As the number of (hyper) parameters is high and we have to repeat the inference many times (over different subsamples for several markets), we need a sampler that is fast and converges rapidly under these circumstances. We use the No-U-Turn-Sampler (NUTS) developed by (Hoffman and Gelman 2014). NUTS is a generalization of the Hamiltonian Monte Carlo algorithm. We also estimate the RS and rRS models using NUTS. As such, all diagnostics are directly comparable. Appendix B provides more technical details on the applied estimation procedure.

**Results**

The indexes resulting from the models specified in the previous Section are given in Fig. 3. The indexes are shown starting from 2005Q1, although the actual data go back further, see “Data and Descriptive Statistics”. The Tables 3 and 4 provide revision statistics, goodness of fit measures, and some MCMC diagnostics for offices in LAI and apartments in SFO respectively.

Glancing at the indexes in Fig. 3 it is clear that replacing the fixed time effects in the repeat-sale model by a structural time series model greatly reduces the index volatility (Goetzmann 1992; Francke 2010; Francke and van de Minne 2017). This is

| Model name                  | Eq.s. | t-dist. | Literature                  |
|----------------------------|-------|---------|------------------------------|
| Standard                   | (RS)  | 2       | No                           |
| Robust                     | (rRS) | 2       | No                           |
| Local Linear Trend         | (LLT) | 2 / 4 / 5 | No                       |
| Local Linear Trend         | (LLT) | 2 / 4 / 5 | Yes                       |
|                            |       |         |                              |
consistent with findings in previous literature and with the simulations in Appendix A.3. The standard deviation of the returns more than halves for both markets, see Tables 4, 5 and 6. More specifically, in LAI (SFO) the volatility of the returns are 0.129 (0.129) and 0.076 (0.062) for the RS and rRS models respectively. The index return volatility subsequently drops to 0.046 (0.035), 0.039 (0.031) and 0.041 (0.032) for the RWd, LLT and LLTt models respectively.

Also note that in both markets the RS and rRS models either result in negative, or close to zero autocorrelation (ACF(1)) in index returns. This suggests that the indexes are less driven by noise when introducing structural time series specifications in a repeat-sales framework. The volatility of the returns (first-order autocorrelation) is lowest (highest) for the LLT method in both markets.

In both markets, the LLTt model has the highest model fit measured by WAIC. Although in LAI the difference is less than 5 points, meaning it is not seen as a considerable difference. The noise is also least for the LLTt model in both LAI and SFO, driven by the $t$-distribution. Indeed, and in line with Francke and van de Minne (2017), the parameter estimates given in Tables 9 and 10 clearly show that the errors in the measurement equation follow a $t$-distribution: The posterior mean of $\nu$ is 8.6 and 5.6 for LAI and SFO respectively, clearly rejecting the normality assumption.

It is clear that the structural time series models (RWd, LLT and LLTt) are ‘better’ than the standard repeat-sales model measured even by the ‘traditional’ metrics. The volatility (standard deviation of the index returns), and first-order autocorrelation (ACF(1)) are more in line with expectation, and the WAIC and standard error of the residuals improve. Again, the only exception being LAI, where the difference in WAIC is less than 5 points between the LLT and LLTt models.

Next we will focus on the revision statistics. As noted in “Empirical Strategy and Revisions Statistics”, we break down the revisions in four groups:

1. revisions in the index *levels* for all periods starting from 2007Q3;
2. revisions in the index *returns* for all periods starting from 2007Q3;
3. revisions in the index *levels* for the final period only;
4. revisions in the index *returns* for the final period only.
Table 3 Revision Statistics for Offices in LAI, existing models

|                      | RS      | rRS     | RWd     | LLT     | LLTt    |
|----------------------|---------|---------|---------|---------|---------|
| **Traditional metrics** |         |         |         |         |         |
| WAIC                 | 188.092 |         | 137.214 | 129.089 | 125.604 |
| Noise (sd)           | 0.284   | 0.280   | 0.046   | 0.039   | 0.041   |
| Volatility           | 0.129   | 0.076   | 0.046   | 0.039   | 0.041   |
| ACF(1)               | −0.370  | 0.038   | 0.596   | 0.929   | 0.920   |
| **Revision, All, Levels** |         |         |         |         |         |
| Wilcoxon             | 0.796   | 0.348   | 0.311   | 0.316   | 0.312   |
| Std.dev              | 0.025   | 0.022   | 0.016   | 0.016   | 0.017   |
| |Mean| | 0.014   | 0.014   | 0.009   | 0.009   | 0.009   |
| Max                  | 0.226   | 0.222   | 0.146   | 0.172   | 0.156   |
| Fails at 5%          | 0.047   | 0.030   | 0.017   | 0.020   | 0.024   |
| **Revision, All, Returns** |         |         |         |         |         |
| Wilcoxon             | 0.310   | 0.305   | 0.296   | 0.268   | 0.286   |
| Std.dev              | 0.028   | 0.018   | 0.009   | 0.008   | 0.008   |
| |Mean| | 0.016   | 0.012   | 0.005   | 0.004   | 0.005   |
| Max                  | 0.218   | 0.151   | 0.083   | 0.071   | 0.064   |
| Fails at 2.5%        | 0.193   | 0.103   | 0.022   | 0.016   | 0.021   |
| **Revision, Final, Levels** |         |         |         |         |         |
| Std.dev              | 0.072   | 0.069   | 0.058   | 0.061   | 0.065   |
| |Mean| | 0.051   | 0.064   | 0.043   | 0.043   | 0.047   |
| Max                  | 0.226   | 0.222   | 0.146   | 0.172   | 0.156   |
| Fails at 5%          | 0.303   | 0.485   | 0.242   | 0.242   | 0.364   |
| **Revision, Final, Returns** |         |         |         |         |         |
| Std.dev              | 0.071   | 0.044   | 0.027   | 0.024   | 0.026   |
| |Mean| | 0.050   | 0.044   | 0.021   | 0.018   | 0.019   |
| Max                  | 0.218   | 0.151   | 0.083   | 0.069   | 0.061   |
| Fails at 2.5%        | 0.576   | 0.636   | 0.303   | 0.212   | 0.333   |
| **MCMC diagnostics** |         |         |         |         |         |
| $\bar{R}$ (mean)    | 1.000   | 1.000   | 1.000   | 1.001   | 1.000   |
| $\bar{R}$ (max)     | 1.001   | 1.001   | 1.000   | 1.014   | 1.008   |
| Eff. Sample Size (%) | 1.000   | 0.997   | 0.992   | 0.979   | 0.978   |

Volatility is the standard deviation of the index returns. ACF(1) denotes first-order autocorrelation. The ‘fails’ statistics represents what percentage of revisions exceeded the pre-defined threshold value. The threshold value is printed in the table. Wilcoxon is the (two tailed) Wilcoxon Signed-Rank test. Number represents the percentage of times an index was similar to its predecessors.

Before discussing the revisions in detail, a word is in order about how to ‘digest’ the results. Our quantitative tabular results of the index quality statistics are quite multi-dimensional. While all the metrics we will show and (at least briefly) discuss may not be necessary in all studies, we feel it is useful to consider them in the current...
### Table 4  Revision Statistics for Apartments in SFO, existing models

|                      | RS         | rRS        | RWd        | LLT        | LLTt       |
|----------------------|------------|------------|------------|------------|------------|
| **Traditional metrics** |            |            |            |            |            |
| WAIC                 | −23.897    | −58.508    | −55.643    | −71.905    |            |
| Noise (sd)           | 0.206      | 0.208      | 0.211      | 0.163      |            |
| Volatility           | 0.129      | 0.062      | 0.035      | 0.031      | 0.032      |
| ACF(1)               | −0.298     | −0.095     | 0.267      | 0.377      | 0.301      |
| **Revision, All, Levels** |            |            |            |            |            |
| Wilcox               | 0.534      | 0.311      | 0.330      | 0.270      | 0.352      |
| Std.dev              | 0.028      | 0.022      | 0.013      | 0.013      | 0.013      |
| |Mean|          | 0.015      | 0.015      | 0.008      | 0.008      | 0.008      |
| Max                  | 0.192      | 0.173      | 0.120      | 0.101      | 0.096      |
| Fails at 5%          | 0.052      | 0.029      | 0.009      | 0.009      | 0.008      |
| **Revision, All, Returns** |            |            |            |            |            |
| Wilcox               | 0.309      | 0.319      | 0.295      | 0.310      | 0.276      |
| Std.dev              | 0.034      | 0.018      | 0.007      | 0.007      | 0.007      |
| |Mean|          | 0.017      | 0.013      | 0.005      | 0.005      | 0.004      |
| Max                  | 0.204      | 0.105      | 0.048      | 0.066      | 0.050      |
| Fails at 2.5%        | 0.169      | 0.114      | 0.016      | 0.014      | 0.009      |
| **Revision, Final, Levels** |            |            |            |            |            |
| Std.dev              | 0.072      | 0.056      | 0.044      | 0.046      | 0.044      |
| |Mean|          | 0.051      | 0.053      | 0.035      | 0.038      | 0.036      |
| Max                  | 0.192      | 0.173      | 0.120      | 0.101      | 0.096      |
| Fails at 5%          | 0.364      | 0.424      | 0.182      | 0.212      | 0.152      |
| **Revision, Final, Returns** |            |            |            |            |            |
| Std.dev              | 0.069      | 0.035      | 0.021      | 0.023      | 0.021      |
| |Mean|          | 0.048      | 0.037      | 0.017      | 0.018      | 0.017      |
| Max                  | 0.190      | 0.105      | 0.048      | 0.066      | 0.050      |
| Fails at 2.5%        | 0.515      | 0.576      | 0.212      | 0.273      | 0.212      |
| **MCMC diagnostics** |            |            |            |            |            |
| $\bar{R}$ (mean)    | 1.000      | 1.001      | 1.000      | 1.004      | 1.000      |
| $\bar{R}$ (max)     | 1.000      | 1.001      | 1.003      | 1.038      | 1.008      |
| Eff. Sample Size (%) | 1.000      | 0.997      | 0.987      | 0.911      | 0.965      |

Volatility is the standard deviation of the index returns. ACF(1) denotes first-order autocorrelation. The ‘fails’ statistics represents what percentage of revisions exceeded the pre-defined threshold value. The threshold value is printed in the table. Wilcoxon is the (two tailed) Wilcoxon Signed-Rank test. Number represents the percentage of times an index was similar to its predecessors.

In context where we are introducing both new index methodologies and a new perspective for evaluating index quality. Admittedly, it will often be the case that most if not all of the separate metrics will ‘tell the same story’ about relative or comparative index quality. We find it most useful to digest the overall results and comparisons by
Table 5  Summary of the newly proposed repeat-sales models

| Model name     | Eqs.  | t-dist. | Literature |
|----------------|-------|---------|------------|
| Autoregressive | (AR)  | 2 / 6   | Yes        | –          |
| Aggregate prices | (AG) | 2 / 7   | Yes        | –          |
| AR + AG        | (ARAG)| 2 / 8   | Yes        | –          |

Table 6  Revision Statistics for Offices in LAI, new models

|                               | LLTt | AR   | AG   | ARAG  |
|-------------------------------|------|------|------|-------|
| **Traditional metrics**       |      |      |      |       |
| WAIC                          | 125.604 | 126.931 | 117.152 | 116.314 |
| Noise (sd)                    | 0.246 | 0.245 | 0.244 | 0.244 |
| Volatility                    | 0.041 | 0.042 | 0.044 | 0.044 |
| ACF(1)                        | 0.920 | 0.923 | 0.908 | 0.932 |
| **Revision, All, Levels**    |      |      |      |       |
| Wilcoxon                      | 0.312 | 0.338 | 0.130 | 0.146 |
| Std.dev                       | 0.017 | 0.016 | 0.014 | 0.014 |
| |Mean| 0.009 | 0.008 | 0.008 | 0.008 |
| Max                           | 0.156 | 0.183 | 0.170 | 0.155 |
| Fails at 5%                   | 0.024 | 0.021 | 0.016 | 0.017 |
| **Revision, All, Returns**   |      |      |      |       |
| Wilcoxon                      | 0.286 | 0.283 | 0.275 | 0.270 |
| Std.dev                       | 0.008 | 0.007 | 0.006 | 0.005 |
| |Mean| 0.005 | 0.004 | 0.003 | 0.003 |
| Max                           | 0.218 | 0.064 | 0.054 | 0.049 |
| Fails at 2.5%                 | 0.021 | 0.017 | 0.008 | 0.006 |
| **Revision, Final, Levels**  |      |      |      |       |
| Std.dev                       | 0.065 | 0.061 | 0.047 | 0.046 |
| |Mean| 0.047 | 0.046 | 0.033 | 0.033 |
| Max                           | 0.064 | 0.077 | 0.170 | 0.155 |
| Fails at 5%                   | 0.364 | 0.364 | 0.242 | 0.212 |
| **Revision, Final, Returns** |      |      |      |       |
| Std.dev                       | 0.026 | 0.025 | 0.020 | 0.018 |
| |Mean| 0.019 | 0.020 | 0.015 | 0.014 |
| Max                           | 0.061 | 0.077 | 0.054 | 0.049 |
| Fails at 2.5%                 | 0.333 | 0.303 | 0.152 | 0.121 |
| **MCMC diagnostics**         |      |      |      |       |
| $\bar{R}$ (mean)             | 1.000 | 1.000 | 1.000 | 1.000 |
| $\bar{R}$ (max)              | 1.008 | 1.002 | 1.004 | 1.008 |
| Eff. Sample Size (%)          | 0.978 | 0.983 | 0.960 | 0.965 |

Volatility is the standard deviation of the index returns. ACF(1) denotes first-order autocorrelation. The ‘fails’ statistics represents what percentage of revisions exceeded the pre-defined threshold value. The threshold value is printed in the table. Wilcoxon is the (two tailed) Wilcoxon Signed-Rank test. Number represents the percentage of times an index was similar to its predecessors.
reference to visual, graphical presentation of the comparisons, particularly the history of revisions in the indexes, as shown in Appendix D. Here we give the historic index revisions (Fig. 6), and path of revisions in both levels (Fig. 7) and returns (Fig. 8), similar in style to Fig. 1. In Fig. 1 we include the results of the RS and LLTt model. (And the ‘new’ ARAG model which will be discussed in “Proposing New Models”.

The index historical revisions graphs integrate and reflect all of the index quality metrics, and distill much of their essence for practical purposes, and they can be related qualitatively and intuitively to actual historical events.

Next we discuss some of the results in more detail. Firstly, we find that the gains are more profound in the returns than in the index levels. For example, the absolute mean revision in index levels for the RS model in LAI, is 1.4%. The absolute mean revision in index levels for the LLTt model decreases to 0.9%, an improvement of 35%. In returns, the absolute mean revision goes from 1.6% to 0.5%, an improvement of 70%. This is persistent over the markets and different models.

It is also clear from the get-go that the final periods tend to revise more severely compared to the overall index, both in levels and in returns. Indeed, compare the absolute mean revision of the ‘Revisions, Final’ with the ‘Revisions, All’, and you will find a factor three difference on average. This is also evident from Figs. 7 and 8, which gives the cumulative revisions in index levels and returns. Note that most of the ‘big’ revisions in Figs. 7 and 8 are early on. After the first 4 to 8 quarters, the revisions are smaller, and tend to be more ‘predictable’. (Except for perhaps the RS model.)

The LLTt model in SFO performs best measured by the revision statistics (Table 4). Almost exclusively this model performs better than any other model. Only the maximum index return revision is slightly lower for the random walk with drift model (RWd). This is interesting, especially considering that the volatility of the returns was lower for the LLT model, and the first-order autocorrelation (ACF(1)) higher, meaning that indeed, these metrics do not tell the complete story on its own.

For LAI (Table 3) it is more difficult to point to one ‘winner’. When comparing the revision statistics of the index levels (both all period and the final period) between the different models, the best model is the RWd. For the returns the LLT model is arguably the best, again measured over all periods and over the final period. The LLTt model is almost always ‘second’ best, making this model an overall good choice.

At this stage, the contribution of the newly proposed revisions statistics over the traditional metrics may appear minimal, as they suggest the same results in the comparison of the index methodologies. (Other than in perhaps LAI.) But we will see shortly that the revisions statistics do allow better discrimination when we introduce

---

5For example, consider Fig. 3. In both markets the crisis is clearly visible. Although the amplitude and the timing for especially the recovery is different. The SFO index started dropping 2 quarters before the LAI index. Subsequently, offices in LAI went down approximately 50%, whereas apartments in SFO ‘only’ decreased 20% in value. However, arguably the biggest difference is that at the end of the sample, the LAI index is still not above its previous peak, whereas the SFO index reached the previous peak already in 2013.
our new methodologies that are specifically designed to improve revision performance. (And as noted in the Introduction, the revisions statistics relate more directly and intuitively to the practical utility of the price indexes in important real world applications.)

Another advantage of the revisions metrics is that they allow us to compare the model-fit between different datasets, SFO and LAI. This is difficult to do using the other statistics, including the WAIC. As it happens, we see that the statistics are in this case pretty similar between the markets. The main exception is the revision of the final period level for the structural time series models. Here, SFO clearly has a better fit.

In terms of index specification overall, it is impressive how much the revisions statistics quantify the improvement that the structural time series repeat-sales models provide over the standard repeat-sales model. Indeed, the revisions in levels go down by a third on average. In returns, the gains are even more profound, on average returns revisions are reduced by almost 70%. In terms of index quality statistics, though, we note that the traditional metrics also favor the structural time series models. For the existing repeat sales models the revision statistics do not represent a large improvement over traditional methods of measuring index quality. However, the importance of revision statistics for measuring index quality, over and above traditional metrics, will be demonstrated in “Revisions in Newly Proposed Models”.

Moreover, there is also some concern regarding the structural time series models. As we are using structural time series in thin markets, we are particularly sensitive about the potential for ‘stickiness’ of indexes around major market turning points. This is because structural time series may be susceptible to missing a ‘break’ in the temporal structure.

If there is a structural break in the time series – a possible example is the 2008 financial crisis) – the specified structural time series model might treat the most recent observations as additive outliers, instead of innovative outliers (Harvey 1989; Durbin and Koopman 2012). As a result the structural break in the market is ‘ignored’, as it happens. This may be of particular concern for a sudden downturn in the market, as the number of observations will drop considerably as well see Fisher et al. (2003) and De Wit et al. (2013) for a detailed explanation on the positive price-volume correlation found in real estate.

In short, stickiness seems to exists in the existing structural time series repeat-sales models. See Fig. 6b and e, which give the full revision history of the LLTt model. The RWd and LLT models give similar results (not shown here to conserve space, they are available upon request). This ‘stickiness’ is not observed in the RS model (Fig. 6a and d). But of course, the RS model has other shortcomings, and obviously with it the market turning point may be initially obscured by random estimation error (noise) as the turning point is happening. As we have noted, market turning point stickiness is of particular concern for important practical applications of price indexes, such as economic statistics reports and derivatives trading support. Hence, the motivation to develop our new methodologies that we will review in the next section (but to cut to the chase, compare now Fig. 6c and f with Fig. 6b and e regarding turning point stickiness).
Finally, we also give some statistics on the performance of the MCMC algorithm. We give the $\bar{R}$, the potential scale reduction statistic (Gelman and Rubin 1992), and the effective sample size as percentage of the total sample size. Our models mix really well. On average, the effective sample size is close to the full sample size. The $\bar{R}$ is well under 1.1 for all models, which is commonly used as the convergence upper-limit (Lunn et al. 2013). Except for the LLT models all parameters even have a $\bar{R}$ of less than 1.01. See Appendix B for more details on the $\bar{R}$ statistic and the definition of ‘effective sample size’.

**Revisions in Newly Proposed Models**

Even though the focus in this paper is on how to use revision statistics as a measure of index quality, it is also interesting to explore new models. In “Proposing New Models” we discuss three variants of structural time series repeat-sales models. Structural time series models are very flexible, and given that they perform so well, it is worth investigating different specifications more carefully. We also put careful consideration at how to solve the ‘stickiness’ observed with the existing structural time series models, as discussed in “Revisions in Existing Models”. Section “Results” gives the revision results for these new models.

**Proposing New Models**

It has been noted in “Empirical Strategy and Revisions Statistics” that there are many reasons to assume that real estate returns should follow a (first-order) autoregressive process (see Case and Shiller; Case and Shiller; Quan and Quigley; Barkham and Geltner; Geltner et al., among others). In the first model we therefore explicitly specify the (log) periodic return in the state equation as a first-order autoregressive [AR(1)] process. To our knowledge, using an autoregressive component for the log index return – both in repeat-sales and hedonic price models – has not been done before.\(^6\) The AR(1) trend specification for $\mu$ in Eq. 2 is given by

$$
\Delta \mu_t \sim N \left( \rho \Delta \mu_{t-1}, \frac{\sigma^2 \eta}{1 - \rho^2} \right),
$$

where $\rho$ is the autoregressive parameter, which we restrict to be between 0 and 1. We denote this model AR.

In addition to the formal recognition of inertia in the private property market, the second model takes advantage of information from broader populations of properties. The basic idea is that granular real estate markets or segments tend to co-move with each other (Geltner et al. 2014, p. 554–556). This may reflect an aspect of the nature of risk in property asset markets. On the one hand there exists idiosyncratic price

---

\(^6\)A partial exception is Gatzlaff and Geltner (1998) which estimated repeat-sales indexes of Florida commercial property based on a ridge regression methodology reflecting an a priori assumption of positively correlated ‘true’ returns.
movement specific to individual assets or granular market segments, largely reflecting the space markets. But just as important, asset-valuation risk reflects changes over time in the capital market that cause changes in the opportunity cost of capital. Time variation in the discount rate causes at least as much volatility in prices (see for example Geltner and Mei 1995). This phenomenon is also widely accepted in the stock market (Shiller 1981). Such capital market based price movements may have more common elements across space market based granular market segments. We therefore add the returns of aggregate price indexes as an input variable for granular index returns.

We estimate and employ two aggregate indexes. The first is a metro-level ‘all property’ index, specific to the metro area of the target granular index but estimated on a pooled sample of all the main property types (offices, apartments, industrial and retail) within the relevant metro area. The second aggregate index is a sector-specific national index estimated separately only for the granular target property type sector, pooled across all the metros in the US.

The benefit of these aggregate indexes is that they use much larger samples and are therefore expected to be more ‘stable’. Also, as there is sufficient data, these indexes will probably not show the stickiness described in “Results”. The first index is referred to as the ‘market’ index, and the second as the ‘sector’ index. The former reflects metro specific factors common to all sectors in the metro, while the latter reflects sector specific factors common to all metros for the target sector. Both indexes are also estimated recursively similar to our markets of interest.

The ‘market’ index for the office indexes in LAI is based on the entire Los Angeles metro area for all properties (apartments, office, retail and industrial), which includes 5,000 pairs in total. The ‘sector’ - office - index is based on more than 10,000 pairs in the entire US. The ‘market’ index for SFO apartments is based on the entire San Francisco metro area for all property types (industrial, office, retail and apartment). In total we observe 2,639 pairs in this aggregate database. The ‘sector’ index consists of all apartments in the US, a total of approximately 14,000 pairs.

The aggregate indexes are first estimated in a separate step, using the AR model described by Eqs. 2 and 6. The results of these indexes are omitted from this paper for brevity, but are available upon request. The trend component is specified as

$$\Delta \mu_t \sim N \left( \sum \lambda_j \Delta p^A_{j,t}, \sigma^2_{\eta} \right), \tag{7}$$

where variable $p^A$ is the log price index at time $t$ for the aggregate market $j$, and it represents both the market and sector aggregate indexes, with corresponding

---

7 Note that Francke and van de Minne (2017) use a similar setup. Their hierarchical repeat-sales (HRS) model has multiple stochastic log price trends with a hierarchical additive structure: One common trend using information from all sectors and markets, and target market specific trends estimated as deviations from the common trend. In the present approach we simply estimate the aggregate indexes separately and use them as explanatory variables in the granular index estimation. This reduces computing time considerably. Moreover, we found that in some cases of data scarcity, the HRS can result in highly correlated indexes.

8 We do provide some of the revision statistics for the aggregate indexes in “Robustness”. 
Fig. 4 Estimated indexes using the new models. LLT is a LLT repeat-sales model with $t$-distributed measurement errors. AR is an autoregressive component in the trend. AG has a trend that is informed by a market and sector trend. ARAG is a combination of both the AR and AG models.

parameter $\lambda$. This model is denoted AG. In principle, other explanatory variables or structures could be added.\(^9\)

Finally, in the third model we mix both the AR and AG model, denoted by ARAG. The trend specification is given by

$$\Delta \mu_t \sim N \left( \rho \Delta \mu_{t-1} + \sum \lambda_j \Delta p^A_{j,t}, \sigma^2_1 - \rho^2 \right). \quad (8)$$

We use non-informative priors as defined earlier. All models assume the error term in Eq. 2 to be $t$-distributed, and are estimated by the NUTS algorithm. A summary of the models is provided in Table 5.

**Results**

The estimated indexes for the new models are given in Fig. 4. The Tables 6 and 7 provide revision statistics, goodness of fit measures, and some MCMC diagnostics for offices in LAI and apartments in SFO respectively. A summary of the posterior distributions (other than the index levels) is given in Tables 9 and 10. For comparison we also include the results of the Local Linear Trend model with $t$-distributed errors. (We pick this model as a benchmark as it is the ‘overall’ best traditional model.)

The model fit measured by the WAIC, improves for the AG and ARAG in LAI compared to the LLTt model, but not in SFO. Indeed, in LAI the gain in WAIC is almost 10 points, whereas in SFO the LLTt model has approximately a similar WAIC as the new models. The other more traditional metrics remain comparable, except for the autocorrelation in the returns in SFO. The AG and ARAG model results in relative

\(^9\)In an earlier version we experimented with transaction volume, regional GDP, and unemployment in the state equation, and allowed the variance of the signal and noise component to be time-varying. For the sake of brevity and readability these results have not been included in the paper. Moreover, the results did not improve much, or not at all.
Using Revisions as a Measure of Price Index Quality in Repeat-Sales...

Table 7  Revision Statistics for Apartments in SFO, new models

|                      | LLTt | AR  | AG  | ARAG |
|----------------------|------|-----|-----|------|
| Traditional metrics  |      |     |     |      |
| WAIC                 | −71.905 | −71.452 | −73.771 | −71.695 |
| Noise (sd)           | 0.163  | 0.162 | 0.166 | 0.168 |
| Volatility           | 0.032  | 0.034 | 0.028 | 0.025 |
| ACF(1)               | 0.301  | 0.322 | 0.559 | 0.705 |

| Revision, All, Levels |      |     |     |      |
| Wilcoxon             | 0.352  | 0.346 | 0.423 | 0.439 |
| Std.dev              | 0.013  | 0.012 | 0.011 | 0.010 |
| |Mean| | | |
| Max                   | 0.092  | 0.096 | 0.071 | 0.072 |
| Fails at 5%           | 0.008  | 0.008 | 0.003 | 0.003 |

| Revision, All, Returns |      |     |     |      |
| Wilcoxon              | 0.276  | 0.286 | 0.310 | 0.292 |
| Std.dev               | 0.007  | 0.007 | 0.006 | 0.005 |
| |Mean| | | |
| Max                   | 0.204  | 0.050 | 0.035 | 0.037 |
| Fails at 2.5%         | 0.009  | 0.014 | 0.009 | 0.007 |

| Revision, Final, Levels |      |     |     |      |
| Std.dev               | 0.044  | 0.043 | 0.035 | 0.035 |
| |Mean| | | |
| Max                   | 0.192  | 0.096 | 0.071 | 0.072 |
| Fails at 5%           | 0.152  | 0.182 | 0.091 | 0.091 |

| Revision, Final, Returns |      |     |     |      |
| Std.dev               | 0.021  | 0.023 | 0.019 | 0.018 |
| |Mean| | | |
| Max                   | 0.050  | 0.042 | 0.035 | 0.037 |
| Fails at 2.5%         | 0.212  | 0.364 | 0.242 | 0.182 |

| MCMC diagnostics      |      |     |     |      |
| R (mean)              | 1.000  | 1.000 | 1.000 | 1.000 |
| R (max)               | 1.008  | 1.007 | 1.003 | 1.002 |
| Eff. Sample Size (%)  | 0.965  | 0.980 | 0.969 | 0.966 |

Volatile is the standard deviation of the index returns. ACF(1) denotes first-order autocorrelation. The ‘fails’ statistics represents what percentage of revisions exceeded the pre-defined threshold value. The threshold value is printed in the table. Wilcoxon is the (two tailed) Wilcoxon Signed-Rank test. Number represents the percentage of times an index was similar to its predecessors.

high first-order autocorrelation, although as noted earlier, this does not necessarily imply better model fit. This is also evident from looking at the indexes in Fig. 4, which seem less ‘saw-tooth’ like than before in SFO.

Considering the traditional metrics only, one might conclude that the new models are not really an improvement over the existing models. However, by comparing the revision statistics, it is clear that the new models almost exclusively outperform the existing models measured by all revision statistics (absolute mean, standard deviation, maximum, and number of fails). This holds in specific for the models...
including aggregate indexes (AG and in particular ARAG), and for the the final period revisions.

Based on the statistics of revision in index levels, the AG and ARAG improve over the existing structural time series models with 30% (18%) in LAI (SFO): The absolute mean final period revision goes from 4.3% for the best model in LAI (RWd/LLT) to 3.3% for the best new model (both AG/ARAG). In SFO this statistic goes from 3.5% to 3.0% for the best models (RWd and AG/ARAG respectively). When comparing the final period index returns, the improvement is slightly less, with 22% and 12% for LAI and SFO receptively. Also the number of times the revisions exceed the predefined threshold value decreases considerably. For example, in 21% of the cases, the final period return revision is more than 2.5% for the best of the existing models (LLT) in LAI, compared to only 12% for the best new model (ARAG). The fit is still slightly better for SFO (final period index level revision), however the differences decrease even further compared to the existing models discussed in the previous Section.

Comparing the ARAG model with LLT model in Figs. 7 and 8 also reveals the gain visually. Especially in index levels (Fig. 7) the cumulative revisions are clearly smaller. Although note that in LAI some index points still revise downward even after a long period of time. We do not observe this pattern in index returns, nor in the other market. Reasons for this systematic downward revisions was given in “Introduction”.

We did two additional runs of the ARAG model, including only one of the aggregate indexes instead of both the sector and market indexes. The results give an indication of which aggregate indexes is dominant. In both LAI and SFO we generally find – based on the revision statistics – that the ARAG model including the sector index performs better compared to the one including the market index. In few instances, just having the sector index is even an improvement over including both aggregate indexes. In particular the final period revision statistics in SFO improve.

Figure 6c and f show the full revision history for LAI and SFO respectively, using the ARAG model. The Figures show that the ARAG indexes are less ‘sticky’ around market turns, like the crisis and subsequent recovery compared to the LLT model.

| Location                  | Property type | Min | Max | Market | Sector |
|---------------------------|---------------|-----|-----|--------|--------|
| Los Angeles Proper        | Office        | 181 | 660 | LA     | US Office |
| San Jose                  | Apartment     | 24  | 115 | SF     | US Apartment |
| United States             | Apartment     | 2,830 | 12,866 | –       | –       |
| United States             | Office        | 2,516 | 9,109 | –       | –       |
| Boston CBD                | Office        | 56  | 171 | BST    | US Office |
| Manhattan                 | Office        | 141 | 433 | NYC    | US Office |

LA stands for Los Angeles metro area, SF gives the San Francisco metro area, NYC is the New York metro area, and finally BST means the Boston metro area. All areas are defined as by RCA in our data.
The timing of the downturn and upturn is very timely, and the index does not revise much. This confirms the idea that granular indexes using aggregate indexes are less prone to this stickiness as there is more data available.

Note that the coefficients on the aggregate indexes go down after including the AR component, see Tables 9 and 10. Especially the coefficient on the sector index in LAI halves after the inclusion of the AR term, $1.600$ for the AG model versus $0.793/(1 - 0.237) = 1.039$ for the ARAG model. The AR coefficient ($\rho$) also decreases when combined with the aggregate indexes. The market index has a higher coefficient in SFO compared to the sector coefficient (0.6 versus 0.3). In LAI it is the other way around, the sector index has the higher coefficient. Finally, like the LLT\(_t\) model in “Results” the parameter estimates given in Tables 9 and 10 clearly show that the errors in the measurement equation follow a \(t\)-distribution: The posterior means of \(\nu\) are between 5 and 9 for all models.

**Robustness**

As a robustness check we also ran our models on a larger set of markets. These markets vary in size (i.e. number of transaction pairs) and geography. More specifically, we ran our model on six other markets and computed the same statistics as in the previous Sections. A summary of these markets can be found in Table 8. In this Section we will quickly report out findings, however a tabular with the most important metrics can be found in Appendix E.

Our first ‘new’ market is Los Angeles proper, which is the biggest market within the Los Angeles metro area, with 181 (660) pairs at the start (end) of the sample. We use the same aggregate indexes here as we did for our LAI market. San Jose is in the San Francisco metro area, similarly to the previously analyzed Oakland market. As such, we use the same aggregate indexes here as well. Its a very small market (the least amount of transaction pairs in this paper), with only 24 transaction pairs at the start of our sample, and only slightly over 100 at the end. Our largest ‘new’ markets are for the United States office and apartment sectors, with each around 10,000 observations over the full sample, and more than 2,000 observations at the start of the sample. These markets have no aggregate indexes ($\Delta p_{j,t}^A$). As a result we cannot estimate the AG and ARAG models here. In fact, the estimated AR indexes themselves are used as aggregate indexes for the other markets. We also estimate indexes for two completely different geographies. Firstly we have offices in the relatively small geographical area of the Boston Central Business District (CBD). The aggregate indexes are an all property type Boston metro area index, and the previously used US office index. We observe between 56 and 171 transaction pairs in said market. The second added market is for offices in Manhattan. Even though the geography is small as well, the amount of transaction pairs is higher than in San Jose, LAI and SFO (between 141 and 433 pairs).

The results follow an expected pattern. According to the traditional metrics, the indexes ‘fit’ better if one uses structural time series modeling. Especially the local linear trend with \(t\)-distributed errors does well (LLT\(_t\)). However, our newly proposed
models do not perform better than the LLTt model, using the traditional metrics of fit only. (A notable exception is in the Los Angeles proper market, where the difference in WAIC is larger than 5 points between the LLTt and AG/ARAG models.) It is only after analyzing the revision statistics, that it becomes obvious that the newly proposed models are indeed an overall improvement. For example, in San Jose the average absolute mean revision goes from 1.1% for the LLTt model, to 0.8% for the ARAG model, a 27% improvement. However, note that the gain is not as large in all markets, if there is a gain at all. Most notably for our largest markets; US apartments and offices. In these markets, the absolute mean revision is approximately similar between all models and markets, including the Bailey et al. (1963) repeat-sales model. The average absolute mean revision hovers around 0.4% and 0.5% per quarter. In other words, the benefit of using structural time series models diminishes if the number of observations becomes larger (and vice versa), which is also noted by Francke (2010). However, the final period revisions do improve overall with structural time series models. For example, the final period revision for the US apartment index returns goes down from 4.3% (RS model) to 3.1% (ARAG model).

Note that the amount of revisions in returns for our ‘best’ San Jose apartments model (ARAG) is equal to those of the standard repeat sale model for US apartments. More specifically, the average absolute mean revision of the returns is 0.4% for both the ARAG model in San Jose and the RS model for US apartments. The mean revisions of the index level are better for the RS model for US apartments (0.4%) compared to the San Jose index (0.8%). However, this is still admirable after realizing that there are 100 times as many transaction pairs available for our US apartment index. (Notice again how we can compare the revision statistics over the different markets in this way, but not so much the traditional metrics.)

Conclusion

In this paper we focus on index revisions in repeat-sales models applied to granular commercial property markets. Absence of revision is desirable for multiple reasons. However, due to the structure of the repeat-sales model and the relative low number of transactions – in particular compared to other markets, like housing – revisions in commercial real estate markets are usually quite substantial. In the main body of text we apply our models to offices in Inland Empire and Orange County (in Los Angeles metro area) and to apartments in Oakland, San Francisco. In both markets we only observe between the 100 to 340 pairs in the revision analysis. As a robustness check also apply our model to multiple other markets with varying amount of observations.

We distinguish between systematic, random and data generating process related revision. The focus in this paper is on the latter two. Random revisions are fundamentally caused by the dispersion in transaction prices, and as a result in property price returns. The filtering of the noise out of the individual returns depends on the specification of the data generating process, in particular with respect to the trend. In this paper we consider some variants of structural time series repeat-sales models, in which time fixed effects are replaced by a stochastic trend specification.
Previous literature already shows that structural time series repeat-sales models are less sensitive to low numbers of observations and outliers. The focus in previous literature has never been on index revision. In this paper we show that different (structural time series) model specifications lead to different indexes, and more importantly different index revision behavior.

Acknowledging the importance of index revisions in practical real estate applications, we argue that revision statistics should be used as important indicators of index quality. For that reason we propose some metrics to measure the severity of index revisions, based on the revisions in the final period and the full history, in both index levels and index returns.

Model selection is typically done by comparing likelihood based information criteria, such as the Watanabe Akaike Information Criterion. However, we observe that different repeat-sales models have almost equal model fit as measured by WAIC, but that they do differ in revision behavior by quite a margin. We advocate that in these situations revision statistics should be taken into account to select the ‘correct’ model. Moreover, information criteria do not allow you to easily compare fit between different markets (different datasets). In contrast, differences in revisions statistics between markets are easy to interpret.

We find that newly proposed models have better performance in revision statistics. The most important properties of these new models are the following. First, we replace the local linear trend specification for the log index level by a first order autoregressive process for the log index return. Second, we add aggregate ‘sector’ and ‘market’ index returns as explanatory variables for the granular returns. And finally, we allow for fat tails by assuming a $t$-distribution for the measurement error, where we estimate the degrees of freedom.

It should be noted that structural time series models are very flexible. Future research could therefore greatly benefit from more experimenting with different index return specifications, including different lag structures, error distributions and explanatory variables.

Acknowledgements. We would like to thank the attendees at the Property Indices session of the 2017 international AREUEA conference in Amsterdam and at the 2016 Hoyt meeting in West Palm Beach. Specifically, we would like to thank Daniel Melser and Martijn Dröes whom both discussed the paper. We also want to thank Real Capital Analytics for providing the data and insights that made this study possible. The comments and observations by Jim Sempere, Elizabeth Szep and Willem Vlaming at Real Capital Analytics were especially appreciated. Finally, we would like to thank the anonymous referee for his comments on this paper.

Appendix A: Revisions in Repeat-Sales Models

A.1 Revisions in the Standard Repeat-Sales Model

The way indexes are revised over time is illustrated by Bailey et al. (1963) and Clapp and Giaccotto (1999). Suppose we have only three periods, $t = 0, 1, 2$, and we are
interested in the revision subsequent to time 1 of an index at time 1, originally es-
imated from transactions in period 0 and 1. That is, consider the revision of the index
history when the transactions for period 2 become available. Let \( \bar{r}_{st} \) be the average
log price change between the period of buy \( s \) and the period of sell \( t \) and let \( \hat{\delta}_{T|\tau} \)
be the estimated log index return for period \( T \) conditional on all information (trans-
actions) up to time \( \tau \), where \( \tau \geq T \). The initial estimate of the log index return for
period 1 is simply the average return of the properties in period 1, \( \hat{\delta}_{1|1} = \bar{r}_{01} \), where \( n_{st} \)
is the number of transaction pairs with time of buy \( s \) and time of sell \( t \). Now the estimated log index return for
period 1 conditional on all information up to time 2 is a weighted average of the original estimate
\( \hat{\delta}_{1|1} \) and new information with respect to \( \delta_1 \) based on the difference
\( \bar{r}_{02} - \bar{r}_{12} \), with variance \( 2\sigma^2_{\epsilon}(1/n_{02} + 1/n_{12}) \),

\[
\hat{\delta}_{1|2} = \frac{(n_{02} + n_{12})n_{01}\bar{r}_{01} + n_{02}n_{12}(\bar{r}_{02} - \bar{r}_{12})}{n_{01}(n_{02} + n_{12}) + n_{12}n_{02}} = w\bar{r}_{01} + (1-w)(\bar{r}_{02} - \bar{r}_{12}),
\]

where the weight \( w = (n_{01}(n_{02} + n_{12})) / (n_{01}(n_{02} + n_{12}) + n_{02}n_{12}) \) depends on the
variances of \( \bar{r}_{01} \) and \( \bar{r}_{02} - \bar{r}_{12} \). The revision \( \hat{\delta}_{1|2} - \hat{\delta}_{1|1} \) can be expressed as \( (1-w)(\bar{r}_{02} - \bar{r}_{12} - \bar{r}_{01}) \).

Let us illustrate the revision by a numerical example. Assume that \( \bar{r}_{01} = 0.08 \), \( \bar{r}_{02} = 0.15 \) and \( \bar{r}_{12} = 0.10 \), and \( n_{01} = n_{02} = n_{12} = 25 \). The initial estimate
\( \hat{\delta}_{1|1} = \bar{r}_{01} = 0.08 \). The estimate of \( \delta_1 \) after transactions in period 2 are available, is
a weighted average of \( \bar{r}_{01} = 0.08 \) and \( \bar{r}_{02} - \bar{r}_{12} = 0.05 \), with weight \( w = 2/3 \), so
\( \hat{\delta}_{1|2} = 0.07 \), and the revision is \(-0.01\).

A.2 Revisions in a Random Walk Repeat-Sales Model

Revisions in structural time series repeat-sales models are different from the standard
(Bailey et al. 1963) model described in A.1. In this appendix we provide an algebraic
derivation of how indexes are revised over time for a random walk repeat-sales model.
For more advanced models the revision expressions become very complicated.

The random walk repeat-sales model for 3 observations \( \bar{r}_{01}, \bar{r}_{02} \) and \( \bar{r}_{12} \) can be
expressed as

\[
\begin{pmatrix}
\bar{r}_{01} \\
\bar{r}_{02} \\
\bar{r}_{12} \\
0 \\
0
\end{pmatrix} =
\begin{pmatrix}
1 & 0 \\
1 & 1 \\
0 & 1 \\
1 & 0 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
\delta_1 \\
\delta_2 \\
\bar{\epsilon}_{01} \\
\bar{\epsilon}_{02} \\
\bar{\epsilon}_{12} \\
\eta_1 \\
\eta_2
\end{pmatrix},
\]

where \( \delta_t = \Delta \mu_t \), \( \text{Var}(\bar{\epsilon}_{st}) = 2\sigma^2_{\epsilon}/n_{st} \), and \( \text{Var}(\eta_t) = \sigma^2_{\eta} \). Rows 1 – 3 are
the observation equations, and rows 4 and 5 represent the random walk specification.
Define the signal-to-noise ratio \( q \) as \( \sigma^2_{\eta}/(2\sigma^2_{\epsilon}) \). If only the transactions up to time
1 are available, the model consists only of rows 1 and 4. Note that for the standard Bailey et al. repeat-sales model only rows 1 – 3 are applicable.

The model can be estimated by weighted least squares, giving

\[
\hat{\delta}_1 | 1 = \frac{n_{01}}{n_{01} + 1/q} \bar{r}_{01}, \quad \text{based on transactions up to time 1}
\]

\[
\hat{\delta}_1 | 2 = \frac{(n_{02} + n_{12})n_{01}\bar{r}_{01} + n_{02}n_{12}(\bar{r}_{02} - \bar{r}_{12}) + 1/q(n_{01}\bar{r}_{01} + n_{02}\bar{r}_{02})}{(n_{01} + 1/q)(n_{02} + n_{12} + 1/q) + (n_{12} + 1/q)n_{02}},
\]

the last equation based on on transactions up to time 2. Note that if \( q \to \infty \), the expressions are identical to the ones in the previous section.

If we assume that the signal-to-noise ratio is equal to \( q = 0.1 \), and using the same numbers as in the previous example, then \( \hat{\delta}_1 | 1 = 0.057 \) (instead of 0.08), and \( \hat{\delta}_1 | 2 = 0.063 \) (instead of 0.07). The revision is equal to 0.006, and is smaller in absolute magnitude than the revision in the standard repeat sales model (0.006 versus 0.01).

A.3 Calculating Revisions for Different Signal-to-Noise Ratios

In this Section we calculate revisions for different values of the signal-to-noise ratios \( q \) in the random walk repeat-sales model as described in A.2. Our purpose here is simply to demonstrate the large role that \( q \) plays in revisions. We use typical values for the signal \( (\sigma^2_\eta) \) and noise \( (\sigma^2_\epsilon) \).

We use the LAI repeat-sales data, see “Data and Descriptive Statistics” to compute the indexes starting from 34 quarters back. We have 338 transaction observation pairs in the final dataset, or about 4.5 per quarter on average. With each new quarter of history, approximately 10 new pairs are added, including second-sales in the new quarter as well as additional pairs earlier in the history. In the calculations we use signal-to-noise ratios \( q \) in the range between 0.02 and than 0.6. This range covers the ratios that we do indeed estimate.

Figure 5 provides revision statistics and the volatility and the autocorrelation of the index returns for different values of the signal-to-noise ratio \( q \). In order to conserve space we only look at the absolute mean revisions in levels and returns, for both the total sample and specific for the final period.

It can be seen that lower values of \( q \) give less revisions. Note that in an extreme case with \( q = 0 \) there would be no revisions; the indexes would be flat. In this case there would also be a first-order autocorrelation of 1, and no volatility. The increase in levels revisions is not linear with an increase in \( q \). For example, with \( q = 0.1 \) the average (absolute) revision is 7 bps, whereas with \( q = 0.5 \), the average (absolute) revision is ‘just’ 10 bps. However, linearity does seem to hold in the average (absolute) return revisions (Fig. 5d). More interesting is that the final period return revision behaves like a concave function of the assumed signal-to-noise ratio (Fig. 5e). Our results show that at \( q = 0.16 \) and \( q = 0.60 \) the final period index level revisions are equally large (equal cumulative historical change). For the returns revisions the concavity seems to be mostly absent.
Fig. 5 Revisions in LAI for different simulated values of the signal-to-noise ratio $q$
Appendix B: Technical Estimation Details

This appendix contains some technical details on the model estimation. We apply the No-U-Turn-Sampler (NUTS) developed by Hoffman and Gelman (2014). It is a generalization of the Hamiltonian Monte Carlo (HMC) algorithm. The HMC avoids the random walk behaviour and sensitivity to correlated parameters that plague other Markov chain Monte Carlo methods by taking a series of steps informed by first-order gradient information. These features allow it to converge to high-dimensional target distributions much more quickly compared to for example the Metropolis-Hastings algorithm and Gibbs sampling. In addition, the NUTS algorithm avoids setting the parameters for the step size and the desired numbers of steps, which the HMC is so sensitive to. In case of many parameters, the manual setting of step sizes and desired number of steps, becomes intractable. The NUTS algorithm has been introduced in R via the software package called ‘Rstan’ (Carpenter et al. 2017).

The difference in convergence between NUTS and Gibbs sampling is clearly illustrated by the the more elaborate ARAG model. In our applications Gibbs sampling, by using the program JAGS (Plummer 2003), could take hours to converge, while NUTS only needed 10 – 20 seconds.

We apply the ‘Matt trick’ to sample \( \mu_t \) by specifying \( \mu_t = \mu_{t-1} + \eta_t \), where the increments \( \eta_t \sim \text{Normal}(0, 1) \) are independent of each other. The effective sample size is increased considerably by doing so, simply because there is less correlation in the index returns compared to index levels.\(^{10}\) Intuitively, without the ‘Matt trick’, if an ‘extreme’ value is sampled for price level in period \( t \), this will affect the price level sampled in period \( t+1, t+2, \) etc. By sampling increments, this relationship is gone. A more technical description is given by Betancourt and Girolami (2015).

We sample 2,000 times, of which we discard the first 1,000 when analysing the chains (i.e. the warm-up stage), in parallel over 6 chains (so \( 6 \times 1,000 = 6,000 \) samples in total). We provide different initial values for each variable in each chain and we do not thin the chains (see Link and Eaton 2012, for our reasons to shy away from thinning). We subsequently evaluate whether or not the model converges based on \( \bar{R} \), the potential scale reduction statistic.\(^{11}\) We use very strict convergence criteria, i.e. average \( \bar{R} \leq 1.01 \), which is more strict than usual. The fact that the models for our granular indices converge even with such strict criteria, shows the power of the NUTS algorithm. As shown in “Results” and “Results”, the effective and total sample size are almost identical for most models. We also keep track of the Monte Carlo (MC) error (Koehler et al. 2009), but these are not presented in this paper to conserve

\(^{10}\)The effective sample size (ESS) is computed as follows; \( \text{ESS} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho(k)} \), where \( n \) is the number of samples and \( \rho(k) \) is the correlation at lag \( k \). One gets a different ESS for every variable. Thus, the difference between the effective sample size and the actual sample size, gives one a measure on how independent the draws are.

\(^{11}\)The intuition behind the \( \bar{R} \) is that the chains should look alike, if the chains converged. First, the Gelman-Rubin diagnostic is computed, which calculates both the between-chain and the within-chain variance. The \( \bar{R} \) is essentially the fraction between the two, see Gelman and Rubin (1992) and Brooks and Gelman (1998) for more details. A value of 1.1 is usually used as an upper limit.
space. Even with a sample size of only 6,000, the MC error divided by the mean or standard deviation of the posterior, remains very low.

### Appendix C: Parameter Estimates and Credible Intervals

#### Table 9  Summary of Posterior Distributions for LAI

|                  | RS  | RWd | LLT | LLTt | AR  | AG  | ARAG |
|------------------|-----|-----|-----|------|-----|-----|------|
| **Parameters**   |     |     |     |      |     |     |      |
| $\rho$           | mean|     |     |      |     |     |      |
|                  |     | 0.598|     |      |     | 0.237|      |
|                  | median| 0.643|     |      |     | 0.222|      |
|                  | 2.5% | 0.110|     |      |     | 0.120|      |
|                  | 97.5%| 0.893|     |      |     | 0.543|      |
| $\nu$            | mean| 8.613| 8.476| 8.328| 8.238|     |      |
|                  | median| 7.900| 7.798| 7.650| 7.541|     |      |
|                  | 2.5% | 4.300| 4.288| 4.250| 4.243|     |      |
|                  | 97.5%| 16.973| 16.488| 16.278| 15.952|     |      |
| $\lambda_{\text{Market}}$ | mean|     |     |      |     |      | 0.281| 0.220|
|                  | median|     |     |      |     |      | 0.240| 0.156|
|                  | 2.5% |     |     |      |     |      | 0.130| 0.800|
|                  | 97.5%|     |     |      |     |      | 0.787| 0.653|
| $\lambda_{\text{Sector}}$ | mean|     |     |      |     |      | 1.600| 0.793|
|                  | median|     |     |      |     |      | 1.400| 0.800|
|                  | 2.5% |     |     |      |     |      | 0.466| 0.340|
|                  | 97.5%|     |     |      |     |      | 1.369| 1.214|
| **Hyperparameters** |     | |     |      |     |      |      |      |
| $\sigma_\epsilon$ | mean| 0.284| 0.280| 0.281| 0.246| 0.245| 0.244| 0.244|
|                  | median| 0.284| 0.280| 0.281| 0.246| 0.245| 0.244| 0.244|
|                  | 2.5% | 0.263| 0.260| 0.261| 0.217| 0.216| 0.216| 0.215|
|                  | 97.5%| 0.380| 0.320| 0.340| 0.274| 0.273| 0.272| 0.271|
| $\sigma_\eta$    | mean| 0.710| 0.240| 0.260| 0.250| 0.230| 0.160|      |
|                  | median| 0.700| 0.200| 0.230| 0.200| 0.200| 0.210| 0.150|
|                  | 2.5% | 0.520| 0.100| 0.100| 0.400| 0.110| 0.600|      |
|                  | 97.5%| 0.960| 0.690| 0.720| 0.690| 0.410| 0.330|      |
Table 10  Summary of Posterior Distributions for SFO

| Parameters | RS | RWd | LLT | LLTt | AR | AG | ARAG |
|------------|----|-----|-----|------|----|----|------|
| ρ          |     |     |     |      |    |    |      |
| mean       | 0.193 |     |     |      |    |    |      |
| median     | 0.148 | 0.131 |     |      |    |    |      |
| 2.5%       | 0.600 |     | 0.980 |      |    |    |      |
| 97.5%      | 0.626 |     | 0.436 |      |    |    |      |
| ν          |     |     |     |      |    |    |      |
| mean       | 5.610 | 5.890 | 5.295 | 5.383 |    |    |      |
| median     | 4.678 | 4.729 | 4.887 | 5.190 |    |    |      |
| 2.5%       | 2.654 | 2.766 | 2.850 | 2.990 |    |    |      |
| 97.5%      | 9.572 | 9.664 | 9.939 | 1.238 |    |    |      |
| λ_Market   |     |     |     |      |    |    |      |
| mean       | 0.650 |     | 0.523 |      |    |    |      |
| median     | 0.620 |     | 0.527 |      |    |    |      |
| 2.5%       | 0.970 | 0.990 |      |      |    |    |      |
| 97.5%      | 1.350 | 0.948 |      |      |    |    |      |
| λ_Sector   |     |     |     |      |    |    |      |
| mean       | 0.287 | 0.255 |      |      |    |    |      |
| median     | 0.250 | 0.222 |      |      |    |    |      |
| 2.5%       | 0.140 | 0.110 |      |      |    |    |      |
| 97.5%      | 0.778 | 0.680 |      |      |    |    |      |
| Hyperparameters |     |     |     |      |    |    |      |
| σ_ε         |     |     |     |      |    |    |      |
| mean       | 0.260 | 0.280 | 0.211 | 0.163 | 0.162 | 0.166 | 0.168 |
| median     | 0.260 | 0.280 | 0.210 | 0.162 | 0.162 | 0.165 | 0.167 |
| 2.5%       | 0.188 | 0.191 | 0.192 | 0.135 | 0.136 | 0.139 | 0.141 |
| 97.5%      | 0.226 | 0.228 | 0.231 | 0.191 | 0.190 | 0.193 | 0.194 |
| σ_η         |     |     |     |      |    |    |      |
| mean       | 0.580 | 0.490 | 0.500 | 0.520 | 0.360 | 0.300 |      |
| median     | 0.570 | 0.490 | 0.510 | 0.530 | 0.340 | 0.280 |      |
| 2.5%       | 0.360 | 0.900 | 0.120 | 0.160 | 0.170 | 0.110 |      |
| 97.5%      | 0.850 | 0.840 | 0.840 | 0.840 | 0.630 | 0.560 |      |
Appendix D: Revisions for a Selection of Indexes

Fig. 6 Revisions for a selection of the indexes in LAI and SFO
Fig. 7  Cumulative revisions of index levels at different periods
Fig. 8 Cumulative revisions of index returns at different periods

(a) RS in LAI.  
(b) LLTt in LAI.

(c) ARAG in LAI.  
(d) RS in SFO.

(e) LLTt in SFO  
(f) ARAG in SFO
### Appendix E: Revision Results of our Robustness Section

#### Table 11 Revisions in our ‘robustness’ markets

| Metric | RS  | rRS | RWd  | LLT | LLTt | AR  | AG  | ARAG |
|--------|-----|-----|------|-----|------|-----|-----|------|
|        |     |     |      |     |      |     |     |      |
| Los Angeles proper office (Obs. 181 – 660) | | | | | | | | |
| [Mean] (level) | 0.014 | 0.012 | 0.009 | 0.009 | 0.009 | 0.009 | 0.007 | 0.007 |
| [Mean] (diff) | 0.016 | 0.011 | 0.006 | 0.005 | 0.004 | 0.004 | 0.003 | 0.002 |
| [Mean] (final, level) | 0.069 | 0.051 | 0.048 | 0.052 | 0.048 | 0.044 | 0.024 | 0.022 |
| [Mean] (final, diff) | 0.067 | 0.033 | 0.024 | 0.024 | 0.020 | 0.020 | 0.012 | 0.011 |
| Std.dev (returns) | 0.113 | 0.066 | 0.036 | 0.030 | 0.032 | 0.032 | 0.033 | 0.032 |
| WAIC | 444.647 | 389.872 | 386.658 | 369.528 | 367.854 | 353.268 | 353.678 | 353.678 |
| Noise (sd) | 0.320 | 0.318 | 0.320 | 0.273 | 0.273 | 0.274 | 0.274 | 0.274 |
| San Jose apartments (Obs. 24 – 115) | | | | | | | | |
| [Mean] (level) | 0.025 | 0.034 | 0.013 | 0.014 | 0.011 | 0.010 | 0.009 | 0.008 |
| [Mean] (diff) | 0.023 | 0.021 | 0.006 | 0.007 | 0.007 | 0.006 | 0.005 | 0.004 |
| [Mean] (final, level) | 0.079 | 0.083 | 0.038 | 0.044 | 0.039 | 0.038 | 0.030 | 0.029 |
| [Mean] (final, diff) | 0.050 | 0.036 | 0.015 | 0.020 | 0.018 | 0.017 | 0.015 | 0.015 |
| Std.dev (returns) | 0.141 | 0.043 | 0.043 | 0.039 | 0.044 | 0.043 | 0.040 | 0.037 |
| WAIC | −40.820 | −39.007 | −36.597 | −53.809 | −54.281 | −56.404 | −56.365 | −56.365 |
| Noise (sd) | 0.163 | 0.184 | 0.186 | 0.126 | 0.127 | 0.130 | 0.132 | 0.132 |
| United States apartment (Obs. 2,830 – 12,866) | | | | | | | | |
| [Mean] (level) | 0.004 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |
Table 11 (continued)

| Metric                     | RS   | rRS  | RWd  | LLT  | LLTt | AR   | AG   | ARAG |
|----------------------------|------|------|------|------|------|------|------|------|
| Mean (diff)                | 0.004| 0.004| 0.003| 0.002| 0.002| 0.002|      |      |
| Mean (final, level)        | 0.015| 0.011| 0.016| 0.013| 0.014| 0.013|      |      |
| Mean (final, diff)         | 0.015| 0.011| 0.013| 0.010| 0.010| 0.010|      |      |
| Std. dev (returns)         | 0.043| 0.040| 0.037| 0.031| 0.031| 0.031|      |      |
| WAIC                       | −3,945.950| −3,931.340| −3,929.420| −3,544.990| −3,543.830| |      |      |
| Noise (sd)                 | 0.328| 0.328| 0.328| 0.262| 0.262|      |      |      |
| United States office (Obs. 2,516 – 9,109) | | | | | | | | |
| Mean (level)               | 0.005| 0.006| 0.004| 0.004| 0.004| 0.005| 0.005| 0.005|
| Mean (diff)                | 0.005| 0.006| 0.003| 0.002| 0.002| 0.003| 0.003| 0.002|
| Mean (final, level)        | 0.016| 0.020| 0.019| 0.018| 0.021| 0.018|      |      |
| Mean (final, diff)         | 0.016| 0.020| 0.014| 0.012| 0.014| 0.012|      |      |
| Std. dev (returns)         | 0.047| 0.047| 0.038| 0.034| 0.034| 0.035|      |      |
| WAIC                       | −4,822.830| −4,800.010| −4,795.900| −4,610.300| −4,609.170| |      |      |
| Noise (sd)                 | 0.409| 0.409| 0.409| 0.349| 0.349|      |      |      |
| Boston CBD office (Obs. 56 – 171) | | | | | | | | |
| Mean (level)               | 0.019| 0.021| 0.012| 0.012| 0.011| 0.010| 0.010| 0.010|
| Mean (diff)                | 0.019| 0.015| 0.004| 0.005| 0.004| 0.003| 0.003| 0.003|
| Mean (final, level)        | 0.092| 0.066| 0.042| 0.047| 0.042| 0.036| 0.026| 0.026|
| Mean (final, diff)         | 0.083| 0.039| 0.013| 0.015| 0.012| 0.011| 0.012| 0.011|
| Std. dev (returns)         | 0.179| 0.059| 0.046| 0.045| 0.038| 0.036| 0.039| 0.039|
| WAIC                       | −81.153 | −69.851 | −70.017 | −41.508 | −41.636 | −37.771 | −37.216 |      |
| Noise (sd)                 | 0.328| 0.336| 0.338| 0.219| 0.220| 0.219| 0.219| 0.218 |
| Metric               | RS    | r_RS   | RWd   | L_LT  | LLT   | LLTt  | AR    | AG    | ARAG  |
|---------------------|-------|--------|-------|-------|-------|-------|-------|-------|-------|
| Manhattan office (Obs. 141 – 433) |       |        |       |       |       |       |       |       |       |
| Mean (level)        | 0.011 | 0.009  | 0.008 | 0.006 | 0.005 | 0.004 | 0.004 | 0.003 | 0.002 |
| Mean (diff)         | 0.012 | 0.009  | 0.005 | 0.004 | 0.003 | 0.002 | 0.002 | 0.001 |       |
| Mean (final, level) | 0.064 | 0.054  | 0.046 | 0.038 | 0.036 | 0.028 | 0.028 | 0.021 |       |
| Mean (final, diff)  | 0.059 | 0.054  | 0.046 | 0.038 | 0.036 | 0.028 | 0.028 | 0.021 |       |
| Std.dev (returns)   | 0.157 | 0.072  | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 |
| WAIC                | 173.475 | 155.674 | 155.912 | 140.242 | 139.522 | 134.200 | 134.631 | 134.631 | 134.631 |
| Noise (sd)          | 0.330 | 0.334  | 0.334 | 0.263 | 0.263 | 0.263 | 0.262 | 0.262 | 0.262 |

Using Revisions as a Measure of Price Index Quality in Repeat-Sales...
References

Abraham, J.M., & Schauman, W.S. (1991). New evidence on home prices from Freddie Mac repeat sales. *Real Estate Economics*, 19(3), 333–352.

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

Barkham, R., & Geltner, D.M. (1995). Price discovery in American and British property markets. *Real Estate Economics*, 23(1), 21–44.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79, 30.

Bokhari, S., & Geltner, D. (2012). Estimating real estate price movements for high frequency tradable indexes in a scarce data environment. *The Journal of Real Estate Finance and Economics*, 43(2), 522–543.

Bokhari, S., & Geltner, D.M. (2011). Loss aversion and anchoring in commercial real estate pricing: empirical evidence and price index implications. *Real Estate Economics*, 39(4), 635–670.

Bourassa, S.C., Cantoni, E., Hoesli, M. (2013). Robust repeat sales indexes. *Real Estate Economics*, 41(3), 517–541.

Brooks, S.P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A. (2017). Stan: a probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32.

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

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

Clapham, E., Englund, P., Quigley, J.M., Redfearn, C.L. (2006). Revisiting the past and settling the score: index revision for house price derivatives. *Real Estate Economics*, 34, 275–302.

Clapp, J.M., & Giaccotto, C. (1999). Revisions in repeat-sales price indexes: here today, gone tomorrow. *Real Estate Economics*, 27, 79–104.

Clements, M.P., & Galvão, A.B. (2017). Predicting early data revisions to us gdp and the effects of releases on equity markets. *Journal of Business & Economic Statistics*, 35, 1–18.

De Wit, E.R., Englund, P., Francke, M.K. (2013). Price and transaction volume in the Dutch housing market. *Regional Science and Urban Economics*, 43(2), 220–241.

Deng, Y., & Quigley, J.M. (2008). Index revision, house price risk, and the market for house price derivatives. *The Journal of Real Estate Finance and Economics*, 37(3), 191–209.

Durbin, J., & Koopman, S.J. (2012). *Time series analysis by state space methods* Vol. 2. Oxford: Oxford Univ Press.

Fisher, J., Gatzlaff, D., Geltner, D.M., Haurin, D. (2003). Controlling for the impact of variable liquidity in commercial real estate price indices. *Real Estate Economics*, 31(2), 269–303.

Francke, M.K. (2010). Repeat sales index for thin markets: a structural time series approach. *Journal of Real Estate Finance and Economics*, 41, 24–52.

Francke, M.K. (2017). Repeat sales models, holding periods and index revision. The Hoyt Group; May 2017 50th Anniversary Program Presentations.

Francke, M.K., & De Vos, A.F. (2000). Efficient computation of hierarchical trends. *Journal of Business and Economic Statistics*, 18, 51–57.

Francke, M.K., & van de Minne, A. (2017). The hierarchical repeat sales model for granular commercial real estate and residential price indices. *The Journal of Real Estate Finance and Economics*, 55(4), 511–532.
Gatzlaff, D.H., & Geltner, D.M. (1998). A transaction-based index of commercial property and its comparison to the NCREIF index. *Real Estate Finance, 15*(1), 7–22.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–72.

Geltner, D., & Mei, J. (1995). The present value model with time-varying discount rates: implications for commercial property valuation and investment decisions. *The Journal of Real Estate Finance and Economics, 11*(2), 119–135.

Geltner, D.M., & de Neufville, R. (2017). Flexibility and real estate valuation under uncertainty. A practical guide for developers. Wiley Blackwell. Available at SSRN: https://ssrn.com/abstract=2998832.

Geltner, D.M., Francke, M.K., Shimizu, C., Fenwick, D., Baran, D. (2017). Commercial property price indicators: sources, methods and issues. Eurostat.

Geltner, D.M., MacGregor, B.D., Schwann, G.M. (2003). Appraisal smoothing and price discovery in real estate markets. *Urban Studies, 40*(5-6), 1047–1064.

Geltner, D.M., Miller, N.G., Clayton, J., Eichholtz, P.M.A. (2014). *Commercial real estate, analysis & investments* Vol. 4. Boston: Cengage Learning.

Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior: evidence from the housing market. *The Quarterly Journal of Economics, 116*(4), 1233–1260.

Goetzmann, W.N. (1992). The accuracy of real estate indices: repeat sale estimators. *Journal of Real Estate Finance and Economics, 5*, 5–53.

Guo, X., Zheng, S., Geltner, D.M., Liu, H. (2014). A new approach for constructing home price indices: the pseudo repeat sales model and its application in China. *Journal of Housing Economics, 25*, 20–38.

Harvey, A. (1989). *Forecasting structural time series models and the kalman filter*. Cambridge: Cambridge University Press.

Hegde, S.P., & McDermott, J.B. (2003). The liquidity effects of revisions to the S&P 500 index: an empirical analysis. *Journal of Financial Markets, 6*(3), 413–459.

Hoffman, M.D., & Gelman, A. (2014). The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593–1623.

Koehler, E., Brown, E., Haneuse, S.J.P.A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician, 63*(2), 155–162.

Link, W.A., & Eaton, M.J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3*(1), 112–115.

Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D. (2013). *The BUGS Book: A practical introduction to bayesian analysis*. Boca Raton: CRC Press.

McMillen, D.P., & Thorsnes, P. (2006). Housing renovations and the quantile repeat-sales price index. *Real Estate Economics, 34*(4), 567–584.

Plummer, M. (2003). JAGS: a Program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd internation workshop on distributed statistical computing, pp. 124.

Quan, D.C., & Quigley, J.M. (1991). Price formation and the appraisal function in real estate markets. *The Journal of Real Estate Finance and Economics, 4*(2), 127–146.

Schwann, G.M. (1998). A real estate price index for thin markets. *Journal of Real Estate Finance and Economics, 16*, 269–287.

Shiller, R.J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review, 71*, 421–436.

Shiller, R.J. (1993). *Macro Markets, Creating institutions for managing society’s largest economic risks*. Oxford: Oxford University Press.

Shrestha, M.L., & Marini, M. (2013). Quarterly GDP revisions in G-20 countries: evidence from the 2008 financial crisis. IMF Working Paper (13/60).

Vehtari, A., Gelman, A., Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1–20.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(Dec), 3571–3594.

Yu, K., & Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics & Probability Letters, 54*(4), 437–447.