J Real Estate Finan Econ (2010) 41:24–52
DOI 10.1007/s11146-009-9203-1

Repeat Sales Index for Thin Markets

A Structural Time Series Approach

Marc K. Francke

Published online: 11 August 2009

Abstract The repeat sales model is commonly used to construct reliable house price indices in absence of individual characteristics of the real estate. Several adaptations of the original model by Bailey et al. (J Am Stat Assoc 58:933–942, 1963) are proposed in literature. They all have in common using a dummy variable approach for measuring price indices. In order to reduce the impact of transaction price noise on the estimates of price indices, Goetzmann (J Real Estate Finance Econ 5:5–53, 1992) used a random walk with drift process for the log price levels instead of the dummy variable approach. The model that is proposed in this article can be interpreted as a generalization of the Goetzmann methodology. We replace the random walk with drift model by a structural time series model, in particular by a local linear trend model in which both the level and the drift parameter can vary over time. An additional variable—the reciprocal of the time between sales—is included in the repeat sales model to deal with the effect of the time between sales on the estimated returns. This approach is robust can be applied in thin markets where relatively few selling prices are available. Contrary to the dummy variable approach, the structural time series model enables prediction of the price level based

M. K. Francke (✉)
Finance Group, Faculty of Economics and Business,
University of Amsterdam Business School,
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: m.k.francke@uva.nl, marc.francke@ortec-finance.com

M. K. Francke
Ortec Finance Research Center, Orlyplein 145c,
1043 DV Amsterdam, The Netherlands

on preceding and subsequent information, implying that even for particular time periods where no observations are available an estimate of the price level can be provided. Conditional on the variance parameters, an estimate of the price level can be obtained by applying regression in the general linear model with a prior for the price level, generated by the local linear trend model. The variance parameters can be estimated by maximum likelihood. The model is applied to several subsets of selling prices in the Netherlands. Results are compared to standard repeat sales models, including the Goetzmann model.

Keywords House prices · Kalman filter · Signal-extraction · Smoothing · State-space models

Introduction

The value of the housing stock is a significant portion of the national wealth. The total value of the private real estate market in the Netherlands was approximately €1,239 billion in 2007, corresponding to 436% of the real disposable household income in 2007, see CPB (2009). For that reason, many organizations and individuals, such as financial institutions or house owners, are interested in house price movements. A frequently used method to model house price movements is the repeat sales approach. For example, the House Price Index of the Kadaster (the Dutch Land Registry) is an application of the repeat sales model for the Netherlands, see Jansen et al. (2008).

Individual house characteristics, like house size, plot size, age, etc., can be omitted from the repeat sales model. This is one of the main advantages of the repeat sales model in absence of individual house characteristics. On a negative side, only selling prices of houses sold more than once can be used in a repeat sales model: all single-sales are not used in the estimation. A more general problem that the sold houses are a non-random selection of the entire housing stock (sample selection bias) is addressed by Gatzlaff and Haurin (1997) and Hwang and Quigley (2004), based on a procedure proposed by Heckman (1979). In this article, however, we do not address the problem of sample selection bias.

Implicit assumption in the repeat sales model is that the house characteristics and their impact on house prices do not change over time. This assumption does not obviously hold true for the age of the house in different selling years. In the repeat sales model the effect of age is embedded into the model's estimates of the time effect, see Cannaday et al. (2005).

In this article we focus on the specification of the time effect in the repeat sales model by specifying it as a structural time series model. A structural time series model is a model in which the trend, seasonal and error terms, plus other relevant components, are modeled explicitly. In particular, we focus on the estimation of price trends in thin markets where the number of repeat sales is relatively low and, hence, the impact of transaction price noise on the estimation of the price trends is high.

Structural time series models have already been used in real estate applications. For example Schwann (1998) estimates a hedonic price index for a thin market, where the periodic returns follow a stationary autoregressive process. Francke and De Vos (2000) estimate hedonic price indices by a hierarchical trend model, in which different trends are simultaneously estimated for different market segments. Both models have the format (in logs): observed series = trend + regression effects + irregular, where some structure for the trend component is assumed. To our knowledge, structural time series models have not previously been used in order to estimate repeat sales price indices.

In the structural time series repeat sales model, the trend component is modeled explicitly using a local linear trend (LLT) model. The LLT model depends on two variance parameters, which are estimated by maximizing the appropriate likelihood function. Small values of these parameters result in smooth price indices. The approach of Goetzmann (1992) is a special case of the LLT repeat sales model. In this paper, we provide an alternative to his two-step estimation procedure.

In the next step, we examine the effect of the time between repeat sales on the estimation of the price level. Empirical evidence shows that large profits are made when the time between sales is relatively short, for example within first 6 months. This can be the result of "flipping houses", that is buying and selling houses for profit in a short period of time or house improvements (about the latter, see Goetzmann and Spiegel 1995). A simple solution would be to drop these sales out of the sample. In this article, we include a variable containing the reciprocal of the time between repeat sales, as an alternative solution to dropping sales within first 6 months.

This article is structured as follows. Section "Repeat Sales Models and Small Samples" discusses existing methods for reducing the impact of transaction price noise on repeat sales price indices in small samples. Section "A Local Linear Trend Repeat Sales Model" describes the local linear trend repeat sales model, which also includes a term for the time between repeat sales, as an alternative solution to the approach given by Goetzmann (1992). The same section provides some background on structural time series models, its relation to the non-parametric methods, and references to real estate applications. Section "Estimation" explains the estimation approach of the LLT repeat sales model. Section "Application" begins with a description of the Kadaster database, containing all selling prices of houses in the Netherlands in the period from January 1993 to May 2009. It continues with a comparison of price indices from the LLT repeat sales model to indices based on the frequently used models of Case and Shiller (1987) and Goetzmann (1992). Price indices are compared for different subsets, varying from all residential selling prices in the Netherlands to a small subset in a specific area code. Finally, the impact of revision in repeat sales price indices is examined for both models. Section "Conclusions" concludes.

Repeat Sales Models and Small Samples

General Specification

The repeat sales model is introduced by Bailey et al. (1963) and a number of adaptations of this original model is given in the real estate literature. A general specification is provided by (see also Kuo 1997)

$$y_{it} = \mu_i + (l_{it} - 1)\gamma_0 + \beta_t + \alpha_{it} + \varepsilon_{it}, \quad t = 1, \ldots, T, \tag{1}$$

$$\alpha_{i,t+1} = \rho\alpha_{it} + \eta_{it}, \quad i = 1, \ldots, M, \tag{2}$$

where y_{it} is the natural logarithm of the selling price of house i at time t (in months), μ_i is the ith house specific time invariant effect, and β_t is the log time index at time t. For identification purposes, it is assumed that $\beta_1 = 0$. T is the number of months and M is the number of houses. The total number of observations is $N = \sum_{i=1}^{M} n_i$, where n_i is the number of sales of house i, and $n_i > 1$. α_{it} is a house specific time trend, specified as a first order autoregressive process. A usual assumption is that α_{it} follows a random walk, implying that $\rho = 1$ in Eq. 2. The transaction price noise ε and the transition noise η are assumed to be independent, and are distributed as $\varepsilon \sim N(0, \sigma^2 I)$ and $\eta \sim N(0, q_\eta\sigma^2 I)$, respectively. l_{it} denotes the number of times house i is sold up to time t, and γ_0 is the time-independent return associated with each sale. Goetzmann and Spiegel (1995) argue that γ_0 in a repeat sales model captures fix-ups immediately after purchase. Shiller (1993, p. 139–140) argues that γ_0 should not be used in the presence of heterogeneity across space. He presents evidence that in the presence of heterogeneity across space γ_0 is too large and the slope of the index too small. The same type of reasoning is followed by Clapp and Giaccotto (1999), therefore they exclude all sales within one or two years.

It is a common practice to rewrite Eq. 1 in ‘first differences’, giving

$$y_{it} - y_{is} = \gamma_0 + (t_i - s_i)^{-1}\gamma_1 + \beta_t - \beta_s + e_{it} - e_{is}, \tag{3}$$

where $t_i > s_i$ and $e_{it} = \alpha_{it} + \varepsilon_{it}$, canceling out the M (fixed effect) levels μ_i.

Instead of simply excluding sales within a short time period or including a constant, an additional term $(t_i - s_i)^{-1}\gamma_1$ is introduced in Eq. 3, which makes the model more flexible. For s_i close to t_i the term $(t_i - s_i)^{-1}$ is large, and for s_i far from t_i the term $(t_i - s_i)^{-1}$ approaches zero. Section “Application” provides empirical evidence that the average periodic return is a decreasing function of time between sales, which can be approximated well by the functional form $(t_i - s_i)^{-1}\gamma_1$. Hence, large profits are made when this period is relatively short. Reasons for these high periodic returns are speculation and/or fix-ups. In the Netherlands there is an additional reason due to the transfer tax reduction for resales within 6 months, see Section “Application”. In our applications we explore several variants of Eq. 3, with and without γ_0 and γ_1.

Table 1 provides several versions of repeat sales models which are proposed in literature. All of them are special cases of the general model (2)–(3). For

Table 1 Restrictions on parameters in repeat sales models (2)–(3)

Model	γ_0	γ_1	ρ	σ_η	σ_ε	β_t
Bailey et al. (1963)	0	0	0	0		Fixed
Case and Shiller (1987, 1989)	0	0	1			Fixed
Hill et al. (1997)	0	0			0	Fixed
Webb (1988)	0	0	1		0	Fixed
Goetzmann and Spiegel (1995)	$\neq 0$	0	0			Fixed
Goetzmann (1992)	0	0	0			Random walk with drift

example, in the Case and Shiller (1987) model it is assumed that in Eq. 2 $\rho = 1$ and in Eq. 3 $\gamma_0 = \gamma_1 = 0$. The model reduces to $\alpha_{i,t+1} = \alpha_{it} + \eta_{it}$ and $y_{it} - y_{is} = \beta_t - \beta_s + \alpha_{it} + \varepsilon_{it} - \alpha_{is} - \varepsilon_{is}$. In the Bailey et al. (1963) model it is assumed that in Eq. 3 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = \sigma_\eta = 0$, so $\alpha_{i,t+1} = 0$. The model reduces to $y_{it} - y_{is} = \beta_t - \beta_s + \varepsilon_{it} - \varepsilon_{is}$. For comparison purposes, the repeat sales model which is used by Clapp and Giaccotto (1992a) does not fit within this framework. Instead of using the time-variant covariance structure α_{it}, they allow for time-invariant covariances $\mathrm{Cov}(\varepsilon_{is}, \varepsilon_{it}) = \rho\sigma^2$ for $s \neq t$.

Small Samples

In the repeat sales models, the specification of the time effect is simply a dummy variable approach with fixed parameters β_t. Conditional on μ_i and α_{it}, the estimate of β_t is the average selling price at time t. This means that the estimate of β_t does not depend on preceding and subsequent periods. However, the estimate of β_t is sensitive to transaction price noise, in particular in small samples when the number of transactions per period is low. This happens, for example, with local price indices, short time periods, and/or in case of severe outliers, when the transaction price differs from its true market value by a large amount. The resulting price indices may then become very volatile.

In order to reduce the impact of transaction price noise on the estimate of β_t, different methods have been proposed. A first group of methods consists of a two-step procedure. In the first step, the log price indices (or periodic returns) are estimated from a version of the repeat sales given by Eqs. 2–3. Let $\hat{\beta}_t$ denote the estimated log price indices. In the second step, these $\hat{\beta}_t$ estimates are inputted into a smoothing algorithm like, for example, a locally weighted regression. Examples in the literature are provided by Cleveland (1979) and Wand and Jones (1995, Chapter 5), who introduce a more general, local polynomial kernel estimators. In comparison, in order to construct local house price indices, Clapp (2004) apply local polynomial regression in a space-time model.

The main drawback of this two-step procedure is that it does not take into account the uncertainty in the estimates of β_t. Therefore, it disregards the precision of parameter $\hat{\beta}_t$, and the covariance matrix, $\mathrm{Cov}(\hat{\beta}_s, \hat{\beta}_t)$. The precision differs over time because the number of observations differs from one period to another, especially in small samples. Another concern is the behavior of the kernel at the boundaries, i.e. at the beginning and the end of the time period, because the kernel window at the boundaries is devoid of data.

A more recent approach in order to handle small datasets in repeat sales models is provided by Baroni et al. (2007). They propose a principal components analysis (PCA) factor repeat sales index, which exploits the relation between the house price indices and other economic and financial explanatory variables. One drawback of this approach is that the estimated price indices depend on the included set of explanatory variables.

A third way to manage a small number of observations is to replace dummy variables by a smooth (continuous) *deterministic* trend function, for example, a cubic spline. A slightly more flexible and equally easy to implement is the Fourier form approach, which depends on only few parameters. For more details, see McMillen and Dombrow (2001) and McMillen and McDonald (2004). This method has also been applied in a hedonic price model literature, see for example Thorsnes and Reifel (2007).

An early and successful signal-extraction approach is provided by Goetzmann (1992), who uses a *stochastic* trend specification. This is a Bayesian approach, based on the work by Lindley and Smith (1972), in which a prior distribution is specified for the periodic returns $\Delta\beta_{t+1} \equiv \beta_{t+1} - \beta_t$, given by $\Delta\beta_{t+1} \sim N(\kappa, \sigma_\zeta^2)$. This is equivalent to expressing β_{t+1} as a random walk with drift,

$$\beta_{t+1} = \beta_t + \kappa + \zeta_t, \quad \zeta \sim N\left(0, \sigma^2 q_\zeta\right), \tag{4}$$

where σ^2 is the variance of ε_{it} in Eq. 1 and q_ζ is the signal-to-noise ratio. The resulting estimates of the periodic returns and the price indices are less sensitive to transaction noise.

The structural time series approach which is applied in this article can be interpreted as a generalization of the Goetzmann (1992) approach. Firstly, structural time series models allow for a more general model specifications of the prior than the Goetzmann's approach. In the random walk with drift model specification, the a priori assumption is that the drift term (κ) is constant over time. However, in successive periods of appreciation and depreciation of the price levels, this assumption is not valid. A more appropriate specification would be to allow κ in Eq. 4 to change over time. An example of such a model is the local linear trend model, which is explored in more detail in Section "A Local Linear Trend Repeat Sales Model".

The second generalization concerns the estimation of the signal-to-noise ratio, given by q_ζ. In Goetzmann's approach, the variances σ^2 and $q_\zeta\sigma^2$ are estimated in an initial step, which sometimes leads to biased estimates of the variances. The resulting signal-to-noise ratio is plugged into the second step of the Bayesian procedure. However, as it is shown in Section "Estimation", it is possible to compute the concentrated loglikelihood and to estimate the signal-to-noise ratio parameters directly by maximization. This can be applied in the

Bayesian procedure as well as in the structural time series approach, avoiding the somewhat ad hoc two–step procedure.

In contrast to the dummy variable approach, the structural time series model enables the prediction of the price level based on preceding and subsequent information. This means that even for particular time periods where no observations are available, an estimate of the price level can be provided. The use of a structural time series model results in a more stable price index and (partly) reduces systematic downward revisions found in the repeat sales indices, see for example Clapp and Giaccotto (1999) and Clapham et al. (2006). Another advantage of these models is that price indices are provided even in continuous time models, avoiding the problem of temporal aggregation, for an example see Englund et al. (1999).

A Local Linear Trend Repeat Sales Model

Model Specification

In the repeat sales model it is typically assumed that the β_t's are fixed unknown parameters. In this article, it is assumed that β_t is a scalar stochastic trend process in the form of a local linear trend model, in which both the level and slope can vary over time. The local linear trend model is given by

$$\beta_{t+1} = \beta_t + \kappa_t + \zeta_t, \qquad \zeta_{it} \sim N\left(0, q_\zeta \sigma^2\right), \tag{5}$$

$$\kappa_{t+1} = \kappa_t + \xi_t, \qquad \varepsilon_{it} \sim N\left(0, q_\xi \sigma^2\right). \tag{6}$$

The local linear trend model includes several specific models. If $\zeta_t = \xi_t = 0$, then $\kappa_{t+1} = \kappa_t = \kappa$ and $\beta_{t+1} = \beta_t + \kappa = t\kappa$ (when $\beta_0 = 0$), hence the trend is exactly linear. If $q_\xi = 0$, then the local linear trend model reduces to a random walk with drift, $\beta_{t+1} = \beta_t + \kappa + \zeta_t$, equivalent to the prior proposed by Goetzmann (1992). If we further assume that $\kappa = 0$, the stochastic trend is simply a random walk $\beta_{t+1} = \beta_t + \zeta_t$. On the other hand, if $q_\zeta \sigma^2 \to \infty$, no time structure is imposed, and the β_t's can be regarded as fixed unknown parameters, similar to the standard repeat sales model. The signal-to-noise ratios (q_ζ and q_ξ) are estimated by maximum likelihood, see Section "Estimation".

The local linear trend repeat sales model is provided by Eqs. 2, 3, 5 and 6. The initial value of κ is an unknown parameter, say κ_1. Similar to the standard repeat sales model, we assume for the purpose of identification that $\beta_1 = 0$.

In order to interpret β_t as the common trend we have to imply that the sum of the individual house trends is zero, i.e. $\sum_{i=1}^{M} \alpha_{it} = 0$ for $t = 1, \ldots, T$. A simpler, equivalent approach is to define the common trend d_t as the sum of the common trend β_t and the average individual house trend α_{it}, such that $d_t = \beta_t + \frac{1}{M} \sum_{i=1}^{M} \alpha_{it}$. If $\sigma_d^2 = \infty$, then $\sum_{i=1}^{M} \alpha_{it} = 0$ for all t. Note that for the standard repeat sales model, for which $q_\xi = \infty$ or $q_\zeta = \infty$, it holds that $\sum_{i=1}^{M} \alpha_{it} = 0$. In practice the term $M^{-1} \sum_{i=1}^{M} \alpha_{it}$ is negligible.

Structural Time Series Models

The local linear trend repeat sales model is an example of a structural time series model, in which the trend, seasonal and error terms, plus other relevant components, are all modeled explicitly. This is in contrast with the Box-Jenkins approach where trend and seasonal effect components are removed by differencing prior to the detailed analysis. The basic univariate structural time series model is

$$y_t = \beta_t + \delta_t + \varepsilon_t,$$

where β_t is a slowly varying component called the trend, δ_t is a periodic component or fixed period called the seasonal, and ε_t is an irregular component called error or disturbance. In this article we do not include the seasonal component and we focus on the specification of the trend component instead. For a detailed description, we refer the reader to Harvey (1989), West and Harrisson (1997), and Durbin and Koopman (2001), who discuss these models as examples of the more general class of *state-space* or *dynamic linear* models.

In the state-space form, the unobserved components can be estimated by the *Kalman filter* algorithm. The Kalman filter also produces the likelihood function, which enables the estimation of the variance parameters. The Ox package SsfPack contains ready-to-use estimation procedures for estimating the state-space models. It can be downloaded for free in order to be used for academic research and teaching purposes, see Koopman et al. (1999). In this article for the reason of experience and practice, the statistical program Gauss is used for the estimation of the local linear trend repeat sales model.

Harvey and Koopman (1999) and Koopman and Harvey (2003) examine the weighting patterns for signal-extraction implied by unobserved components models. These weighting patterns may be compared to the kernels used in non-parametric time series trend estimation. The signal-to-noise ratio in structural time series models has the same role as the bandwidth in a non-parametric approach. Harvey and Koopman summarize the following advantages of using the structural time series models in comparison to non-parametric methods:

- Different models can be compared by likelihood based criteria;
- Inference about the parameters, including the signal-to-noise ratio, can be based on the likelihood;
- Appropriate weights are implicitly provided. The weights depend on the position of the observations (begin, middle, or end of series) and magnitude of outlying observations. However, they are not necessarily symmetric.
- Root mean square errors can be computed for the estimated trend;
- The models can be made robust to outliers by specifying t-distributions;
- By formulating a model in continuous time, the optimal weighting for irregularly spaced observations is automatically carried out.

Figure 1 shows a simple example of weighting functions for a local linear trend model. The dependent variable y_t is the average of log selling prices per

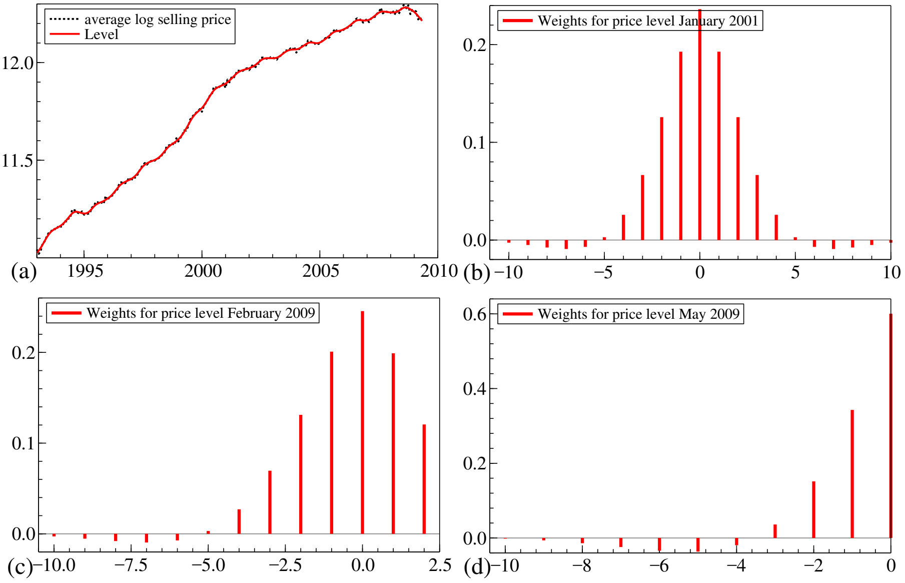

Fig. 1 Weighting functions for a local linear trend model (**a**–**d**)

month from January 2003 to May 2009. A description of the data is provided in Section "Data". The model is given by $y_t = \beta_t + \varepsilon_t$ and Eqs. 5–6. For simplicity it is assumed that ε_t has constant variance σ^2. The model is formulated in state-space form and estimated by the Kalman filter. Estimation results and weighting functions are generated using the Structural Time Series Analyser, Modeller and Predictor (STAMP) software; see Koopman et al. (2007). Panel (a) in Fig. 1 shows the series of averages and the estimated level β_t. In panel (b)–(d) some examples of weighting functions are given, depending on the estimated values of the signal-to-noise ratios q_ζ and q_ξ, and the position of the observations in the series. Panel (b) provides an almost symmetric weighting function, for the level in January 2001, in the middle of the series. Panel (c) and (d) give asymmetric weighting functions for the levels at the end of the series, February and May 2009 respectively. Note that for the estimation of the levels β_t it is not necessary to calculate the weighting functions; rather they can be derived from the output of the Kalman filter.

Real Estate Applications of Structural Time Series Models

Structural time series models, or more generally state space models, have already been used in real estate applications. Schwann (1998) estimates a hedonic price index for a thin market using a Kalman filter, where the periodic returns $\Delta\beta_t$ follow a stationary autoregressive process. Francke and De Vos (2000) estimate a hierarchical trend model, in which different trends are simultaneously estimated for different market segments. The trend specification is decomposed into a common trend, a region specific trend, and a house type specific trend. The region specific trend and the house type specific trend are modeled in deviation from the common trend. These models are efficiently estimated combining ordinary least squares and the Kalman filter, see also Francke and Vos (2004) and Francke (2008). Schulz and Werwartz (2004) provide a state-space model for house prices in Berlin. They include explanatory variables like inflation rates, mortgage rates, and building permissions in order to model the common price movement. Hannonen (2005, 2008) uses a structural time series model to analyze and predict urban land prices. To our knowledge, state-space models and the Kalman filter have not previously been used in order to estimate repeat sales price indices.

Estimation

A structural time series model can be put into a state-space format and efficiently estimated by the Kalman filter, see for example Durbin and Koopman (2001). In the local linear trend repeat sales model, the size of the state vector, which is the number of unknown parameters apart from the variances, becomes very large and is equal to $M + T + 2$, where M is the number of houses and T the number of time periods (for γ_0 and γ_1). In the application provided in the next section the number of houses is approximately 500,000. Including all these variables in the state vector is not feasible, as it would require storage and inversion of 500,000 × 500,000 matrices.

As shown in the previous section an alternative to the repeat sales model Eq. 1 is the specification in 'first differences' (3), canceling out the M levels μ_i. Unfortunately, model (2), (3), (5), and (6) cannot (easily) be put into the state-space format, because the data depend on the difference of the state vector in two moments in time, with varying time spans. The state-space approach assumes that the state vector is a Markov chain. Therefore we have to rely on another estimation procedure.

One option is to use the Expectation Maximization (EM) algorithm for the model in levels as given in Eq. 1. Conditional on μ_i, the Kalman filter can straightforwardly be applied to estimate the log price index β_t and other parameters. In an additional step, the parameters μ_i can be estimated by means of the EM algorithm. This results in a recursive estimation procedure, where it is guaranteed that the algorithm converges to at least a local optimum, see Dempster et al. (1977). The EM algorithm is proposed by Shumway and Stoffer (1982) and Watson and Engle (1983). The main advantage of this approach is that more general time specifications including, for example, more complex trend specifications and seasonal components, can easily be dealt with.

In this article a different approach for the estimation of the local linear trend repeat sales model is put forward. The local linear trend repeat sales model 'in differences' is estimated by an empirical Bayesian procedure. The model can then be expressed as a linear regression model with a prior for β, induced by the local linear trend model (5)–(6). Conditional on the parameters ρ, q_η, q_ζ, and q_ξ, the posteriors for β and σ easily follow. Estimates of ρ, q_η, q_ζ, and q_ξ are obtained by maximizing the likelihood of the 'differenced' data.

The repeat sales model (3) can be written as

$$\widetilde{y}_i = \mathbf{i}\gamma_0 + \mathbf{p}_i^{-1}\gamma_1 + \mathbf{p}_i\kappa_1 + \widetilde{X}_i\boldsymbol{\beta} + \widetilde{e}_i = \widetilde{Z}_i\delta + \widetilde{e}_i, \tag{7}$$

$$\widetilde{e}_i = \widetilde{\alpha}_i + \widetilde{\varepsilon}_i \sim N\left(0, \sigma^2\widetilde{\Omega}_i(\theta_1)\right), \tag{8}$$

for $i = 1, \ldots, M$, where $\widetilde{y}_i$ is a $n_i - 1$ vector of 'differenced' log selling prices, $\mathbf{i}$ is a vector of ones, $\mathbf{p}_i$ is a vector of the differences between the selling dates of repeat sales of the same house i, with typical elements $(t_i - s_i)$ for $t > s$, $\widetilde{X}_i$ is a $(n_i - 1) \times (T - 1)$ matrix containing elements 0, -1 and 1, and $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_T)'$. $n_i - 1$ is the number of transaction pairs of house i. In most cases one only pair of repeat sales is available, hence $n_i - 1$ equals 1. Note that β_t is slightly redefined in the sense that in Eq. 6 it is now assumed that $\kappa_1 = 0$, leading by repeated substitution in Eqs. 5–6 to the term $(t_i - s_i)\kappa_1$ in Eq. 7.

The $(n_i-1) \times (n_i-1)$ covariance matrix $\widetilde{\Omega}_i(\theta_1)$ depends on the unknown parameters $\theta_1 = (\rho, q_\eta)'$. If it is assumed that $\rho = 1$, the covariance matrix $\widetilde{\Omega}_i(\theta_1)$ has a typical form

$$\sigma^2\widetilde{\Omega}_i(\theta_1) = \mathrm{Var}(\widetilde{\varepsilon}_i) + \mathrm{Var}(\widetilde{\alpha}_i) = \sigma^2\left(\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} + q_\eta\begin{bmatrix} t_i - s_i & 0 & 0 \\ 0 & s_i - \tau_i & 0 \\ 0 & 0 & \tau_i - \varsigma_i \end{bmatrix}\right), \tag{9}$$

where $t_i > s_i > \tau_i > \varsigma_i$. Appendix A provides an expression for $\mathrm{Var}(\widetilde{\alpha}_i)$ when $|\rho| < 1$. Note that $\mathrm{Cov}(\widetilde{e}_i, \widetilde{e}_j) = 0$ for $i \neq j$.

The prior distribution for $\boldsymbol{\beta}$ comes from the local linear trend model. It follows from (5)–(6) that the prior for $\Delta\boldsymbol{\beta}$ is given by

$$\Delta\boldsymbol{\beta} \sim N\left(0, \sigma^2 q_\zeta I + \sigma^2 q_\xi \begin{pmatrix} 0 & 0 \\ 0 & A_1 \end{pmatrix}\right), \tag{10}$$

where A_1 is defined below. Observing that $\Delta\beta_2 + \ldots + \Delta\beta_t = \beta_t - \beta_1 = \beta_t$ when $\beta_1 = 0$, the prior for $\boldsymbol{\beta}$ easily follows and is given by

$$\boldsymbol{\beta} \sim N\left(0, \sigma^2\Sigma(\theta_2)\right), \tag{11}$$

$$\Sigma(\theta_2) = A_2\left(q_\zeta I + \sigma^2 q_\xi \begin{pmatrix} 0 & 0 \\ 0 & A_1 \end{pmatrix}\right)A_2', \tag{12}$$

where

$$A_1 = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & T-2 \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{13}$$

No prior information for the parameters γ_0, γ_1 and κ_1 is available, hence the precision matrix, which is the inverse of the variance matrix Ψ for the regression parameters $\delta = (\gamma_0, \gamma_1, \kappa_1, \boldsymbol{\beta}')'$, is given by $\Psi^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma^{-1} \end{bmatrix}$.

The posterior of δ is provided by

$$\delta | \widetilde{y}, \widetilde{Z}, \theta, \sigma^2 \sim N\left(\delta^*, \sigma^2 \mathrm{V}(\delta^*)\right), \tag{14}$$

$$\mathrm{V}(\delta^*) = \left(\sum_{i=1}^{M} \widetilde{Z}_i' \widetilde{\Omega}_i^{-1} \widetilde{Z}_i + \Psi^{-1} \right)^{-1}, \tag{15}$$

$$\delta^* = \mathrm{V}(\delta^*) \sum_{i=1}^{M} \widetilde{Z}_i' \widetilde{\Omega}_i^{-1} \widetilde{y}_i, \tag{16}$$

where $\theta = (\theta_1', \theta_2')'$.

Conditional on θ, the variance parameter σ^2 can be estimated analytically. σ^2 can be concentrated out of the marginal likelihood function, leading to

$$\begin{aligned} \sigma^2_{\mathrm{ML}} &= \frac{\mathrm{RSS}}{m} \\ &= \frac{\sum_{i=1}^{M} \widetilde{y}_i' \widetilde{\Omega}_i^{-1} \widetilde{y}_i - \sum_{i=1}^{M} \widetilde{y}_i' \widetilde{\Omega}_i^{-1} \widetilde{Z}_i \left(\sum_{i=1}^{M} \widetilde{Z}_i' \widetilde{\Omega}_i^{-1} \widetilde{Z}_i + \Psi^{-1} \right)^{-1} \sum_{i=1}^{M} \widetilde{Z}_i' \widetilde{\Omega}_i^{-1} \widetilde{y}_i}{N - M - T - 2}. \end{aligned} \tag{17}$$

The remaining parameters θ can be estimated by maximizing the concentrated marginal likelihood function (with respect to σ^2 and δ), given by

$$\begin{aligned} -2 \log L_c^M &= m \left(\log 2\pi + \log \sigma^2_{\mathrm{ML}} + 1 \right) \\ &\quad + \sum_{i=1}^{M} \log \left| \widetilde{\Omega}_i \right| + \log \left| \sum_{i=1}^{M} \widetilde{Z}_i' \widetilde{\Omega}_i^{-1} \widetilde{Z}_i + \Psi^{-1} \right| + \log |\Sigma|. \end{aligned} \tag{18}$$

The estimation procedure can be summarized as follows:

1. Conditional on θ, an estimate of δ and σ^2 is provided by Eqs. 14–17. The terms $\sum_{i=1}^{M} \widetilde{y}_i' \widetilde{\Omega}_i^{-1} \widetilde{y}_i$, $\sum_{i=1}^{M} \widetilde{y}_i' \widetilde{\Omega}_i^{-1} \widetilde{Z}_i$, $\sum_{i=1}^{M} \widetilde{Z}_i' \widetilde{\Omega}_i^{-1} \widetilde{Z}_i$, and $\sum_{i=1}^{M} \log |\widetilde{\Omega}_i|$ can be computed per house observation. The precision matrix Ψ^{-1} follows from Eq. 12.
2. The parameters θ can be estimated by maximizing the likelihood function (18). All terms in the likelihood function are available from step 1.
3. Finally, the log price index and log return are given by $(t-1)\kappa_1^* + \beta_t^*$ and $\kappa_1^* + \beta_t^* - \beta_{t-1}^*$ respectively. The corresponding variances (and covariances) can be computed from Eqs. 14–15 straightforwardly. The price indices and returns are obtained by taking the antilog, and have a lognormal distribution.

Note that for $q_\zeta = q_\xi = \infty$ in Eq. 12, the precision matrix $\Psi = 0$, hence it assumes no prior information. Therefore, conditional on θ_1, the estimation results coincide with standard repeat sales models. When $q_\xi = 0$ in Eq. 12, conditional on θ_1 and q_ζ, the estimation results are equivalent to the approach of Goetzmann (1992).

The main difference between the Goetzmann's approach and the local linear trend repeat sales model approach is the estimation of the parameters θ. For example, in Goetzmann's approach, σ^2 and $q_\zeta \sigma^2$ are estimated in a two–step procedure. In the local linear trend repeat sales model, they are estimated in one step by maximum likelihood.

The slope parameters κ_t can be estimated in a similar fashion as the trend parameters β_t. The computation requires submatrices already computed in step 1. More details can be found in Appendix B.

Application

Data

The Kadaster (Dutch Land Registry Office) is responsible for the registration of real estate properties. The database covers all transactions within the Netherlands. The number of selling prices of owner-occupied houses in the period from January 1993 until May 2009 is more than 3.5 million. In the following cases, the transactions are not used for the calculation of the price index:

- sales between relatives;
- transactions where the buyer is a legal entity;
- if the same lot is sold more than once in one transaction;
- no full ownership or long lease;
- more than one purchase price in one transaction;
- unlikely purchase price.

The remaining number of transactions of houses sold more than once is approximately 1.5 million, covering 644 thousand different houses, which is 17% of the owner-occupied housing stock. The total number of houses by the end of 2007 is slightly above 7 million, and 53.3% of the housing stock is owner-occupied.

Some characteristics of these transactions are provided in Table 2. In Table 3 the number of observations of houses sold more than once are given, and in Table 4 the number of observations per house type are provided.

Figure 2 shows the frequency of the time between repeat sales in the Netherlands. The mode, the median, and the mean are 36, 52, and 58 months, respectively. A sharp decline of the frequency after 6 months can be explained by the fact that there is a large reduction in transfer tax if a house is re-sold within first 6 months. In this case the transfer tax of 6% is only applied on the (positive) difference between the second and the first transaction price, instead of applying it on the transaction price itself. For example, when the house is

Table 2 Overview of transactions in the period January 1993–May 2009

Description	
Total number of selling prices	3,481,390
Total number of selling prices after screening	3,188,622
Number of selling prices (at least two sales of the same house)	1,536,407
Number of different houses	643,904
Municipalities	443
Zip code (4 digits)	3,804

first sold for €250,000 and within 6 months for €290,000, the first buyer has to pay 6% of €250,000 = €15,000 in transfer tax, while the second buyer has to pay only 6% of €40,000 = €2,400 in transfer tax. This tax system can have considerable impact on selling prices. Note that a substantial part of the repeat sales, 4%, are within 6 months.

The database is used by the Kadaster to construct a monthly weighted repeat sales index, based on the method by Case and Shiller (1987). Indices are provided on a national level as well as on regional and house type levels. More details on the index construction method and the database can be found in Jansen et al. (2008).

In 2008 the weighted repeat sales is replaced by a monthly Sales Price Appraisal Ratio (SPAR) index. The index is published by the CBS (Statistics Netherlands) in cooperation with the Kadaster. The appraisal value is the WOZ-value (Waardering Onroerende Zaken), a yearly assessed value used for property tax. The WOZ law requires that the determined appraisal value is also used for other legal purposes, such as for the levy which the water boards can raise, and income taxes levied by the central government. The SPAR method has been applied in New Zealand since the early 1960s, see Bourassa et al. (2006). A more general treatment of assessed value price indices methods is provided by Clapp and Giaccotto (1992a, b). De Vries et al. (2007) provides a detailed description of the application of the SPAR method for the Netherlands.

Table 3 Number of sales

Number of sales	Number of observations
2	455,503
3	140,701
4	37,428
5	8,393
6	1,584
7	251
8	38
9	4
10	2

Table 4 House types

House type	Number of observations
Apartments	511,943
Terraced houses and corner houses	737,754
Semi-detached houses	151,347
Detached houses	135,363

Comparison of Indices

In this section the results using four different price index methods are compared. Price indices are constructed using

1. the Case and Shiller (1987) (CS) repeat sales model;
2. the Goetzmann (1992) repeat sales model;
3. the random walk with drift (RWD) repeat sales model;
4. the local linear trend (LLT) repeat sales model.

The only difference between the Goetzmann and the RWD repeat sales model concerns the estimation of the variance parameters σ^2 and $\sigma^2 q_\zeta$ in Eqs. 5 and 7. In the Goetzmann model, these variances are estimated in an initial step using the CaseShiller repeat sales model (the two-stage Bayesian variant). In the RWD repeat sales model, the variances are estimated by maximization of the likelihood function (18); the signal-to-noise ratio q_ξ in Eq. 6 is equal to zero for the RWD model. Similarly to the CS and the Goetzmann models, it is assumed that in the RWD and the LLT models in Eq. 2 $\rho = 1$ and in Eq. 7 $\gamma_0 = \gamma_1 = 0$.

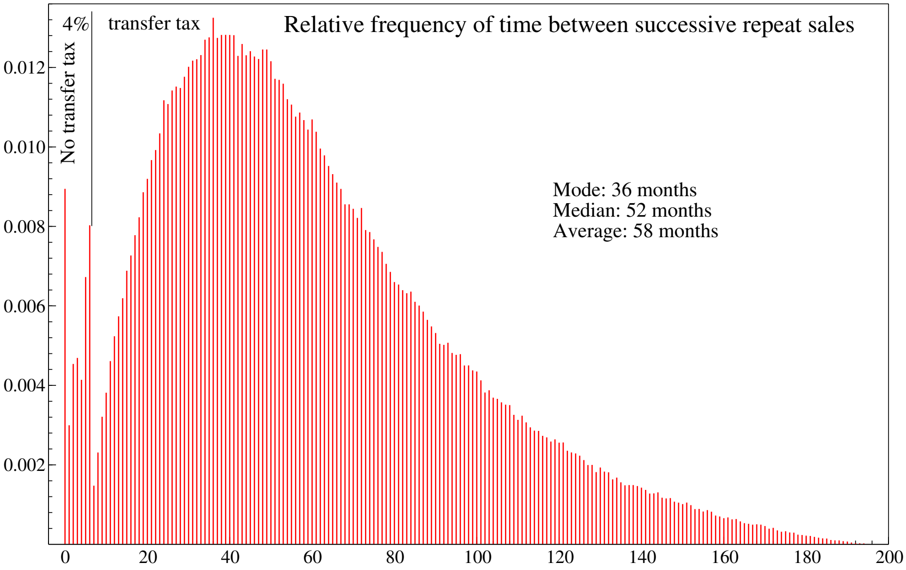

Fig. 2 Relative frequency of time between repeat sales

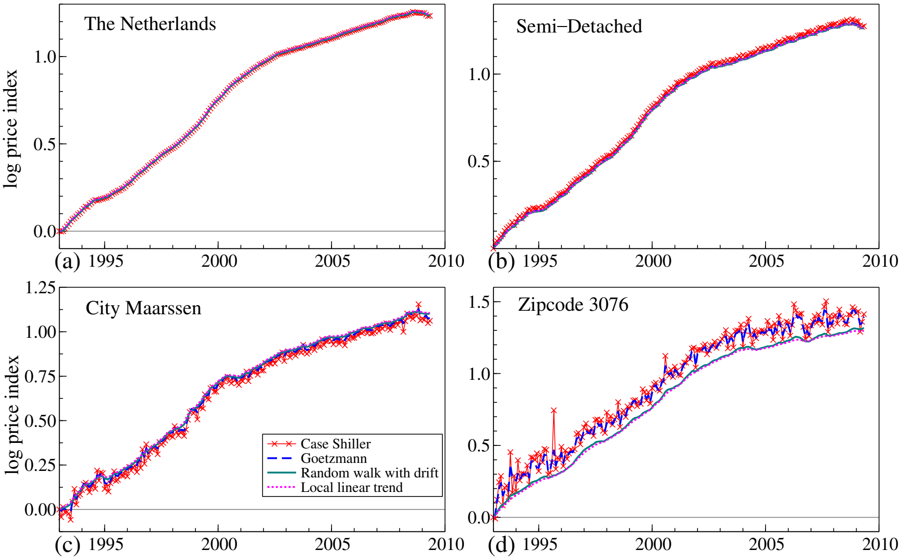

Fig. 3 Log price indices $\hat{\beta}_t$ for different subsets. The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 7 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = 1$. The remaining models are special cases of the local linear trend repeat sales model

The methods are compared using four different data sets, with varying number of observations:

a. all selling prices in the Netherlands (846,439 observations);
b. the selling prices of semi-detached houses (70,471 observations);
c. the selling prices of a small city of Maarssen (2,234 observations);
d. the selling prices in a specific area code (991 observations).

The time period ranges from January 1993 to May 2009. Price indices are constructed on a monthly basis. The average number of observations per month for the four data sets are 4297, 358, 11, and 5, respectively. Sales within first 6 months are excluded.

Table 5 gives the estimation results for different data sets and models. Figure 3 shows the log price level for the whole sample. In Fig. 3a the price indices for the Netherlands, using the four models, almost coincide: the LLT index is somewhat smoother than the other indices. This can also be seen from the standard deviation of the estimated monthly returns $\Delta\hat{\beta}_t$: it is 0.0047, 0.0043, 0.0043, and 0.0039 for the CS, Goetzmann, RWD, and LLT models, respectively.

In comparison, in Fig. 3d, the differences between the zip code area indices are substantial. The LLT price index is smooth, and is virtually the same as the RWD price index, while the CS index is very volatile, due to the small number of observations. The standard deviations of $\Delta\hat{\beta}_t$ for the CS, Goetzmann, RWD, and LLT models are 0.0895, 0.0429, 0.0069, and 0.0043, respectively.

For the semi-detached houses the differences between the indices are not substantial, but the CS price index is slightly irregular, see Fig. 3b. As can be seen from Fig. 3c, the city level price indices show the same pattern as the zip code indices, however they are less extreme.

Note that in Fig. 3c, the RWD and LLT price index are above the CS and Goetzmann index, and in Fig. 3d it is vice versa. This results from the fact that the CS price index is much more sensitive to outliers than the RWD and LLT index, particularly at the begin of the period, where the log price index value

Table 5 Estimation results from repeat sales models for different subsets

	Case Shiller	Goetzmann	Random walk with drift	Local linear trend
The Netherlands				
$12\kappa_1$		0.076 (18.95)	0.076 (18.47)	0.067 (4.08)
σ	0.075	0.075	0.075	0.075
$\sqrt{q_\eta}\sigma$	0.015	0.015	0.015	0.015
$\sqrt{q_\zeta}\sigma$		0.005	0.005	0.000
$\sqrt{q_\xi}\sigma$				0.001
st. dev. $\Delta\hat{\beta}_t$	0.0047	0.0043	0.0043	0.0039
Loglikelihood	389,010.0	389,766.3	389,766.3	389,854.7
$N-M$	846,439			
Semi-Detached				
$12\kappa_1$		0.077 (9.73)	0.077 (12.98)	0.166 (7.07)
σ	0.076	0.076	0.076	0.076
$\sqrt{q_\eta}\sigma$	0.015	0.015	0.015	0.015
$\sqrt{q_\zeta}\sigma$		0.009	0.007	0.002
$\sqrt{q_\xi}\sigma$				0.001
st. dev. $\Delta\hat{\beta}_t$	0.0092	0.0057	0.0050	0.0041
Loglikelihood		33,150.0	33,156.8	33,190.5
$N-M$	80,162			
Maarssen				
$12\kappa_1$		0.066 (2.22)	0.068 (7.10)	0.094 (3.96)
σ	0.045	0.045	0.044	0.044
$\sqrt{q_\eta}\sigma$	0.011	0.011	0.011	0.011
$\sqrt{q_\zeta}\sigma$		0.035	0.011	0.009
$\sqrt{q_\xi}\sigma$				0.000
st. dev. $\Delta\hat{\beta}_t$	0.0345	0.0164	0.0054	0.0046
Loglikelihood	1,644.4	2,024.2	2,065.6	2,068.8
$N-M$	2,511			
Zip code 3076				
$12\kappa_1$		0.085 (1.11)	0.081 (6.01)	0.126 (4.21)
σ	0.063	0.063	0.071	0.072
$\sqrt{q_\eta}\sigma$	0.014	0.014	0.013	0.013
$\sqrt{q_\zeta}\sigma$		0.090	0.015	0.009
$\sqrt{q_\xi}\sigma$				0.001
st. dev. $\Delta\hat{\beta}_t$	0.0895	0.0429	0.0069	0.0043
Loglikelihood	280.4	484.9	546.5	549.6
$N-M$	991			

The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 7 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = 1$. The remaining models are special cases of the local linear trend repeat sales model. t-values are provided between the brackets is assumed to be zero. This also holds for the Goetzmann index, although to a lesser extent. In the CS repeat sales model the outliers are absorbed by the initial price levels.

Other results provided by Table 5 are as follows. The standard deviation σ is approximately 7.5%, except for the city Maarssen where the standard deviation is only 4.4%. Note that this is a standard deviation of an individual house transaction, hence it should not to be larger for smaller numbers of observations. The standard deviations for the individual house random walks, $\sqrt{q_\eta}\sigma$, approximately 1.5%, are relatively constant over the models and analysed samples.

The estimated values of $\sqrt{q_\zeta}\sigma$ are identical for the Goetzmann and RWD models in the national Dutch data set. For the zip code area data, the estimates are very different: 0.090 (Goetzmann) versus 0.015 (RWD). This is also reflected in the standard deviations of $\Delta\hat{\beta}_t$: 0.0429 (Goetzmann) versus 0.0069 (RWD). Note that for the RWD and LLT models, the standard deviation of $\Delta\hat{\beta}_t$ is not very sensitive to the sample size, whereas for the CS and Goetzmann model the standard deviation decreases with the number of observations: in the CS (Goetzmann) model it varies between 0.0047 (0.0043) for the Netherlands and 0.0895 (0.0429) for the area code level. The high standard deviations imply that the CS and Goetzmann model cannot be used to construct detailed price indices and returns. The monthly standard deviations in the CS and Goetzmann model are respectively more than 13 and 6 times as large as the average monthly returns, where the average yearly return is in the order of 0.08. For the RWD and the LLT models, these figures are more reasonable: the monthly standard deviation at the area code level is respectively 1.0 and 0.6 times the average monthly return.

In all samples, the local linear trend model has a higher loglikelihood at the cost of only one additional parameter q_ξ, as compared to the Goetzmann and the RWD models. The loglikelihood for the Goetzmann and the RWD models are identical for the national Dutch data set. At the city and area code levels, the RWD model has a substantially higher loglikelihood than the Goetzmann model: the two-step procedure results in suboptimal estimates of σ and q_η. The suboptimal estimates result in more volatile log price indices, in comparison to the RWD price indices, for which the maximum likelihood estimates of the parameters have been used.

Figure 4 focuses on the log price levels from January 2008 to May 2009 in order to track turning points. A turning point can be defined as a change from a positive (negative) to a negative (positive) value of $\Delta\hat{\beta}_t$. From Fig. 4a it can be concluded that prices in the Netherlands began to decline from August 2008: the price decrease starting from August 2008 until May 2009 is only slightly more than 2% points. The local linear trend repeat price index is lagging one month compared to other indices. For the semi-detached houses in Fig. 4b, the picture is less clear. All indices, except for the LLT index, have a dip in September 2008, followed by a price increases in October and November. The LLT index reports a fall in prices from October 2008, hence it is leading the other indices by one month. For the more detailed price indices, the differences

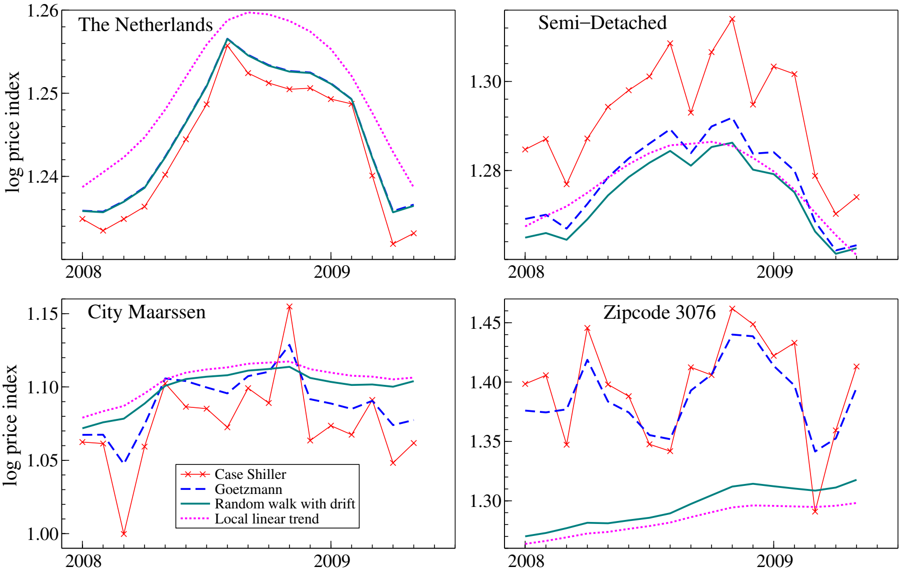

Fig. 4 Log price indices $\hat{\beta}_t$ from 2008 for different subsets. The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 7 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = 1$. The remaining models are special cases of the local linear trend repeat sales model

between the indices are substantial. However, it is difficult to conclude from these small samples that prices are falling.

A nice feature of the LLT repeat sales model is that estimates of the slope parameter κ_t, $t = 1 \ldots, T - 1$ are available. The estimation procedure is closely related to the estimation of β_t and is given in Appendix B. In the Goetzmann and the RWD models, the drift parameter (κ_1) is assumed to be constant over time. The annualized estimates $\kappa_1 \times 12$ are also provided in Table 5. Figure 5 provides the annualized slope parameter estimates $\kappa_t \times 12$ and the corresponding 95% confidence bounds for the LLT model. It can be concluded that for the Netherlands (Fig. 5a) and the semi-detached houses (Fig. 5b) the assumption of a constant drift, similar to the Goetzmann's approach, is not valid: for example, the difference in the slope parameter in 1994 and 1999 is significant. For the more detailed indices in Fig. 5c and 5d, the assumption of constant drift cannot be rejected.

The estimates of the slope parameters κ_t can also be used for tracking turning points. For the LLT model, a turning point can alternatively be defined as a change from a positive (negative) slope parameter to a negative (positive) value. Following this definition, the turning points for the Netherlands and the semi-detached houses are September 2008.

Above examples suggest that in case of many observations the log price level estimates coincide for the four methods. When only a few observations per time period are available, the CS price index is extremely volatile and sensitive to transaction price noise, while the LLT price index remains stable. In general,

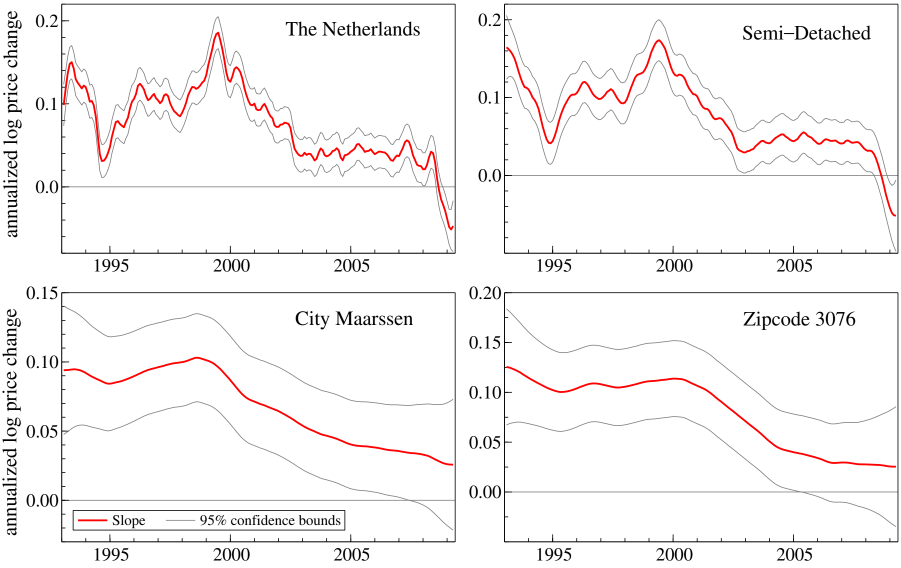

Fig. 5 Slope $\hat{\kappa}_t$ for semi-detached houses in the local linear trend repeat sales model. The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 7 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = 1$. Estimates of the slope parameters κ_t are provided by Eqs. 24–26

the results from the RWD model are close to the LLT model. However, the LLT model has a better model fit, as measured by the loglikelihood. In case of many observations, the Goetzmann and the RWD approaches produce the same results. When the number of observations is small, the two-step Goetzmann estimation procedure leads to suboptimal estimates of the the variances σ^2 and $q_\zeta\sigma^2$. The resulting indices are more volatile than the RWD estimates. This is in line with the findings of Goetzmann (1992), who states that the two-stage Bayes procedure leads to an overestimation of q_ζ.

Time Between Repeat Sales

In this subsection, the impact of the time between repeat sales on price indices is examined, using the LLT repeat sales model. Figure 6 shows the averages of the residuals per time between sales $(t - s)$, measured in months. The residuals $e_{it} - e_{is}$ are calculated from Eqs. 7–8 for the LLT model. The model is estimated on the repeat sales for the Netherlands, excluding the sales within one month. The difference between the minimum and maximum value is approximately 0.20. Note that for large values of $(t - s)$, only a few number of observations is available (see Fig. 2). From Fig. 6, it can be concluded that the between sales averages of the residuals per time can be approximated reasonably well by the function $(t - s)^{-1}\gamma_1$.

Two different semi-detached houses datasets are used in order to examine the inclusion of the non-temporal component γ_0 and the time between sales

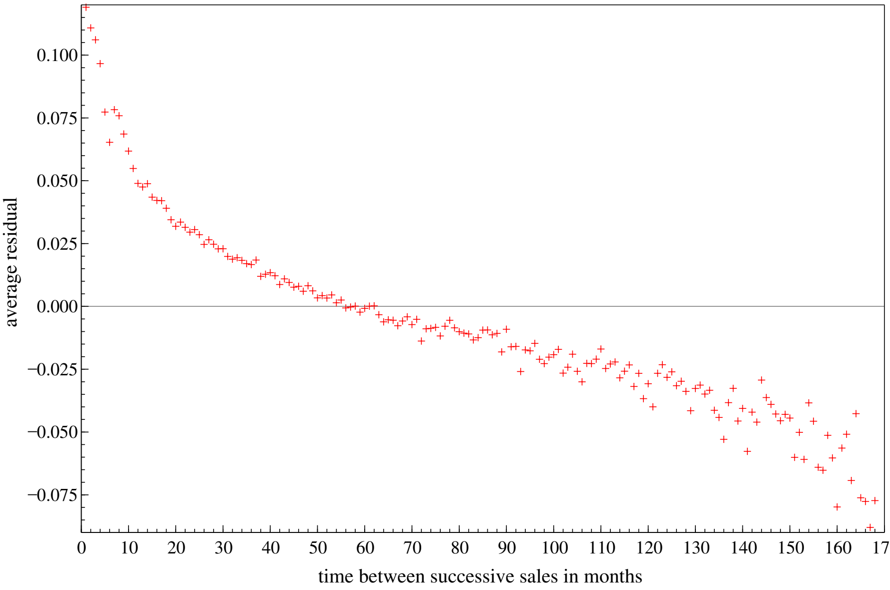

Fig. 6 Averages of residuals per time between sales in the local linear trend repeat sales model for the Netherlands. The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 7 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = 1$

component $(t - s)^{-1}\gamma_1$. In the first dataset, all sales within six months are excluded, while in the second dataset, only the sales within one month are excluded, resulting in 1,843 additional observations.

Based on the first dataset, two different LLT models are estimated: (a) without an intercept ($\gamma_0 = 0$) and (b) with an intercept ($\gamma_0 \neq 0$). In both models the time between sales component is absent ($\gamma_1 = 0$). Based on the second dataset, another two LLT models are estimated: (c) without an intercept ($\gamma_0 = 0$), and (d) with an intercept ($\gamma_0 \neq 0$). In both models the time between sales component ($\gamma_1 \neq 0$) is included.

Table 6 Estimation results from local linear trend repeat sales model for semi-detached houses

Coefficient	$\gamma_0 = \gamma_1 = 0$	$\gamma_0 = 0, \gamma_1 \neq 0$	$\gamma_0 \neq 0, \gamma_1 \neq 0$	$\gamma_0 \neq 0, \gamma_1 = 0$
γ_0			0.032 (28.44)	0.037 (37.58)
γ_1		0.1871 (39.47)	0.097 (17.32)	
σ	0.076	0.076	0.074	0.071
$\sqrt{q_\eta}\sigma$	0.015	0.015	0.016	0.016
$\sqrt{q_\zeta}\sigma$	0.002	0.002	0.002	0.002
$\sqrt{q_\xi}\sigma$	0.001	0.001	0.001	0.001
$N - M$	80,162	82,005	82,005	80,162
Loglikelihood	33,190.5	34,874.6	35,269.3	33,867.3

The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that $\rho = 1$ in Eq. 2. t-values are provided between the brackets

Table 6 provides the estimation results. All estimates of the constant term (γ_0) and the time between sales component (γ_1) are highly significant (t-values are provided between the brackets). The constant term is around 0.03 to 0.04. The time between sales component γ_1 almost doubles (from 0.097 to 0.1871), depending on the fact whether γ_0 is included in the model or not. The coefficient 0.1871 is in line with the difference between the maximum and the minimum value of the averages of the residuals per time between sales in Fig. 6.

Figure 7 gives the estimated price indices. The impact of the constant in the model is large. In the model including a constant term, the slope of the index is smaller compared to the model excluding a constant term. The difference in log index value on May 2009 is around 0.10. The differences between the models excluding and including the time between sales component are very small. The price indices are virtually the same for both models, including and excluding a constant term.

It can be concluded that the inclusion of a constant term has a large downward impact on the estimated price indeces. These results are in accord with the findings of, for example, Shiller (1993) and Clapp and Giaccotto (1999). We conclude that a feasible alternative is to keep all sales in the dataset and explicitly model them, rather than delete all within–short–period repeat sales and, therefore, ignore information in the data.

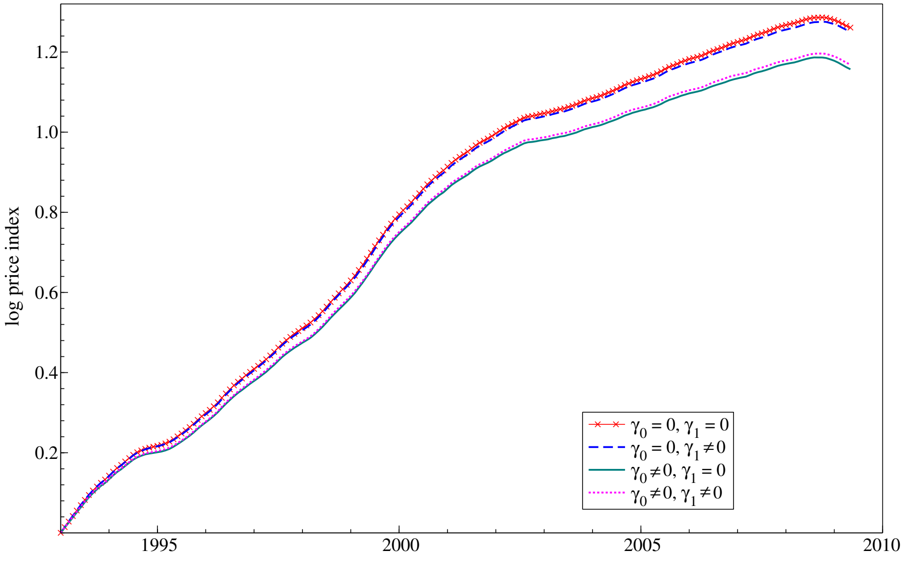

Fig. 7 Log price indices $\hat{\beta}_t$ for semi-detached houses in the local linear trend repeat sales model. The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 2 $\rho = 1$

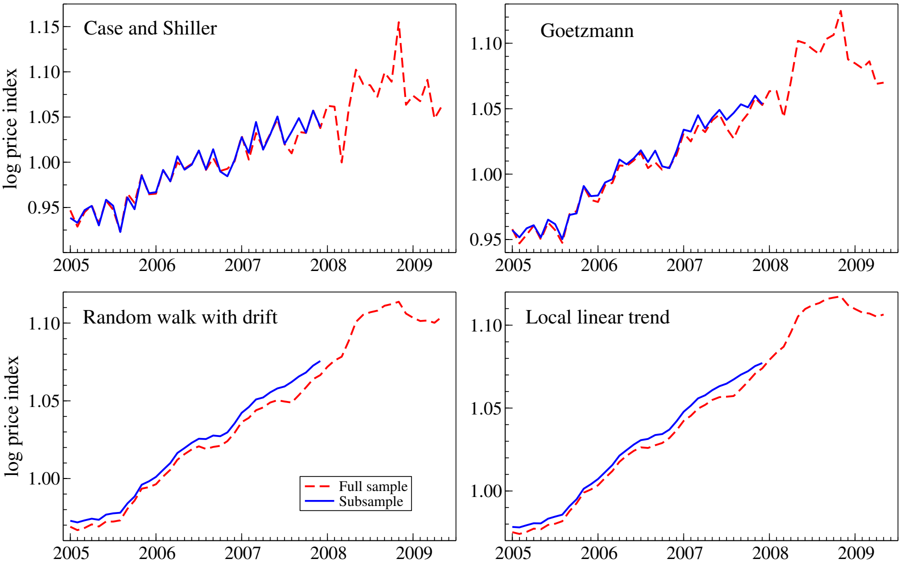

Fig. 8 Revision effects for the log price indices $\hat{\beta}_t$ in the repeat sales model for Maarssen. The local linear trend repeat sales model is provided by Eqs. 5–8. It is assumed that in Eq. 7 $\gamma_0 = \gamma_1 = 0$ and in Eq. 2 $\rho = 1$. The remaining models are special cases of the local linear trend repeat sales model

Revision

In this subsection the impact of revision for the four different repeat sales models is examined. The repeat sales models are estimated on the full sample of the sales in the city Maarssen and on a subsample, where the last 17 months are left out. The repeat sales within first 6 months are omitted. In all models, the parameters γ_0 and γ_1 in Eq. 7 are restricted to be zero. Figure 8a–d show the full sample and the subsample log price levels for the CS, Goetzman, RWD and LLT models, respectively, from January 2005. For all models it holds that the revision effect is not large in this relatively small sample. Table 7 provides summary statistics of the revision effect for the whole sample, defined as the absolute value of the differences in $\hat{\beta}_t$ between the full sample and subsample log price index. For the LLT model the revision effect is the smallest: the average effect is 0.0021 and the maximum effect is 0.0101. The revision effect is the largest for the CS model: the average effect is 0.0035 and the maximum effect is 0.0239. For larger samples, the revision effect for all models is much smaller than this.

Table 7 Average and maximum revision effects in the repeat sales model for the city Maarssen

	CS	Goetzmann	RWD	LLT
Average	0.0035	0.0024	0.0027	0.0021
Maximum	0.0239	0.0195	0.0133	0.0101

The revision effect is defined as the absolute value of the differences in $\hat{\beta}_t$ between the full sample and subsample log price index

Prediction

To illustrate the meaning of the estimated standard deviations in Table 6, we provide an example on the dataset for the semi-detached houses. We take the model in the final column of Table 6 ($\gamma_0 \neq 0, \gamma_1 = 0$) as a base model for the calculations. Let us assume that a house i is sold for €100,000 in May 2004. We want to answer the question what the value of this house will be in May 2007. To be more precise, we want to answer what is the expectation and the standard deviation of the value in 2007. The components that influence the value of the house are the price trend (β_t and κ_t), the individual house trend α_i, and the transaction price noise ε_{it}. The estimated price increase (β_t and κ_t) is 0.136 (14.5%). The random walk has zero expectation, hence the expectation of the value in May 2007 equals €114,500. The standard deviation consists of three independent parts, (1) the standard deviation of the measurement error (0.071), (2) the standard deviation of the random walk of an individual house (0.016), and (3) the standard deviation of the price movements between May 2003 and 2007. The last one can be calculated from Eq. 15 and equals 0.0038. The total standard deviation is $\sqrt{2 \times 0.071^2 + 36 \times 0.016^2 + 0.0038^2} = 0.139$, approximately 14.79%.

Conclusions

In this paper we estimate the local linear trend repeat sales model, as an alternative to repeat sales models in which the log price levels are fixed unknown parameters. For large samples, the differences between the different models are small. It does not matter whether a priori a structure is imposed (random walk with drift and local linear trend model) or not (Case and Shiller model); the estimation results do entirely depend on the data, and not on the a priori structure. However, the local linear trend repeat sales model can also be used to construct price indices in thin markets, with only a small number of repeat sales, and for short time intervals. The impact of transaction price noise on the estimation of the house price trends is considerably reduced using the local linear trend repeat sales model. As a result of the underlying trend model, the estimated price indices are stable.

The local linear trend repeat sales model can be interpreted as a modification of the Goetzmann (1992) approach. The 'constant' appreciation rate assumption (random walk with drift) is replaced by a more realistic 'time varying' appreciation rate (local linear trend model). A second modification is the estimation of the signal-to-noise ratios by maximizing the concentrated likelihood function, thus avoiding the somewhat ad hoc two-step procedure that results in overestimation of the signal-to-noise ratio, and hence in more volatile return series.

In the local linear trend and the random walk with drift repeat sales model, both estimated by maximum likelihood, the standard deviation of the estimated monthly returns is almost insensitive to the sample size: for the local linear trend model it varies between 0.0039 (for $n = 846, 439$) and 0.0046 ($n = 2, 511$). This is about 0.6–0.7 times the average monthly return. In the Case and Shiller and Goetzmann models, the standard deviation decreases with the number of observations: for the Goetzmann model it varies between 0.0043 and 0.0429 and for the Case and Shiller it varies between 0.0047 and 0.0895. In these models, the monthly standard deviation is 6 to 13 times as large as the average monthly return. This implies that the Case and Shiller and the two-stage Bayes variant of the Goetzmann models cannot be used to construct reliable detailed price indices and returns.

In addition, the local linear trend repeat sales model allow us to examine the effect of the time between repeat sales on the estimation of the price level. Empirical evidence shows that large profits are made when the time between sales is relatively short, say within first 6 months. For that reason, a new variable is included in the repeat sales model, containing the reciprocal of the time between sales, providing a satisfactorily description of the empirical evidence.

The structural time series approach that is used in this article allows for more generalizations, such as the inclusion of seasonal effects and specifications of hierarchical trends (see Francke and De Vos 2000) or common factors for different market segments. The impact of outliers can also be reduced by assuming the transaction price noise to have a t-distribution. As part of future research, these generalizations can also be dealt with within the state-space framework.

Acknowledgements I would like to thank two unknown referees, David Geltner, the participants of the Maastricht-MIT-NUS 2008 symposium, and Sunčica Vujić for helpful comments.

Appendix A: Covariance Matrix

The covariance matrix of $\widetilde{\alpha}_i$ depends on the unknown parameters σ and $\theta_1 = (\rho, q_\eta)'$. For $|\rho| < 1$, covariance stationarity is assumed, hence $\alpha_{i1} \sim N(0, q_\eta \sigma^2/(1 - \rho^2))$. For $\rho = 1$, the process (2) starts at an unknown level, say $\alpha_{i1} = \psi_i$. Note that in the 'differenced' data the parameter ψ_i cancels out. The covariance matrices for $\widetilde{\alpha}_i$ for $|\rho| < 1$ have the typical form

$$\mathrm{Var}(\widetilde{\alpha}_i) = \frac{q_\eta \sigma^2}{1 - \rho^2} \times \begin{bmatrix} 2-2\rho^{t-s}, & -1+\rho^{t-s}-\rho^{t-\tau}+\rho^{s-\tau}, & \rho^{t-\tau}-\rho^{s-\tau} - \rho^{t-\varsigma}+\rho^{s-\varsigma} \\ & 2-2\rho^{s-\tau}, & -1+\rho^{s-\tau}-\rho^{s-\varsigma}+\rho^{\tau-\varsigma} \\ & & 2-2\rho^{\tau-\varsigma} \end{bmatrix}, \tag{19}$$

where $t > s > \tau > \varsigma$, see for example Hwang and Quigley (2004). By noting that

$$\frac{1-\rho^m}{1-\rho^2} = \frac{\sum_{i=0}^{m-1} \rho^i}{\rho+1},$$

the covariance matrix $\mathrm{Var}(\widetilde{\alpha}_i)$ can equivalently be expressed as

$$\frac{q_\eta \sigma^2}{1+\rho} \begin{bmatrix} 2S(t,s) & S(t,\tau)-S(t,s)-S(s,\tau) & S(s,\tau)+S(t,\varsigma)-S(t,\tau)-S_\rho(s,\varsigma) \\ & 2S(s,\tau) & S(s,\varsigma)-S(s,\tau)-S(\tau,\varsigma) \\ & & 2S(\tau,\varsigma) \end{bmatrix}, \tag{20}$$

where $S(m,n) = \sum_{i=0}^{|m-n|-1} \rho^i$. Note that for $\rho = 1$ the diagonal elements of $\mathrm{Var}(\widetilde{\alpha}_i)$ are $q_\eta \sigma^2$ $(t-s, s-\tau, \tau-\varsigma)$ and the off-diagonal elements are 0.

Equation 20 is well defined and finite for $-1 < \rho \leq 1$. Contrary to the profile likelihood of (1)–(2), the likelihood of the 'differenced data' (7) is well defined in the unit root. For a detailed discussion of the properties of the marginal likelihood and the likelihood of the 'differenced data' in the regression model with first order autoregressive disturbances, see Francke and De Vos (2007).

Appendix B: Estimation of the Slope Parameters

The slope parameters $\kappa_1, \ldots, \kappa_{T-1}$ from the local linear trend repeat sales model can be estimated in a similar way as the trend parameters β_t in Section "Estimation". The prior for $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_T)'$ conditional on $\boldsymbol{\kappa} = (\kappa_2, \ldots, \kappa_{T-1})'$ can be expressed as

$$\boldsymbol{\beta}|\boldsymbol{\kappa} = \begin{pmatrix} 0 \\ I \end{pmatrix} \boldsymbol{\kappa} + \zeta, \tag{21}$$

where $\zeta \sim N(0, A_1)$, and A_1 is defined in Eq. 13, see Eqs. 5–6. Equation 21 can be substituted in Eq. 7, leading to

$$\widetilde{y}_i = \mathbf{i}\gamma_0 + \mathbf{p}_i^{-1}\gamma_1 + \mathbf{p}_i\kappa_1 + \widetilde{X}_i^* \boldsymbol{\kappa} + \widetilde{X}_i \zeta + \widetilde{e}_i = \widetilde{Z}_i^- \lambda + \widetilde{X}_i \zeta + \widetilde{e}_i, \tag{22}$$

where $\widetilde{X}_i^* = \widetilde{X}_i \begin{pmatrix} 0 \\ I \end{pmatrix}$, and $\lambda = (\gamma_0, \gamma_1, \kappa_1, \boldsymbol{\kappa})$. Note that the matrix $\widetilde{Z}_i^-$ is a submatrix of $\widetilde{Z}_i$; the only missing column is the column corresponding to β_2 in $\widetilde{Z}_i$.

The prior for $\boldsymbol{\kappa}$ is given by

$$\boldsymbol{\kappa} \sim N(0, \sigma^2 q_\xi A_1). \tag{23}$$

No prior information for the parameters γ_0, γ_1 and κ_1 is available, hence the precision matrix which is the inverse of the variance matrix Ψ^*, for all regression parameters $\delta = (\gamma_0, \gamma_1, \kappa_1, \boldsymbol{\kappa}')'$ is given by $\Psi^{*-1} = \begin{bmatrix} 0 & 0 \\ 0 & (q_\xi A_1)^{-1} \end{bmatrix}$.

The posterior of λ is provided by

$$\lambda | \widetilde{y}, \widetilde{Z}^-, \theta, \sigma^2 \sim N\left(\lambda^*, \sigma^2 \mathrm{V}(\lambda^*)\right), \tag{24}$$

$$V(\lambda^{*})=\left(\sum_{i=1}^{M}(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{Z}_i^{-})-\sum_{i=1}^{M}(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{X}_i^{\prime})Q^{-1}\sum_{i=1}^{M}\left(\widetilde{X}_i^{\prime}\Omega_i^{-1}\widetilde{Z}_i^{-}\right)+\Psi^{*-1}\right)^{-1}, \tag{25}$$

$$\lambda^{*}=\mathrm{V}(\lambda^{*})\left(\sum_{i=1}^{M}\left(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{y}_i\right)-\sum_{i=1}^{M}\left(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{X}_i^{\prime}\right)Q^{-1}\sum_{i=1}^{M}\left(\widetilde{X}_i^{\prime}\Omega_i^{-1}\widetilde{y}_i\right)\right), \tag{26}$$

where $Q=(q_\zeta A_1)^{-1}+\sum_{i=1}^{M}(\widetilde{X}_i^{\prime}\Omega_i^{-1}\widetilde{X}_i)$. We used the fact that the inverse of the covariance matrix of $\widetilde{X}\zeta+\widetilde{e}$ in Eq. 22 is given by

$$\left(\widetilde{X}q_\zeta A_1\widetilde{X}^{\prime}+\Omega\right)^{-1}=\Omega^{-1}-\Omega^{-1}\widetilde{X}^{\prime}\left((q_\zeta A_1)^{-1}+\widetilde{X}^{\prime}\Omega^{-1}\widetilde{X}\right)^{-1}\widetilde{X}^{\prime}\Omega^{-1}. \tag{27}$$

The slope parameters for $t>1$ are given by $\kappa_1^{*}+\kappa_t^{*}$.

Note that the data dependent matrices $\sum_{i=1}^{M}(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{Z}_i^{-})$, $\sum_{i=1}^{M}(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{X}_i^{\prime})$, $\sum_{i=1}^{M}(\widetilde{Z}_i^{-\prime}\Omega_i^{-1}\widetilde{y}_i)$, and $\sum_{i=1}^{M}(\widetilde{X}_i^{\prime}\Omega_i^{-1}\widetilde{y}_i)$ are submatrices from their counterparts in Eqs. 14–16. They do not have to be evaluated separately.

References

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.

Baroni, M., Barthélémy, F., & Mokrane, M. (2007). A PCA repeat sales index for apartment prices in Paris. *Journal of Real Estate Research, 29*, 137–158.

Bourassa, S. C., Hoesli, M., & Sun, J. (2006). A simple alternative house price index model. *Journal of Housing Economics, 15*, 80–97.

Cannaday, R. E., Munneke, H. J., & Yang, T. L. (2005). A multivariate repeat-sales model for estimating house price indices. *Journal of Urban Economics, 57*, 320–342.

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

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. (2004). A semiparametric method for estimating local house price indices. *Real Estate Economics, 32*, 127–160.

Clapp, J. M., & Giaccotto, C. (1992a). Estimating price indices for residential property: A comparison of repeat sales and assessed value methods. *Journal of the American Statistical Association, 87*, 300–306.

Clapp, J. M., & Giaccotto, C. (1992b). Estimating price trends for residential property: A comparison of repeat sales and assessed value methods. *Journal of Real Estate Finance and Economics, 5*, 357–374.

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

CPB (2009). *Centraal economisch plan*. The Hague: Netherlands Bureau for Economic Policy Analysis.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*, 1–38.

De Vries, P., Mariën, G., de Haan, J., & Van der Wal, E. (2007). A house price index based on the SPAR method. In *Tech. rep., Paper presented at the Cambridge—UNC Charlotte Symposium on Real Estate Risk Management*.

Durbin, J., & Koopman, S. J. (2001). *Times series analysis by state space methods*. Oxford: Oxford University Press.

Englund, P., Quigley, J. M., & Redfearn, C. L. (1999). The choice of methodology for computing housing price indexes: Comparisons of temporal aggregation and sample definition. *Journal of Real Estate Finance and Economics, 19*, 91–112.

Francke, M. K. (2008). The hierarchical trend model. In T. Kauko, & M. Damato (Eds.), *Mass appraisal methods; an international perspective for property valuers* (pp. 164–180). New York: Wiley-Blackwell RICS Research.

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., & De Vos, A. F. (2007). Marginal likelihood and unit roots. *Journal of Econometrics, 137*, 708–728.

Francke, M. K., & Vos, G. A. (2004). The hierarchical trend model for property valuation and local price indices. *Journal of Real Estate Finance and Economics, 28*, 179–208.

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

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

Goetzmann, W. N., & Spiegel, M. (1995). Non-temporal components of residential real estate appreciation. *The Review of Economics and Statistics, 77*, 199–206.

Hannonen, M. (2005). An analysis of land prices: A structural time-series approach. *International Journal of Strategic Property Management, 9*, 145–172.

Hannonen, M. (2008). Predicting urban land prices: A comparison of four approaches. *International Journal of Strategic Property Management, 12*, 217–236.

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

Harvey, A., & Koopman, S. J. (1999). Signal extraction and the formulation of unobserved components models. *Econometrics Journal, 3*, 84–107.

Heckman, J. (1979). Sample selectivity bias as a specificiation error. *Econometrica, 47*, 153–161.

Hill, R. C., Knight, J. R., & Sirmons, C. F. (1997). Estimating capital asset price indexes. *The Review of Economics and Statistics, 79*, 226–233.

Hwang, M., & Quigley, J. M. (2004). Selectivity, quality adjustment and mean reversion in the measurement of house values. *Journal of Real Estate Finance and Economics, 28*, 161–178.

Jansen, S., De Vries, P., Coolen, H., Lamain, C., & Boelhouwer, P. (2008). Developing a house price index for the Netherlands: A practical application of weighted repeat sales. *Journal of Real Estate Finance and Economics, 37*, 163–186.

Koopman, S. J., & Harvey, A. (2003). Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics & Control, 27*, 1317–1333.

Koopman, S. J., Shephard, N., & Doornik, J. A. (1999). Statistical algorithms for models in state space using ssfpack 2.2. *Econometrics Journal, 2*(1), 113–166.

Koopman, S. J., Harvey, A. C., Doornik, J. A., & Shephard, N. (2007). *STAMP, structural time series analyser, modeller and predictor* (8th ed.). London: Timberlake Consultants Press.

Kuo, C. L. (1997). A Bayesian approach to the construction and comparison of alternative house price indices. *Journal of Real Estate Finance and Economics, 14*, 113–132.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34*, 1–41.

McMillen, D. P., & Dombrow, J. (2001). A flexible Fourier approach to repeat sales price indexes. *Real Estate Economics, 29*, 207–225.

McMillen, D. P., & McDonald, J. (2004). Reaction of house prices to a new rapid transit line: Chicago's midway line, 1983–1999. *Real Estate Economics, 32*, 463–486.

Schulz, R., & Werwartz, A. (2004). A state space model for Berlin house prices: Estimation and economic interpretation. *Journal of Real Estate Finance and Economics, 28*, 37–57.

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. (1993). *Macro markets, creating institutions for managing society's largest economic risks*. Oxford: Oxford University Press.

Shumway, R. H., & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis, 3*, 253–264.

Thorsnes, P., & Reifel, J. W. (2007). Tiebout dynamics: Neighborhood response to a central-city/suburban house-price differential. *Journal of Regional Science, 47*, 693–719.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall/CRC.

Watson, M. W., & Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor. Mimic and varying coefficient regression models. *Journal of Econometrics, 23*, 385–400.

Webb, C. (1988). A probabilistic model for price levels in discontinuous markets. In W. Eichhorn (Ed.), *Measurement in economics*. Heidelberg: Physic-Verlag.

West, M., & Harrisson, J. (1997). *Bayesian forecasting and dynamic models* (2nd ed.). New York: Springer.