The Annals of Applied Statistics
2017, Vol. 11, No. 2, 808–839
DOI: 10.1214/17-AOAS1019

CLUSTERING CORRELATED, SPARSE DATA STREAMS TO ESTIMATE A LOCALIZED HOUSING PRICE INDEX[1]

BY YOU REN, EMILY B. FOX AND ANDREW BRUCE

University of Washington

Understanding how housing values evolve over time is important to policy makers, consumers and real estate professionals. Existing methods for constructing housing indices are computed at a coarse spatial granularity, such as metropolitan regions, which can mask or distort price dynamics apparent in local markets, such as neighborhoods and census tracts. A challenge in moving to estimates at, for example, the census tract level is the scarcity of spatiotemporally localized house sales observations. Our work aims to address this challenge by leveraging observations from multiple census tracts discovered to have correlated valuation dynamics. Our proposed Bayesian nonparametric approach builds on the framework of latent factor models to enable a flexible, data-driven method for inferring the clustering of correlated census tracts. We explore methods for scalability and parallelizability of computations, yielding a housing valuation index at the level of census tract rather than zip code, and on a monthly basis rather than quarterly. Our analysis is provided on a large Seattle metropolitan housing dataset.

1. Introduction. The housing market is a large part of the global economy. In the United States, roughly half of household wealth is in residential real estate [Iacoviello (2011)]. Understanding how housing value changes over time is important to policy makers, consumers, real estate professionals and mortgage lenders. Valuation is relatively straightforward for commoditized sectors of the economy, such as energy or nondiscretionary spending. By contrast, valuation of residential real estate is intrinsically difficult due to the individual nature of houses and the changing composition of houses sold from one time period to the next. Consequently, economists and public policy researchers have devoted considerable effort to developing a meaningful index to measure the change in housing prices over time.

The most common approach to constructing a housing price index is the repeat sales model, first proposed by Bailey, Muth and Nourse (1963) and then extended in numerous ways over the years [cf. Case and Shiller (1987, 1989), Gatzlaff and Haurin (1997), Shiller (1991), Goetzmann and Peng (2002)]. The main idea is to

Received April 2015; revised December 2016.

[1]Supported in part by a gift provided by Zillow, NSF CAREER Award IIS-1350133, the TerraSwarm Research Center sponsored by MARCO and DARPA, and DARPA Grant FA9550-12-1-0406 negotiated by AFOSR.

Key words and phrases. Bayesian nonparametrics, clustering, housing price index, multiple time series, state space models.

use a pair of sales for the same house to model the price trend over time, largely circumventing the problem caused by the change in composition of houses sold. The repeat sales model is the basis for the Case–Shiller home value index, published by Core-Logic and widely disseminated by the media.

One drawback of a repeat sales model is that houses with only a single sales transaction get discarded from the dataset. In growing metropolitan areas, single sales can make up a vast majority of sales; for example, 93%–97% during the 16-year period studied by Case and Shiller (1987). Furthermore, repeat sales properties tend to be older, smaller and more modest than single-sale properties [Englund, Quigley and Redfearn (1999), Meese and Wallace (1997)], presenting a sampling selection bias. Case and Quigley (1991) instead propose a hybrid model that combines repeat sales with house-level covariates (*hedonics*) to make use of all sales. More recently, Nagaraja, Brown and Zhao (2011) propose an autoregressive repeat sales model that also utilizes all sales data, but without the need for hedonic information.

However, even repeat sales models that use all of the transactions are only appropriate when fit to relatively large areas. This is due to the fact that—despite the large number of house sales observations in aggregate—there are very few spatiotemporally-localized sales. For example, in our dataset described in Section 2, most census tracts (114 out of 140) have fewer than 5 sales per month on average and more than 10% of tracts have fewer than 1 sale on average per month (see Table 1). The scarcity of transactions makes it challenging to obtain stable parameter estimates for small regions, and thus repeat sales models lack predictive accuracy. Even the advanced approach of Nagaraja, Brown and Zhao (2011) only produces an index estimated quarterly at the zip code level rather than monthly at the census tract level. This is a significant limitation: the value of real estate is intrinsically local and coarse-scale estimates may mask or distort key phenomena.

The main contribution of this paper is developing a model-based approach to creating housing indices on a finer spatiotemporal granularity than current methods. The indices are valuable for direct analysis and also as input to house-level models. Our formulation is based on a dynamic model that introduces a latent process to capture the census-tract-level housing valuation index on a monthly basis. [The ideas scale to finer spatiotemporal resolutions, as demonstrated in Supplement G.5 of Ren, Fox and Bruce (2017).] This latent process is informed by all

TABLE 1
Number of census tracts in Seattle City that have less than single digit transactions per month on average

Average monthly sales	**< 1**	**< 3**	**< 5**	**< 7**	**< 9**
Number of tracts	16	58	114	136	139
Proportion of tracts	0.11	0.41	0.81	0.97	0.99

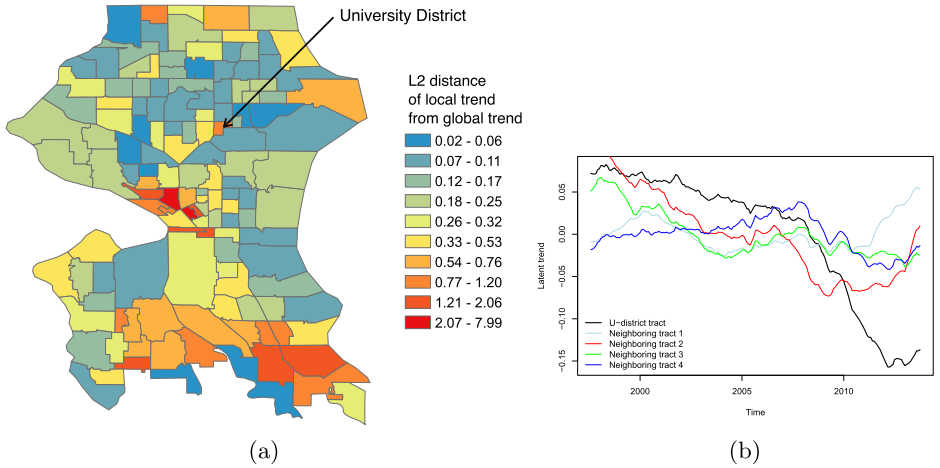

FIG. 1. (a) *Map of inferred tract-specific latent price dynamics, where the color shows how different the local trend is from the global trend, measured in L2 distance over time.* (b) *The University District's latent price dynamics* (black), *which vary significantly from its neighboring census tracts* (other colors). *More details are in Section 6.*

individual house sales within the census tract, including detailed information of sales prices and house hedonics. To overcome the sparseness of sales within a census tract, we inform the latent price trends based on sales in multiple census tracts discovered to have correlated dynamics.

Unlike many spatiotemporal processes, modeling the between-track correlations as a function of Euclidean distance is not appropriate since spatially disjoint regions can behave quite similarly while neighboring census tracts can have significantly different value dynamics. Indeed, our analysis of house sales in Seattle (further described in Section 6) indicates such structure. Figure 1(a) shows a map of deviations of each census tract's inferred local price dynamics from a global trend. We clearly see spatially abrupt changes between neighboring regions. One example is the University District (U-District). Figure 1(b) shows that the price trend in the U-District behaves differently compared to its neighboring census tracts. This census tract is heavily populated by University of Washington students and has a higher crime rate than neighboring tracts. Instead of relying on an explicit spatial model, we develop a Bayesian nonparametric clustering approach to infer the relational structure of the census tracts based solely on observed house sales prices (after accounting for associated hedonics). Within a cluster, the latent value dynamics are correlated whereas census tracts in different clusters are assumed to evolve independently. By leveraging Bayesian nonparametrics—specifically a Dirichlet process prior on the factor process of a latent factor model—our formulation enables a flexible, data-driven method for discovering these clustered dynamics, including the number of clusters.

Our formulation represents a fundamentally different approach to clustering time series. Standard methods are based on similarities in the observed processes, or based on sharing dynamical model parameters in a way that results in each series being a noisy version of a canonical series; see [Liao (2005)] for a survey. We instead cluster based on structure in a latent process—critical to handling our multiple and missing data structure—and furthermore define clusters based on *correlation*. A cluster can then capture, for example, one latent trend rising as another decreases.

The approach taken also offers several advantages over existing housing index methods. Our hierarchical Bayesian nonparametric model efficiently shares information between clustered series—a critical feature to attain high resolution. In particular, our approach provides a form of multiple shrinkage, improving stability of our estimates in this data-scarce scenario. Likewise, the joint Bayesian framework considers all uncertainties together in the clustering, latent price inference and model parameter estimation.

Our paper is organized as follows. Section 2 introduces our Seattle house transaction data and an exploratory data analysis to motivate our modeling choices. Section 3 describes the dynamical model for each census tract individually, and then the correlation structure introduced to couple the tract dynamics. The prior distributions are also specified. Section 4 provides an outline of the posterior sampling steps, and Section 4.5 discusses some of the computational challenges and a strategy to implement the algorithm in parallel. A simulation study is provided in Section 5 and a detailed analysis of our Seattle housing dataset is in Section 6. Section 7 details how the global nonstationary trend can be jointly modeled and estimated.

2. Exploratory analysis of house transaction data. Our house sales data consists of 124,480 transactions in the 140 census tracts of the City of Seattle from July 1997 to September 2013. Foreclosure sales are not included. For each house sale, we have the jurisdiction of the house (i.e., census tract FIPS code, zip code), month and year of the sale, the sales price and house covariates; the latter are commonly referred to as *hedonics* in the housing literature. Our hedonic variables include `number of bathrooms`, `finished square feet` and `square feet of the lot size`.

The scarcity of data localized in space and time is summarized in Table 1, here at the granularity of census tracts and months. To motivate the importance of considering related tracts jointly in this data-scarce regime, we performed the following data analysis. Using the per-tract dynamical model of equations (3.3)–(3.4), we independently analyzed each tract (whereas in Section 3 the focus is on joint modeling of tracts). The latent state sequence represents the underlying price evolution of a given region—our desired index—and the observations are the individual house sales in terms of log price. For this exploratory analysis, we infer the latent state sequence jointly with the model parameters using a Kalman smoother embedded in

an expectation maximization (EM) algorithm. We compare the performance of this independent, per-tract analysis to that of jointly analyzing related tracts. For the sake of exploratory analysis, the latter is determined here by a hierarchical clustering approach based on a variance-adjusted L_2 distance between the independently Kalman-smoothed estimates of the latent state sequences. Specifically, the Kalman smoother generates the mean and variance processes for the latent state sequence, denoted as $\mu_{t,i} \equiv \mathrm{E}(X_{t,i}|Y_{1:T,i})$ and $V_{t,i} \equiv \mathrm{Var}(X_{t,i}|Y_{1:T,i})$ for census tract i. The variance-adjusted L_2 distance between latent state sequences for census tracts i and j is defined as $\sum_t \frac{(\mu_{t,i}-\mu_{t,j})^2}{V_{t,i}+V_{t,j}}$. After performing the hierarchical clustering and cutting the tree by specifying the number of clusters, we consider a multivariate latent state model as in equation (3.3) where all tracts i falling in the same cluster have correlated innovations, $\varepsilon_{t,i}$; that is, $\boldsymbol{\varepsilon}_t^{(k)} \sim N(0, \Sigma_k)$ for Σ_k full, where $\boldsymbol{\varepsilon}_t^{(k)}$ is the vector of $\varepsilon_{t,i}$ for tracts i in cluster k. The observation model remains as in equation (3.4). We then applied a Kalman-smoother-within-EM algorithm to the resulting collection of cluster-specific multivariate state space models. Unsurprisingly, without sharing observations from similar tracts, the baseline independent approach does not perform well when the observations are sparse, as shown in Figure 2(a). In contrast, by leveraging observations from other tracts, the hierarchical clustering-based latent price dynamics are smoother and with narrower intervals, as shown in Figure 2(b).

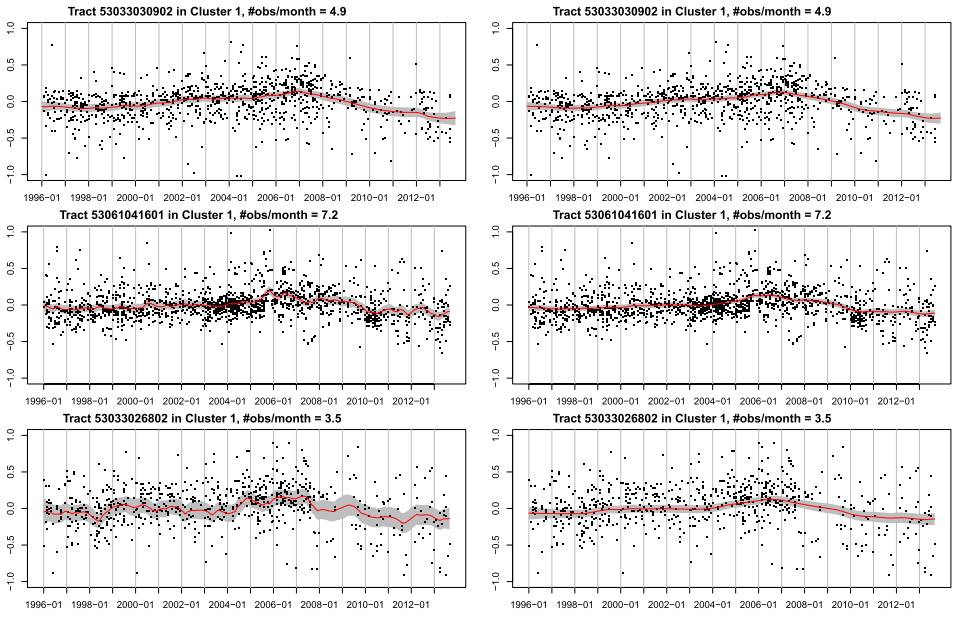

FIG. 2. *Illustrating the importance of joint modeling*: (left) *univariate Kalman smoother applied independently to time series of each census tract*, (right) *multivariate Kalman smoother applied jointly to tracts in the same cluster inferred using hierarchical clustering.*

Although this exploratory analysis motivates the importance of considering related tracts jointly, the hierarchical clustering approach considered in this section is ad hoc since it divides the clustering and estimation into three stages rather than one a unified framework. For example, errors in the independent state estimation stage can propagate to the clusterings inferred at the second stage, which are used for the multivariate analysis in the third stage. Additionally, the proposed multivariate model does not scale well to large clusters due to the associated large number of parameters represented by Σ_k. In Figure 2(b), we simply consider a cluster with 3 tracts. Moreover, the approach requires the user to specify the number of clusters (tree level) and distance metric used in the hierarchical clustering. Regardless, the insights and intuition from this exploratory analysis—clustering and correlating time series—motivates the unified statistical model for relating multiple time series presented in Section 3.

3. A local-level housing index model. Our modeling strategy for handling the scarcity of data locally in space and time is to discover price dynamics shared between region-specific data streams, allowing us to leverage observations from related regions. We first describe a model for the individual housing valuation indices and then describe a clustering-based framework for correlating these indices across regions. Throughout, we will assume that our geographic unit of interest is a census tract.

3.1. *Per-region dynamics*. We model the dynamics of the log house sales prices within a census tract via a state space model. Each census tract i may have multiple house sale observations $\tilde{y}_{t,i,l}$ (log price) at time t, $t = 1, \ldots, T$. We assume that these sales are noisy, independent observations of the latent census tract value $\tilde{x}_{t,i}$ after accounting for house-level hedonics U_ℓ (e.g., square feet):

$$\tilde{x}_{t,i} = g_t + a_i(\tilde{x}_{t-1,i} - g_{t-1}) + \varepsilon_{t,i}, \qquad \varepsilon_{t,i} \sim \mathcal{N}\big(0, \sigma_i^2\big), \tag{3.1}$$

$$\tilde{y}_{t,i,l} = \tilde{x}_{t,i} + f_i(U_l) + v_{t,i,l}, \qquad v_{t,i,l} \sim \mathcal{N}(0, R_i). \tag{3.2}$$

Our discrete-time model is indexed monthly and g_t is the global market trend that captures overall, nonstationary behavior of the time series. To account for the hedonics, we use a census tract-specific regression $f_i(\cdot)$.

To focus and simplify the discussion on capturing dynamics in small geographic regions, in this section we assume that the global trend g_t is known or precalculated based on all transactions in the market. We turn to the modeling and joint estimation of the global trend in Section 7. To specify our model of the *deviation* of the latent dynamics of census tracts from the global market trend, we define $x_{t,i} \equiv \tilde{x}_{t,i} - g_t$ and $y_{t,i,l} \equiv \tilde{y}_{t,i,l} - g_t$. For simplicity, we assume the house feature function $f_i(\cdot)$ is composed of linear basis functions. The resulting model is given by

$$x_{t,i} = a_i x_{t-1,i} + \varepsilon_{t,i}, \qquad \varepsilon_{t,i} \sim \mathcal{N}\big(0, \sigma_i^2\big), \tag{3.3}$$

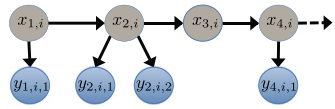

FIG. 3. *Graphical model associated with equations* (3.3)–(3.4) *for census tract i's data stream.*

$$y_{t,i,l} = x_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,h} + v_{t,i,l}, \qquad v_{t,i,l} \sim \mathcal{N}(0, R_i). \tag{3.4}$$

It is worthwhile noting that if the goal is to perform house-level predictions, a more sophisticated hedonic regression model would be appropriate. For example, [Brunauer, Lang and Umlauf (2013)] consider a multilevel structured additive regression model to leverage the hierarchical structure of neighborhood attributes in addition to house-level hedonics. The model also incorporates nonlinear effects, smooth spatial effects and interaction terms. Models, such as these, that more fully account for hedonic effects would be straightforward to incorporate into our methodology.

We refer to the latent $x_{t,i}$ order 1 autoregressive process [AR(1)] in equation (3.3) as the *intrinsic price dynamics* for each census tract. Since we are modeling the deviation from the global trend, the choice of a stationary process is reasonable. Equations (3.3) and (3.4) are akin to a standard linear-Gaussian state space model, but with a varying number (potentially 0) of observations $y_{t,i,l}$ of a given state $x_{t,i}$, as illustrated in Figure 3.

3.2. *Clustering region-specific data streams.* The evolution of intrinsic price dynamics are correlated across tracts, which can be captured by treating the $\varepsilon_{t,i}$ jointly rather than independently across i. Let $\boldsymbol{\varepsilon}_t$ be the vector of $\varepsilon_{t,i}$ for $i = 1, \ldots, p$. The most general correlation structure would assume $\boldsymbol{\varepsilon}_t \sim N(0, \Sigma)$ for Σ a full $p \times p$ positive semidefinite matrix.[2] However, both statistically and computationally, we cannot handle a model with an arbitrary $p \times p$ covariance matrix Σ: we have insufficient data to estimate such a large matrix, and even if we could, the resulting computations involved in estimating the intrinsic price dynamics would involve prohibitively costly $O(p^3 T)$ operations. One alternative is to assume conditional independencies between tracts based on spatial adjacencies (sparsity in Σ^{-1}). However, as depicted in Figure 1, Euclidean distance is not the right metric for describing relationships between tracts in this application.

Instead, we seek to *discover* clusters of correlated latent time courses. This correlation structure is induced by a latent factor model, leading to a low-rank covariance decomposition within clusters, and assumed independence of dynamics

[2]Note that our focus here is on the *instantaneous* or *conditional covariance* Σ, rather than directed relationships determined by the $a_{i,j}$ relating $x_{t-1,j}$ to $x_{t,i}$. In this application, we imagine synchrony in market changes across tracts driven by external factors rather than price changes in one tract *leading* to price changes in another.

between clusters. The idea for clustering multiple data streams has two justifications. From a data generating perspective, housing price dynamics are naturally clustered due to a number of factors, including the composition of homes, number of foreclosures, school district boundaries, crime rate, and the proximity to parks, waterfront and other amenities. From a statistical inference perspective, clustering census tracts increases power and precision in parameter estimation by pooling the observations from grouped data streams. In essence, this is what real estate agents commonly do: If there are no recent house sales in a given neighborhood, they look to house sales occurring in other neighborhoods they deem related.

To arrive at a model of clusters of correlated time series, we take $\boldsymbol{\varepsilon}_t^{(k)} \sim \mathcal{N}(0, \Sigma_k)$ for Σ_k nondiagonal and $\boldsymbol{\varepsilon}_t^{(k)}$ the vector of innovations $\varepsilon_{t,i}$ for census tracts i in cluster k. We assume $\boldsymbol{\varepsilon}_t^{(k)}$ is independent of $\boldsymbol{\varepsilon}_t^{(j)}$ for all $j \neq k$. Stacking up all $\boldsymbol{\varepsilon}_t^{(k)}$, $k = 1, \ldots, K$, into a large $\boldsymbol{\varepsilon}_t$ vector of length p (the number of census tracts), our model is equivalent to $\boldsymbol{\varepsilon}_t \sim N(0, \Sigma)$ for Σ block diagonal with blocks Σ_k. A key question is how to discover this clustering structure from data. This equates to the challenging task of inferring the number of blocks, size of blocks and ordering of census tracts in $\boldsymbol{\varepsilon}_t$.

Both for a parsimonious specification of the correlation structure within clusters—crucial to our data-scarce scenario—and to yield a framework in which to discover the cluster memberships, we assume a *latent factor model* for $\boldsymbol{\varepsilon}_t^{(k)}$. In particular, for all tracts i in cluster k, we specify

$$\varepsilon_{t,i} = \lambda_{ik}\eta_{t,k}^* + \tilde{\varepsilon}_{t,i}, \qquad \tilde{\varepsilon}_{t,i} \sim \mathcal{N}\big(0, \sigma_0^2\big), \qquad \eta_{t,k}^* \sim \mathcal{N}(0, 1). \tag{3.5}$$

Here, $\eta_{t,k}^*$ is the latent factor associated with cluster k at time t, λ_{ik} is the factor loading for census tract i assuming it is in cluster k, and $\tilde{\varepsilon}_{t,i}$ is idiosyncratic noise drawn independently over time and tracts. We can then write $\boldsymbol{\varepsilon}_t = (\Lambda \cdot Z)\boldsymbol{\eta}_t^* + \tilde{\boldsymbol{\varepsilon}}_t$, where Λ is a $p \times K$ real-valued matrix, Z is an indicator matrix with (i, k)th entry equal to 1 if tract i is in cluster k and 0 otherwise, $\boldsymbol{\eta}_t^* \sim \mathcal{N}_K(0, I)$ and $\tilde{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}_p(0, \sigma_0^2 I)$. Here, $A \cdot B$ represents the element-wise product. Conditioned on the factor loading matrices Λ and Z, the covariance for $\boldsymbol{\varepsilon}_t$ is $\Sigma = (\Lambda \cdot Z)(\Lambda \cdot Z)^T + \sigma_0^2 I$. Equivalently,

$$\operatorname{cov}(\varepsilon_{t,i}, \varepsilon_{t,i'} | \Lambda, Z) = \begin{cases} \lambda_{ik}\lambda_{i'k} + \sigma_0^2\delta(i, i'), & k \text{ s.t. } Z_{ik} = Z_{i'k} = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

Here, $\delta(i, i') = 1$ if $i = i'$ and zero otherwise. From equation (3.6), the conditional covariance for $\boldsymbol{\varepsilon}_t$ is a block-diagonal matrix defined by the clusterings specified by Z; that is, data streams within the same cluster will have correlated dynamics, and those in different clusters will evolve independently. This model, along with our prior specification of Section 3.3, is related to that of Palla, Ghahramani and Knowles (2012), but specified here for the time series domain.

3.3. *Bayesian nonparametrics for discovering the clustering structure.* To infer the clustering of tract-specific data streams, we propose a Bayesian nonparametric approach using a Dirichlet Process (DP) prior on the parameters of a mixture model. This approach leads to an adaptive, data-driven clustering allowing for an unknown number of blocks (clusters) in the covariance. As described in Section 3.2, the quantity defining each cluster is the latent factor process $\eta^*_{1:T,k} = (\eta^*_{1,k}, \ldots, \eta^*_{T,k})$, with $\eta^*_{t,k}$ as in equation (3.5). Our mixture model is then defined in terms of an infinite set of mixture weights $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ and cluster centers $\theta^*_k = \eta^*_{1:T,k}$ for $k = 1, 2, \ldots$. We specify a DP prior on these parameters.

A DP [Blackwell and MacQueen (1973), Ferguson (1973)] is a distribution over countably infinite discrete probability measures. A draw $G \sim \mathrm{DP}(\alpha, G_0)$, with concentration parameter α and base measure G_0, can be constructed as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_k}, \qquad \theta^*_k \sim G_0, \tag{3.7}$$

where π_k are sampled via a stick breaking construction [Sethuraman (1994)]:

$$\pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \qquad v_k \sim \mathrm{Beta}(1, \alpha). \tag{3.8}$$

We denote the stick breaking process as $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$. The DP prior produces clusters of $\theta_i \sim G$, $i = 1, \ldots, p$, due to the fact that G is a discrete probability measure (i.e., multiple θ_i are sampled with identical values θ^*_k). Equivalently, we can introduce cluster indicators $z_i \sim \boldsymbol{\pi}$ such that $z_i = k$ implies that θ_i takes the unique value θ^*_k; that is, $\theta_i = \theta^*_{z_i}$. Recall that in our housing application, the cluster-specific parameter θ^*_k equates with $\eta^*_{1:T,k}$.

Integrating out the stick breaking measure $\boldsymbol{\pi}$, the predictive distribution of z_i given the memberships of tracts $z_1, \ldots, z_{i-1}$ is

$$P(z_i = k | \mathbf{z}_{-i}, \alpha) \propto \begin{cases} \dfrac{n_k}{i - 1 + \alpha}, & \text{for } k = 1, \ldots, K, \\ \dfrac{\alpha}{p - 1 + \alpha}, & \text{for } k = K + 1, \end{cases} \tag{3.9}$$

where K indicates the number of unique values in $z_1, \ldots, z_{i-1}$ and n_k the number assigned to cluster k; that is, tract i joins an existing cluster with probability proportional to the size of the cluster, n_k, or starts a new cluster with probability proportional to α. The resulting sequence of partitions is referred to as the *Chinese Restaurant Process* (CRP) [Pitman (2006)].

In summary, our Bayesian nonparametric clustering model defines mixture weights $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$, cluster-specific parameters $\eta^*_{1:T,k} \sim G_0$, and cluster indicators $z_i \sim \boldsymbol{\pi}$. The base measure G_0 is specified as a multivariate normal distribution $\mathcal{N}_T(0, I)$ such that $\eta^*_{t,k} \sim N(0, 1)$ for $t = 1, \ldots, T$, $k = 1, 2, \ldots$. Note that the cluster indicators z_i fully specify the matrix Z of equation (3.6). The result of

this model specification is the ability to learn an unknown number of clusters of *correlated* time series, with cluster-specific correlation structure specified in equation (3.6). In contrast, the Bayesian nonparametric time series clustering model of Nieto-Barajas and Contreras-Cristán (2014) clusters based on the underlying state sequence (our $x_{1:T,i}$) and observation covariate effects (our $\beta_{i,1:H}$) in a state space model similar to equations (3.3)–(3.4). The result is that time series within a cluster are assumed to be noisy versions of the same underlying process, which represents a fundamentally different notion of time series clustering. For example, our model can capture negatively correlated series, which would not be identified as similar according to the model of Nieto-Barajas and Contreras-Cristán (2014). One could also imaging including the AR parameters a_i in our clustering. However, from an exploratory data analysis using the hierarchical clustering approach of Section 2, we did not see evidence of the AR parameters being distinguished between clusters.

3.4. *Prior on other model parameters.*

Latent AR parameters. The latent AR(1) process in equation (3.3) governing the intrinsic price dynamics—using the innovation structure of equation (3.5)—has an autoregressive parameter a_i, factor loadings λ_{ik}, and the idiosyncratic noise variance σ_0^2. We place conjugate priors on these parameters, respectively:

$$a_i \sim \mathcal{N}(\mu_a, \sigma_a^2), \qquad i = 1, \ldots, p, \tag{3.10}$$

$$\lambda_{ik} \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2), \qquad i = 1, \ldots, p, k = 1, 2, \ldots, \tag{3.11}$$

$$\sigma_0^2 \sim \mathrm{IG}(\alpha_{\varepsilon 0}, \beta_{\varepsilon 0}). \tag{3.12}$$

The hyperparameters μ_a, σ_a^2, μ_λ, σ_λ^2 are also given priors. These hyperpriors and settings for hyperparameters $\alpha_{\varepsilon 0}$, $\beta_{\varepsilon 0}$ are provided in Supplement E.1 of Ren, Fox and Bruce (2017).

Emission parameters. Recalling the emission process in equation (3.4), we place conjugate priors on the tract-specific hedonic parameters $\beta_{i,h}$ and observation variance R_i:

$$\beta_{i,h} \sim \mathcal{N}(\mu_h, \sigma_h^2), \qquad i = 1, \ldots, p, h = 1, \ldots, H, \tag{3.13}$$

$$R_i \sim \mathrm{IG}(\alpha_{R0}, \beta_{R0}), \qquad i = 1, \ldots, p. \tag{3.14}$$

We further assume priors on μ_h and σ_h^2. These hyperpriors and the values of the hyperparameters α_{R0} and β_{R0} are provided in Supplement E.2 of Ren, Fox and Bruce (2017).

Figure 4 shows the graphical model representation of the resulting model.

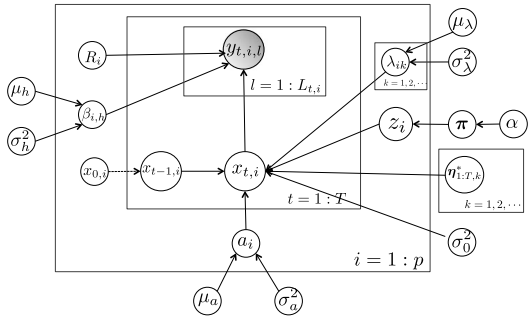

FIG. 4. *Graphical model of our Bayesian nonparametric dynamical model. Boxes indicate replication of random variables and shaded nodes the observations. Note that $x_{1:T,i}$ forms a length T Markov chain; our box here is an abuse of notation used for compactness.*

4. MCMC posterior computations. Our posterior computations are based on a Gibbs sampler, with steps outlined below and detailed derivations in the supplemental article [Ren, Fox and Bruce (2017)]. Scaling this sampling strategy to our large housing dataset is discussed in Section 4.5.

Let $\boldsymbol{\Theta} = \{\mathbf{a} = \{a_i\}, \boldsymbol{\lambda} = \{\lambda_{ik}\}, \mathbf{R} = \{R_i\}, \boldsymbol{\beta} = \{\beta_{i,h}\}, \sigma_0^2\}$ and $\boldsymbol{\Theta}^{(k)}$ the associated subset of parameters corresponding to the kth cluster based on assignments $\mathbf{z} = \{z_i\}$. Throughout, we use ϕ_{-i} to denote the removal of tract i's contribution to some set ϕ. Our Gibbs sampler iterates between:

1. Sample $z_i = k|\mathbf{z}_{-i}, \alpha, \mathbf{y}, \boldsymbol{\Theta}$. We marginalize the stick-breaking random measure $\boldsymbol{\pi}$, the latent housing valuation processes $\mathbf{x}^{(k)}$ and the cluster latent factor processes $\boldsymbol{\eta}^{*(k)}$.
2. Impute $\mathbf{x}$ and $\boldsymbol{\eta}^*$ as auxiliary variables. Specifically, block sample $\mathbf{x}$, $\boldsymbol{\eta}^*$ as $\mathbf{x}^{(k)}|\mathbf{z}, \mathbf{y}^{(k)}, \boldsymbol{\Theta}^{(k)}$ and $\boldsymbol{\eta}^*|\mathbf{z}, \mathbf{x}, \boldsymbol{\Theta}$.
3. Sample $\boldsymbol{\Theta}^{(k)}|\mathbf{z}, \mathbf{y}^{(k)}, \mathbf{x}^{(k)}, \boldsymbol{\eta}^{*(k)}$.
4. Discard $\mathbf{x}$, $\boldsymbol{\eta}^*$ and sample hyperparameters conditional on $\boldsymbol{\Theta}$, $\mathbf{z}$.

4.1. *Step* 1: *Sampling the cluster membership.* The full conditional for the cluster indicator z_i marginalizing $\boldsymbol{\pi}$, $\boldsymbol{\eta}^* = \{\eta^*_{1:T,k}\}$ and $\mathbf{x} = \{x_{1:T,i}\}$ is

$$\begin{aligned} &P(z_i = k|\mathbf{z}_{-i}, \mathbf{y}_{1:T}, \boldsymbol{\Theta}, \alpha) \\ &\quad \propto P(z_i = k|\mathbf{z}_{-i}, \alpha) P\big(\mathbf{y}_{1:T,i}|z_i = k, \mathbf{z}_{-i}, \mathbf{y}^{(k)}_{1:T,-i}, \boldsymbol{\Theta}^{(k)}\big). \end{aligned} \tag{4.1}$$

The first factor represents the CRP prior of equation (3.9) (using exchangeability of the z_i). The second factor is the likelihood of the data stream for tract i assuming membership to cluster k. The marginalization over $\mathbf{x}$ and $\boldsymbol{\eta}^*$ results in a dependence upon all other data streams in cluster k, $\mathbf{y}^{(k)}_{1:T,-i}$. Note that $\boldsymbol{\Theta}^{(k)}$ includes parameters for tract i when conditioning upon $z_i = k$.

A message passing scheme along the entire sequence of length T is required to compute the likelihood of the ith data stream conditioned on all others in cluster k, integrating over the intrinsic dynamics $\mathbf{x}_{1:T}^{(k)}$. This algorithm is essentially a Kalman filter, but allows for a varying number of observations per time step, including no observations for some time periods. The detailed algorithm is provided in Supplement B.1 of Ren, Fox and Bruce (2017).

For the special case of census tract i creating a new cluster, that is, $z_i = K+1$, the prior belief follows the CRP prior of equation (3.9). The likelihood becomes simply $P(\mathbf{y}_{1:T,i}|\mathbf{z}, a_i, \Sigma^{(K+1)}, R_i, \beta_{i,h})$, where $\Sigma^{(K+1)} = \lambda_{i,K+1}^2 + \sigma_0^2$, having sampled $\lambda_{i,K+1} \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2)$ for all tracts, but marginalizing $\eta_{1:T,K+1}^*$. This represents a variant of Neal's Algorithm 8 for sampling from DP models [Neal (2000)].

4.2. *Step* 2: *Block-sampling the intrinsic price dynamics* $\mathbf{x}$ *and cluster latent factor processes* $\boldsymbol{\eta}^*$. To block sample $(\mathbf{x}, \boldsymbol{\eta}^*)$, we first sample the intrinsic price dynamics $\mathbf{x}_{1:T}^{(k)}$ jointly for all tracts in cluster k, analytically marginalizing $\boldsymbol{\eta}^*$. To do this, we use a forward-filter backward-sampler (FFBS) outlined in Supplement C.1 of Ren, Fox and Bruce (2017). We then sample $\boldsymbol{\eta}^*$ given $\mathbf{x}$. By conjugacy, we sample the cluster-specific latent factor $\eta_{t,k}^*$ for time period $t = 1, \ldots, T$ and K existing clusters as follows:

$$\boldsymbol{\eta}_t^* | \boldsymbol{\lambda}, \mathbf{z}, \mathbf{x}, \mathbf{a}, \sigma_0^2 \sim \mathcal{N}_K \left\{ \begin{array}{c} \Omega \dfrac{1}{\sigma_0^2} (\Lambda \cdot Z)^T (\mathbf{x_t} - A\mathbf{x_{t-1}}), \\ \Omega = \left[I_K + \dfrac{1}{\sigma_0^2} (\Lambda \cdot Z)^T (\Lambda \cdot Z) \right]^{-1} \end{array} \right\}. \tag{4.2}$$

The derivation is provided in Supplement C.2 of Ren, Fox and Bruce (2017).

4.3. *Step* 3: *Sampling the dynamic model parameters*. Having sampled $\mathbf{x}$, we can form $\varepsilon_{t,i} = x_{t,i} - a_i x_{t-1,i}$ [see equation (3.3)]. Assuming in Step 1 we sampled $z_i = k$, we have T "covariate/response" pairs $(\eta_{t,k}^*, \varepsilon_{t,i})$ from which to inform the full conditional of λ_{ik}, as if it were the regression coefficient in a standard Bayesian regression model; see equation (3.5). Via conjugacy, this full conditional is a normal distribution specified in Supplement C.4 of Ren, Fox and Bruce (2017). For $j \neq k$, we sample λ_{ij} from its normal prior.

The full conditional from which a_i is sampled follows similarly via conjugacy: Combining equations (3.3) and (3.5), we can form "covariate/response" pairs $(x_{t-1,i}, x_{t,i} - \lambda_{iz_i} \eta_{t,z_i}^*)$, and treat a_i as the regression coefficient. The resulting full conditional (again a normal distribution) is in Supplement C.4 of Ren, Fox and Bruce (2017). For the variance σ_0^2, we have Tp "observations" $x_{t,i} - a_i x_{t-1,i} - \lambda_{iz_i} \eta_{t,z_i}^*$ distributed as $N(0, \sigma_0^2)$ that inform the inverse gamma full conditional specified in Supplement C.4 of Ren, Fox and Bruce (2017).

Finally, the emission parameters R and β can be sampled straightforwardly by treating $y_{t,i,l} - x_{t,i}$ as the response in a regression model with covariates $U_{l,h}$. The full conditionals, which are inverse gamma and normal distributions, respectively, are specified in Supplement C.5 of Ren, Fox and Bruce (2017).

4.4. *Step* 4: *Sampling hyperparameters*. The hyperparameters $\mu_\lambda, \sigma_\lambda^2, \mu_a, \sigma_a^2$ and μ_h, σ_h^2 for $h = 1, \ldots, H$ can be sampled straightforwardly via conjugacy results; see Supplement C.6 of Ren, Fox and Bruce (2017). We additionally assume a hyperprior for the DP concentration parameter $\alpha \sim \text{Gamma}(\alpha_\alpha, \beta_\alpha)$ and follow the sampling procedure of Escobar and West (1995); see Supplement C.7 of Ren, Fox and Bruce (2017).

4.5. *Computational challenges and strategies*. Although marginalizing $\boldsymbol{\pi}$, $\mathbf{x}$, and $\boldsymbol{\eta}^*$—that is, considering a *collapsed* sampler—reduces the dimensionality of the posterior we explore in our sampling, the marginalization of $\boldsymbol{\pi}$ induces dependencies between the cluster z_i. As such, we must rely on the CRP-based sequential sampling described in Section 4.1. Involved in this sampling is a computationally intensive likelihood evaluation. In particular, for each census tract i we must consider adding the tract to each existing cluster k, each of which involves a Kalman-filter-like algorithm. Naively, just harnessing the Woodbury matrix identity yields a computational complexity of $O((\min\{n^{(k)}, p^{(k)}\})^3 T)$, where $n^{(k)}$ is the maximum number of observations at any time t aggregated over census tracts in cluster k and $p^{(k)}$ is the number of census tracts in cluster k. In most cases, we have $n^{(k)} \gg p^{(k)}$.

To address the computational challenge of coupled z_i—which at first glance seems to imply reliance on single machine serial processing—we adopt the clever trick of Williamson, Dubey and Xing (2013) for parallel collapsed MCMC sampling in DP mixture models (DPMM). A similar approach was proposed by MacLaurin and Adams (2014). The conventional DPMM assumes that observations x_i with emission distribution $F(\cdot)$ are drawn as

$$\begin{aligned} G &\sim \text{DP}(\alpha, G_0), \\ \theta_i \mid G &\sim G, \\ x_i \mid \theta_i &\sim F(\theta_i). \end{aligned} \tag{4.3}$$

In order to do exact but parallel MCMC sampling for the DPMM on some P processors, Williamson, Dubey and Xing (2013) proposed the following auxiliary variable representation:

$$\begin{aligned} G_j &\sim \text{DP}(\alpha/P, G_0), \\ \boldsymbol{\phi} &\sim \text{Dirichlet}(\alpha/P, \ldots, \alpha/P), \end{aligned} \qquad \begin{aligned} \gamma_i \mid \phi &\sim \text{Multinomial}(\boldsymbol{\phi}), \\ \theta_i \mid G, \gamma_i &\sim G_{\gamma_i}, \\ x_i \mid \theta_i &\sim F(\theta_i). \end{aligned} \tag{4.4}$$

The auxiliary variable γ_i assigns data point i to processor γ_i. Williamson, Dubey and Xing (2013) proves that for $\boldsymbol{\phi}$ and G_j defined as in equation (4.4), $G := \sum_j \phi_j G_j \sim \mathrm{DP}(\sum_j \alpha/P, \frac{\sum_j (\alpha/P) G_0}{\sum_j \alpha/P}) = \mathrm{DP}(\alpha, G_0)$. Therefore, the marginal distributions for θ_i and x_i remain the same as in the original DPMM representation. Importantly, conditional on the processor allocations $\boldsymbol{\gamma}$, the data points are distributed as independent DPMMs on P machines, which enables independent sampling of cluster indicators in parallel. In our case, we leverage this auxiliary variable framework in order to allocate entire data streams to multiple machines. The resulting steps of parallel MCMC sampling of the cluster indicators z_i are described in Supplement D of Ren, Fox and Bruce (2017).

Beyond parallelizing the sampler, we also ameliorate the computational burden associated with the likelihood evaluations by deriving a simplified Kalman filter exploiting the specific structure of our model. In particular, for each data stream we only need two sufficient statistics $(\bar{\psi}_{t,i}, L_{t,i})$ instead of all of the house-level transactions, where $\psi_{t,i,l}$ is the adjusted sales price for the lth sale in tract i at time t after removing the hedonic effects. The sufficient statistic $\bar{\psi}_{t,i}$ is the mean of the adjusted individual sales prices and $L_{t,i}$ the number of sales for tract i at time t. We can think of the simplified Kalman filter as a filter with observation sequence given by the $p^{(k)}$-dimensional vector of mean sales prices for census tracts in that cluster. This algorithm then has complexity $O((p^{(k)})^3 T)$. Although the complexity of the algorithm has not changed (assuming $p^{(k)} < n^{(k)}$), the practical implementation details are simplified leading to significant runtime speedups. We experimented on empirical data that has one cluster of 21 census tracts, with 15,855 observations over 195 months. We repeat the likelihood evaluation 1000 times. The Kalman filter utilizing the Woodbury identity takes 499 seconds, while the simplified Kalman filter with sufficient statistics only takes 232 seconds, saving more than half of the compute time. This optimized Kalman filtering algorithm for performing likelihood evaluations using sufficient statistics is provided in Supplement B.2 of Ren, Fox and Bruce (2017).[3]

5. Model validation by simulation.

5.1. *Settings*. We first validate our model using simulated data with aspects set to match our real data analysis of Section 6. Specifically, we simulated 20 data streams corresponding to sales in 20 census tracts from January 1997 to September 2013, a period of 213 months. The 20 tracts are pre-assigned to four clusters of size 4, 4, 4 and 8 census tracts, respectively. First, we generated latent price processes, $x_{1:T,i}$, for each tract according to equations (3.3) and (3.5) (see Figure 5). Note that the tracts within each cluster have similar price dynamics, as intended by our

[3]Code is available at: https://github.com/shirleyuw/hyperlocalHouseIndex. See Supplement B.2 of Ren, Fox and Bruce (2017) for further details.

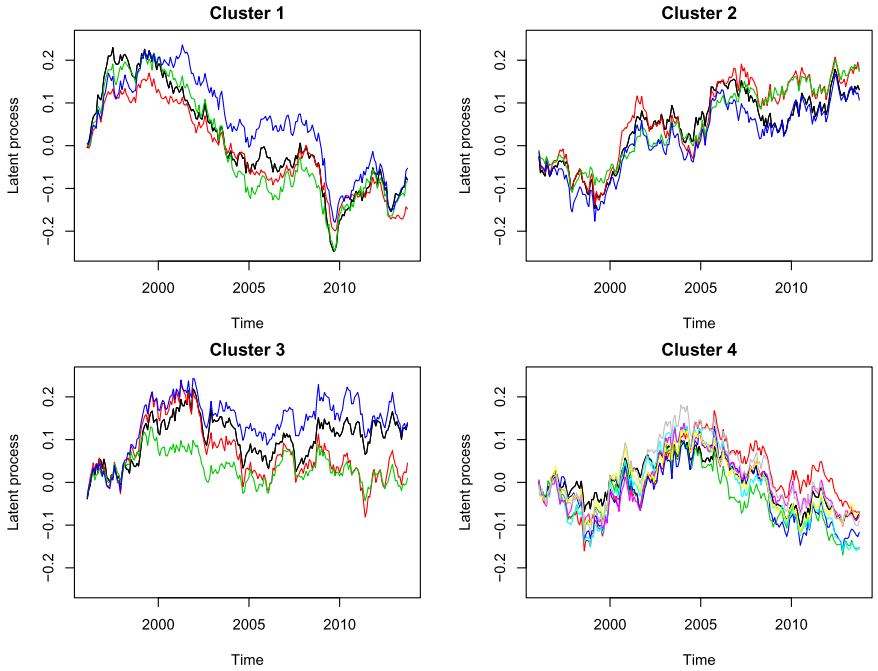

FIG. 5. *Simulated latent price processes for* 20 *census tracts from* 4 *clusters. Traces within each plot correspond to specific census tracts in each cluster.*

model. Second, we generated the observed sales prices, $y_{t,i,l}$, according to equation (3.4). The sales dates and house hedonics are taken from 20 randomly sampled tracts in the City of Seattle, so as to match the real-data frequency of observations and house characteristics. We repeat this process of generating latent price processes and house sales observations 50 times, resulting in 50 replicate time series. One replicate of generated sales prices is shown in Figure 6. For each replicate, the clustering structure and pattern of houses sold is kept fixed. See Supplement F.2 of Ren, Fox and Bruce (2017) for an experiment with a different clustering setup where all tracts are in one cluster and show that we can recover this structure.

5.2. *Results.* For each replicate, we ran the MCMC sampler for 1200 iterations. We used normalized Hamming distance to assess the clustering performance, which measures the proportion of tracts assigned to incorrect clusters after an optimal mapping of estimated to ground truth labels [Munkres (1957)]. Figure 7 demonstrates that we successfully recover the underlying clusters, with trends indicating that our sampler converges very rapidly. As further evidence of convergence, we ran three chains with different initializations, and the scale reduction factor of Gelman and Rubin (1992) indicated convergence of the chains after 1200 iterations.

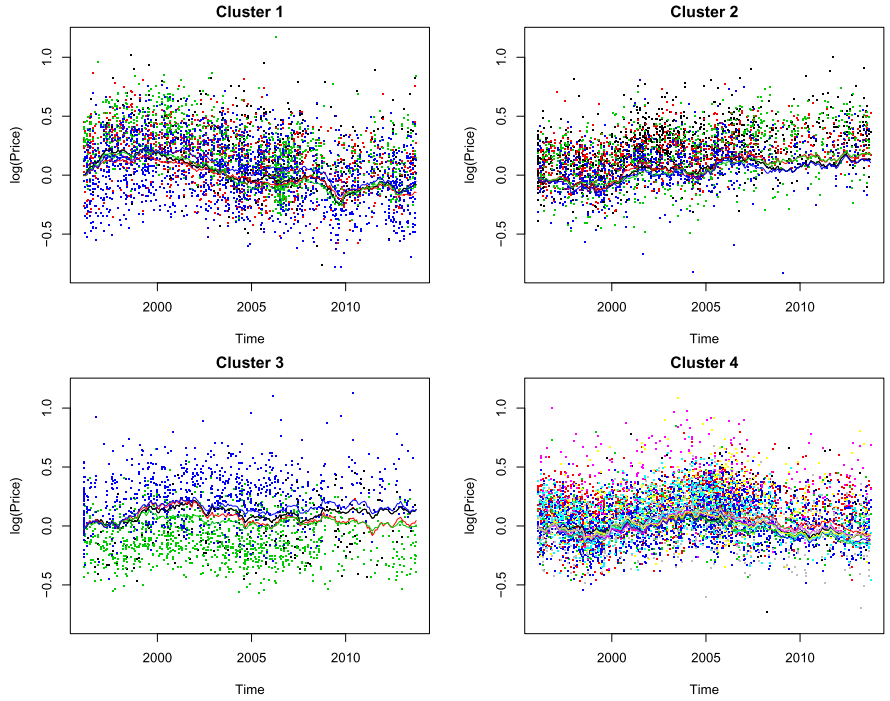

FIG. 6. *For a randomly selected replicate, simulated latent processes* (solid lines) *and sales prices* (dots) *for the* 20 *clustered census tracts for each of the* 4 *ground truth clusters.*

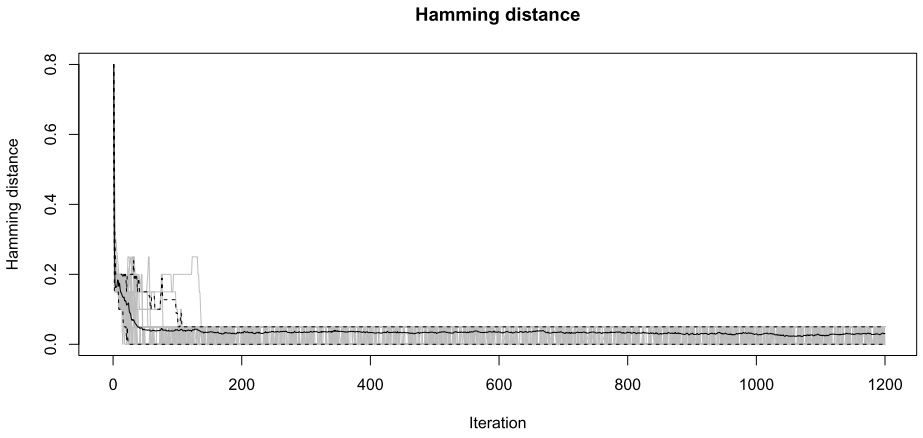

FIG. 7. *For each replicate* (*gray traces*), *normalized Hamming distance between posterior samples of cluster indicators and true cluster memberships* (*after an optimal mapping*) *as a function of Gibbs iteration. The mean and* 95% *intervals are indicated in black and dashed-black lines, respectively.*

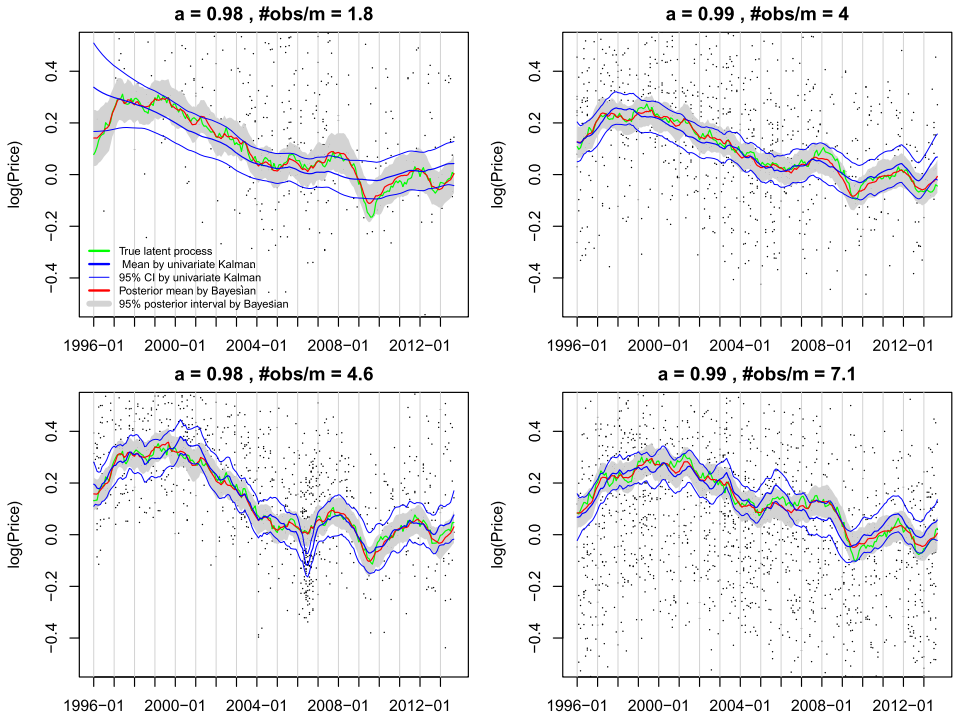

FIG. 8. *For the selected replicate of Figure 6, plots of the estimated intrinsic price dynamics relative to the true* $x_{t,i}$ (*green*) *for the* 4 *census tracts in Cluster* 1. *Compare the posterior mean* (*red*) *and* 95% *posterior intervals* (*shaded gray*) *of our proposed model to the independent Kalman-smoother-within-EM baseline approach* (*blue*), *which performs poorly when the number of observations per month* (*average indicated by #obs/m*) *is low.*

Given sparse (simulated) observations per month at the census tract level, Figure 8 demonstrates that our posterior estimate of the intrinsic price dynamics nicely tracks the true latent dynamics for each tract. As a baseline comparison, we considered applying a Kalman-smoother-within-EM algorithm independently on each tract, as in our exploratory data analysis of Section 2. Unsurprisingly, without sharing observations from similar tracts, the baseline approach fails when the observations are sparse; see Supplement F.2 of Ren, Fox and Bruce (2017) for results on the other census tracts.

An alternative baseline is our hierarchical Bayesian dynamical model, but assuming each tract is in its own cluster. Implementation-wise, we simply fix $z_i = i$ and do not resample these cluster indicators in our MCMC. Figure 9 shows the RMSE for the estimated latent trends $\mathbf{x}$ as a function of the number of observations in the census tract. For tracts with fewer observations, the clustering method provides a substantial reduction in error. As expected, when observations are abundant, the improvement diminishes. These experiments confirm the benefits of the

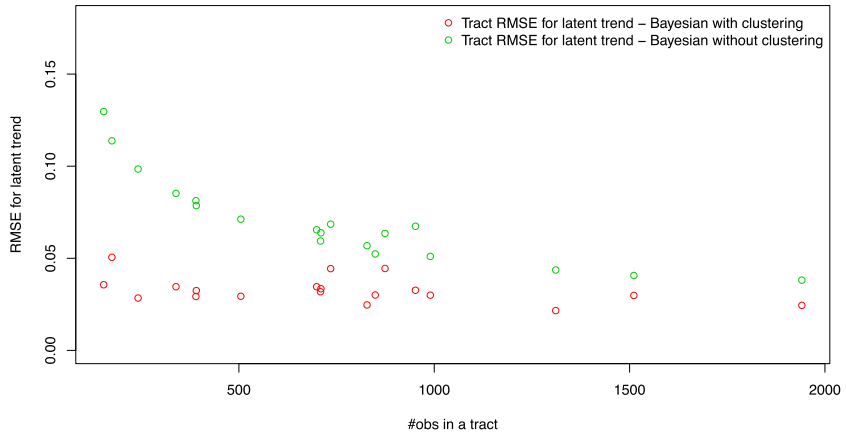

FIG. 9. *For the replicate of Figure 6, RMSE of estimated latent trend per tract using clustering (red) or no clustering (green) as a function of the number of observations.*

DP-based clustering beyond just hierarchical modeling in structured, data-scarce scenarios.

We also experimented with other simulation scenarios, summarized in Table 2. When the latent factor processes have relatively large factor loadings (large μ_λ) leading to large noise variance on the intrinsic price dynamics, the benefits of using clustering for predicting latent trends $\mathbf{x}$ are very significant compared to the model without clustering. However, even under such scenarios, the improvement in predicting the log sales prices $y_{t,i,l}$ themselves is not as large since the hedonic effects dominate the observed price. Importantly, we note that *house-level prediction is not our goal*; instead we are interested in the intrinsic price dynamics $\mathbf{x}$ themselves, which form our fine-resolution index.

TABLE 2

For three simulated scenarios and 50 *replicates per scenario, results on out-of-sample prediction of latent trends $x_{1:T,i}$ and house prices $y_{t,i,l}$. We compare our proposed Bayesian model both with and without the DP-based nonparametric clustering component*

		No clustering	**Clustering**	**Improvement**	
		(Mean)	**(Mean)**	**(Mean)**	**(95% interval)**
$\mu_a = 0.99$	RMSE in x	0.0258	0.0235	8.7%	[6.1%, 11.4%]
$\mu_\lambda = 0.015$	RMSE in y	0.1032	0.1029	1.2%	[−0.6%, 2.9%]
$\mu_a = 0.99$	RMSE in x	0.0747	0.0348	53.5%	[51.0%, 56.0%]
$\mu_\lambda = 0.15$	RMSE in y	0.1147	0.1051	12.8%	[9.9%, 15.7%]
$\mu_a = 0.60$	RMSE in x	0.0800	0.0333	58.4%	[56.0%, 61.0%]
$\mu_\lambda = 0.15$	RMSE in y	0.1155	0.1060	11.9%	[7.5%, 18.0%]

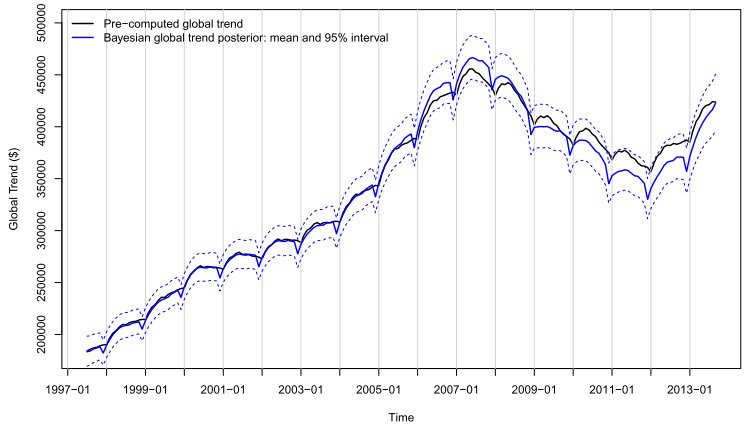

FIG. 10. *Estimated global trend* (*black line*) *using the seasonality decomposition approach of Cleveland et al.* (1990), *after adjusting for hedonic effects. See Supplement* G *of Ren, Fox and Bruce* (2017) *for further details. The posterior mean* (*blue*, *solid*) *and* 95% *credible intervals* (*blue*, *dashed*) *for our jointly estimated trend* (*see Section* 7) *are shown for comparison.*

6. Housing data analysis. We now turn to our housing data analysis based on the City of Seattle data described in Section 2. To focus on our main modeling contributions, here we assume the global trend is separately estimated and removed as a preprocessing step. Computing a fairly good estimate of a global trend is relatively straightforward since we have sufficient data in aggregate. The estimated global trend is shown in Figure 10, with details in Supplement G of Ren, Fox and Bruce (2017). We notice a small but significant seasonal effect, which can be mostly attributed to the changing supply of houses during the year: very few homes are listed in November and December so that transactions that occur in that period are leftover inventory or have other special circumstances. In Section 7, we return to joint estimation of the global trend to properly account for uncertainty in this estimate.

To assess our model, we randomly split the sales *per census tract* into a 75% training and 25% test sets. On the training set, we ran three MCMC chains for 15,000 iterations from different initial values, discarding the first half as burn-in and thinning the remaining samples by 5. We used the scale reduction factor of Gelman and Rubin (1992) to check for convergence.

Figure 11 provides an illustration of the resulting 16 census tract clusters associated with the maximum a posteriori (MAP) sample (i.e., the sample with largest joint probability). The log intrinsic price dynamics associated with each of these clusters, averaged over census tracts assigned to the cluster, are shown in Figure 12. Clusters 15 and 16 have the most dramatic trend. They include census tracts from the downtown Seattle area where the houses are almost exclusively condos and have unique supply and demand dynamics. Clusters 11 and 13 are mostly low-income areas with less expensive housing where the housing recovery has been

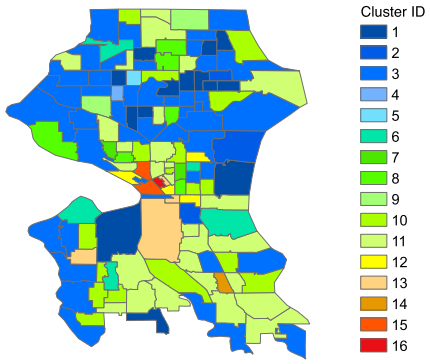

FIG. 11. *Map of cluster assignments under the MAP sample. Cluster labels and associated colors are ordered based on the deviance of the cluster's average* (*across tracts*) *latent trend from the global trend, with blue* (1) *representing smallest and red* (16) *largest.*

slower. The biggest difference between the clusters occurs during the 2006–2012 time period which spanned the housing boom followed by the bust. Intuitively, different regions were affected differently by this highly volatile period. Supplement G of Ren, Fox and Bruce (2017) shows the cluster average index in raw price scale.

For the MAP clustering depicted in Figure 11, the University District (U-District) census tract highlighted in Section 1 gets assigned to Cluster 3—the largest cluster—driven by "the rich get richer" property of the CRP prior. However, when examining all collected posterior samples, 57% of the time the U-District does not share a cluster with *any* of its neighbors and 86% of the time it does not share a cluster with more than one neighbor. The lack of a hard-coded spatial structure in our model is what enables such heterogeneous spatial effects to appear; instead, our DP-based cluster model allows for a flexible dependence structure by discovering regions with correlated price dynamic patterns. Importantly, in forming our index, we average over the uncertainty in the clustering structure. In particular, instead of using a single sample as in the visualizations of Figures 11–12, we compute the posterior mean trend per tract by averaging the tract-specific intrinsic price dynamics across MCMC samples.

6.1. *Comparison with other methods.* We compared our Bayesian nonparametric approach with the Case–Shiller housing index [Case and Shiller (1987)] described in Section 1. Even though our goal is not house-level prediction, it is one metric by which we can assess our fit. Since the Case–Shiller method is based on repeat sales only and does not include hedonics, it is not well suited to predicting house-level prices. In order to fairly compare our approach with Case–Shiller, we treated the Case–Shiller index as the intrinsic price process *with the global trend* $\tilde{\mathbf{x}}$ in our model, and then fit a regression model with tract-specific hedonic effects as in equation (3.4) using observations $\tilde{y}_{t,i,l}$. The estimated hedonic effects

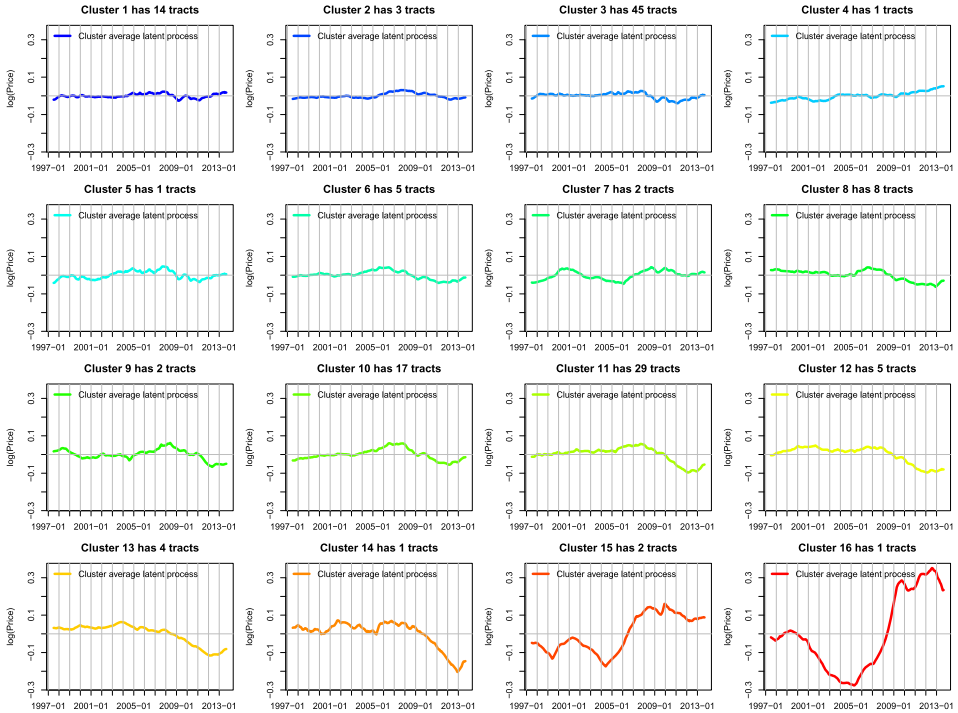

FIG. 12. *For the MAP sample, cluster-average intrinsic price dynamics computed by averaging* $\mathbf{x}_{1:T,i}$ *over all* i *with* $z_i = k$ *for* $k = 1, \ldots, 16$. *The color scheme is as in Figure* 11.

together with Case–Shiller index are then used to predict the house prices. The Case–Shiller index also does not model seasonal effects, so for a fair comparison we remove the seasonal component from our estimated global trend and simply use the resulting smooth trend when forming our predictions. We note that all of our predictive performance numbers are better *with* the seasonal component, as displayed in Table 7.

Due to the scarcity of repeat sales localized at tract level, the Case–Shiller index can only be computed at 8 of the 140 tracts. To maintain a tract-level comparison, if the Case–Shiller index is not available for a given tract, we continue up the spatial hierarchy examining zip code and city levels until there is a computable index that can serve as $x_{t,i}$ in our prediction; that is, we use the finest resolution Case–Shiller index available at any house location to predict house prices. In Table 3, we summarize the number of house-level predictions that are based on the Case–Shiller city, zip code or tract level indices; we also include the number of tracts for which our analyses relied on city and zip code levels, or were able to use tract-level indices directly.

Our Bayesian model can successfully produce value indices for all tracts. To predict house-level prices, we use the posterior predictive distribution approxi-

TABLE 3

For our predictive performance comparison summarized in Table 4, the number of tracts and individual houses (in test set) that rely on using city, zip code or tract-level indices with the Case–Shiller method. Our Bayesian method always uses a tract-level index

	Case–Shiller			Bayesian
	City	**Zip code**	**Census tract**	**Census tract**
# tracts using	11	121	8	140
# observations using	1294	26,576	3248	31,118

mated by our MCMC posterior samples:

$$P\big(y^*_{t,i}|\mathbf{Y}\big) = \int_\theta P\big(y^*_{t,i}|\theta\big) P(\theta|\mathbf{Y})\, d\theta \approx \sum_{m=1}^{M} p\big(y^*_{t,i} \mid \theta^{(m)}\big), \tag{6.1}$$

where y^* is the new data point, $\mathbf{Y}$ denotes the training data and θ represents parameters with $\theta^{(m)}$ the mth MCMC sample. Since $p(y^*_{t,i} \mid \theta^{(m)})$ does not have an analytic form, we simulate a set of $y^*_{t,i}$ for each $\theta^{(m)}$ using equation (3.4). We then use the mean of these posterior predictive samples as the prediction for any house in the test set.

For all of our comparisons, we used the same training and test split. In Table 4, we summarize the out-of-sample predictive performance with five metrics: root mean squared error (RMSE) in price, mean/median/90th percentile of absolute percentage error (Mean APE, Median APE, 90th APE), and the popular industry metric of proportion of house sales within 10% error (P10). Importantly, we highlight again that house sales predictions are largely hedonics driven. Since we constructed all methods using the same hedonics model, we do not expect to see large differences in numbers. Regardless, we see notable improvements using our proposed index, with uniformly better predictive performance as compared to the Case–Shiller index at the finest resolution available. Over all houses in the test set, our method has an 11.2% improvement in RMSE and about 5% improvement in other metrics.

We then look at the breakdown of the analysis by deviation of the inferred latent trend from the global trend. For the 5% of census tracts with the most dramatic intrinsic price dynamics (as measured by L2 norm of posterior mean latent trend over time), we see even larger improvements over Case–Shiller: a 15.5% decrease in RMSE and 21.7% in 90th percentile APE. The latter measure indicates a significant reduction in the tail of the error distribution; that is, not only are we better able to capture these more volatile tracts, we are also having the most dramatic improvements on the hardest-to-predict houses. These effects can be explained as follows. By not hard-coding spatial relationships via adjacencies of tracts, we see in Figure 1 that certain regions (e.g., the U-District) do not get shrunk to trends in

TABLE 4

Predictive performance comparison of index methods using various measures: *root mean squared error* (*RMSE*), *mean absolute percentage error* (*Mean APE*), *median absolute percentage error* (*Median APE*), 90*th percentile absolute percentage error* (90*th APE*) *and proportion within* 10% *error* (*P*10)

	Case–Shiller index at finest resolution w/tract hedonic effects	**Bayesian index at census tract level**	**Improvement**
All observations in test set (31,118 data points)			
RMSE	137,600	122,139	11.2%
Mean APE	0.1734	0.1636	5.6%
Median APE	0.1294	0.1236	4.5%
90th APE	0.3607	0.3427	5.0%
P10	0.3985	0.4190	5.1%
Top 5% tracts with most dramatic latent trends (1111 data points)			
RMSE	91,627	77,399	15.5%
Mean APE	0.2045	0.1748	14.5%
Median APE	0.1403	0.1259	10.3%
90th APE	0.4699	0.3679	21.7%
P10	0.3816	0.4113	7.8%

neighboring tracts. At the same time, our hierarchical Bayesian model with clustering still enables sharing of information to improve estimates, as we see in Table 4. It is not surprising to see the most significant improvements for the most highly volatile tracts: These are the tracts for which providing a robust fine-scale index is so important in order to capture the deviation from the global trend.

Table 5 lists the improvement in predictive performance of our Bayesian tract index over the Case–Shiller index when the latter index is computed at a city or zip code level. The analysis is further broken down by the level of data scarcity in the tract. Not surprisingly, the most significant improvement is for houses in tracts with fewer sales; for example, we see a 16% improvement in 90th percentile APE for these data-scarce tracts, for which the tail of the error distribution is important and hard to characterize. We might expect that our method provides less improvement over the Case–Shiller index at the zip code than city level. Interestingly, as the spatial resolution goes finer from city to zip code level, the Case–Shiller index suffers from worse predictive performance in most cases. This result validates that this popular index method is ill-suited to the task of constructing a housing index for small regions where transactions are scarce.

We now examine the impact of our Bayesian model over alternative approaches using the same tract-specific dynamical model of equations (3.3)–(3.4), both with and without hedonics. In particular, we compare against a Kalman-smoother-within-EM algorithm applied *independently* to each census tract (again, as in Section 2). The results are summarized in Table 6. (Note that the last column of Table 6

TABLE 5
Predictive improvement of our Bayesian tract index over Case–Shiller city and zip code indices for tracts of different sales frequency using Mean APE and 90th APE metrics

	Improvement over Case–Shiller city index	**Improvement over Case–Shiller zip code index**
Top 5% tracts with most sales (3569 data points)		
Mean APE	3.1%	4.8%
90th APE	1.2%	2.9%
Middle 50% tracts (14,507 data points)		
Mean APE	4.6%	7.2%
90th APE	5.1%	7.1%
Lower 5% tracts with least sales (188 data points)		
Mean APE	8.5%	5.4%
90th APE	15.5%	16.0%

coincides with the second column of Table 4, and is repeated for readability.) We see the benefit of incorporating hedonics, but that gain is actually not as large as the improvements seen from our Bayesian approach to joint modeling of the tracts, despite the clear importance of hedonics in driving house predictions. Additionally, as motivated by the results of Table 2, we would expect even larger improvements in the estimation of the target index $\mathbf{x}$, though such an evaluation is not feasible here since we do not have the true index value.

6.2. *Qualitative assessment of the indices*. We now turn to the central focus of the paper and assess the quality of the index itself. Since there is no ground truth or direct performance metric, we use the Zillow Home Value Index (ZHVI®) [Zillow (2014)] as a reference for comparison of the indices. The ZHVI is a bottom up, empirical approach to computing an index:

TABLE 6
Predictive performance comparison of independent Kalman-smoother variants to the proposed Bayesian nonparametric model using the same metrics as in Table 4

	Univariate Kalman smoother without hedonics	**Univariate Kalman smoother with hedonics**	**Bayesian clustering**
RMSE	262,075	194,562	122,139
Mean APE	0.3698	0.2746	0.1636
Median APE	0.2854	0.2238	0.1236
90th APE	0.7634	0.5584	0.3427
P10	0.1907	0.2346	0.4190

1. Using recent house sales data, the value of each home is estimated using a hedonic regression model (known popularly as a Zestimate®).
2. The ZHVI for a given region is defined as the median of the Zestimate for all homes in a region.

ZHVI is appealing due to its straightforward and intuitive nature. Unlike weighted repeat sales methods, the ZHVI is not impacted by the changing composition in types of homes that are sold over different periods of time. In addition, the ZHVI is stable for even very small geographic regions, such as a census tract. The ZHVI can be viewed, informally, as a semi-supervised approach: the index is based on all the data by inferring the value from the smaller set of labeled data (house transactions). The Zestimate is based on a proprietary, sophisticated regression model using a variety of data sources including user-edited data, house listings, tax assessments, transactions, parcel information and geographical features (e.g., proximity to waterfront). The Zestimate is unbiased and in Seattle, the median absolute error of the Zestimate is around 6% (see http://www.zillow.com/zestimate/#acc for published accuracy of the Zestimate). Due to the high accuracy of the Zestimate, and the empirical nature of the approach, the ZHVI provides a valuable basis against which to compare the Bayesian and Case–Shiller indices.[4]

In addition to comparing our Bayesian index to Case–Shiller, we also consider a model in which the the DP-based clustering is removed, treating each census tract as its own cluster, as in Section 5.2. Recall that this model still represents a hierarchical Bayesian dynamic model. Since the Case–Shiller method is not computable for most of the census tracts, we focus our analysis at the zip code level. For the Bayesian index with or without DP-based nonparametric clustering, the zip code index is constructed by averaging the census tract indices within a zip code.

Figure 13 shows that the Bayesian indices with and without the DP-based nonparametric clustering component have significantly different performance during the 2006–2007 period, and to a lesser extent in 2010–2011. In 2006–2007, the Seattle housing market was in a boom period with high sales and volatility [see Figure 6 in Supplement G.3 of Ren, Fox and Bruce (2017)]. After the bust, the housing market started to stabilize in 2010–2011. The market boom and subsequent stabilization were manifested in the different housing sectors in disparate ways. The DP clustering-based index is more closely aligned with the ZHVI, especially in the highly volatile year of 2007, since it is better able to capture the dynamics of the change in value for different housing sectors. This is because the nonclustering Bayesian hierarchical model shrinks the census tracts with few observations toward a global mean, whereas our clustering model allows atypical census tracts to be shrunk toward a more informed structure, such as the one shown in Figure 12.

[4]One of the motivations for this paper is to achieve the properties of the ZHVI using a rigorous, model-based approach, which opens up new opportunities for statistical inferences and further model development.

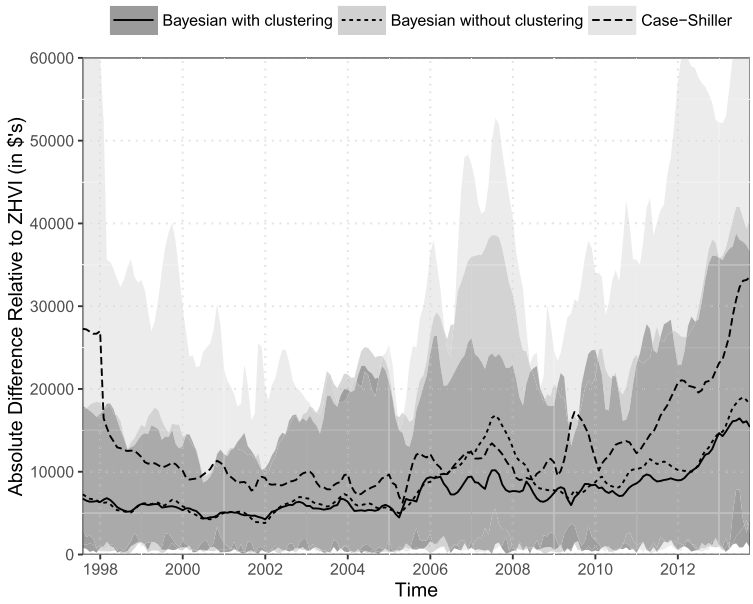

FIG. 13. *Differences of various index methods relative to the ZHVI at the zip code level. Examining performance across zip codes, the mean absolute difference* (black line) *and* 90% *interval* (shaded dark gray) *of our proposed Bayesian index is compared to that of the Bayesian index without clustering* (short dash *and* shaded medium gray) *and the Case–Shiller zip code index* (long dash *and* shaded light gray). *The differences for the Bayesian methods are based on posterior mean estimates. A* 3*-panel figure separating these components is in Supplement* G.4 *of Ren, Fox and Bruce* (2017).

Figure 13 also compares the zip code Case–Shiller index, which differs much more significantly from the ZHVI at all times than the proposed Bayesian index. Without any kind of sharing information and shrinkage across different regions, the Case–Shiller index has the widest interval among the three methods. Furthermore, the beginning and the end of the study periods are extremely challenging for Case–Shiller index due to lack of repeated sales available at these boundaries. In the middle of the series, during the highly volatile period of 2007, the difference between Case–Shiller and the ZHVI is especially large.

Digging into this volatile 2007 period, Figure 14 provides a more detailed comparison of the differences in Figure 13 during 2007. We see that Case–Shiller has a long-tailed distribution of absolute differences of individual zip code indices relative to ZHVI, in addition to a hump at this right tail. The shrinkage provided by the other two Bayesian methods leads to much lighter tails and removes this high-error hump, with the clustering approach clearly the closest match to ZHVI. In particular, looking at the cumulative distribution of Figure 14(b), we see that the Bayesian model *without clustering* has a lighter tail than Case–Shiller, improving these outlying estimates via shrinkage induced by the hierarchical Bayesian model. However, this model also has fewer small-error zip codes relative to the

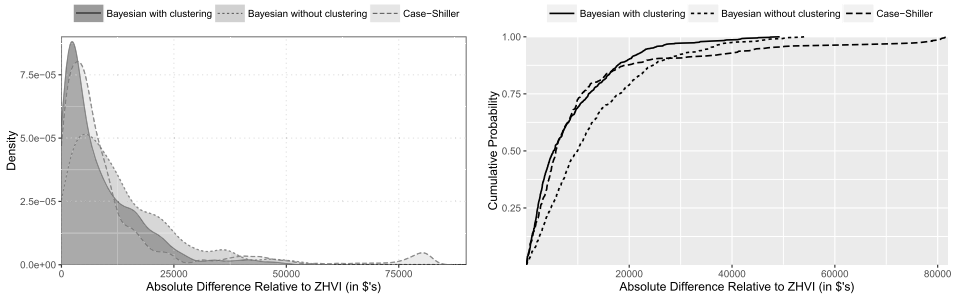

FIG. 14. *A more detailed examination of the differences in Figure* 13 *during* 2007. (Left) *Estimated density and* (right) *cumulative distribution of the absolute differences. A* 3*-panel figure separating the components is in Supplement* G.4 *of Ren, Fox and Bruce* (2017).

Case–Shiller baseline. By contrast, our proposed Bayesian nonparametric clustering index has as many small-error zip codes as Case–Shiller, but also has many fewer large-error zip codes than either of the comparison methods. Thus, we see the importance not only of a hierarchical Bayesian approach, but one that leverages structured relationships between regions.

7. Joint model with the global trend. Instead of a fixed, pre-calculated global trend extracted from the data as a preprocessing step, in this section we propose to jointly model and estimate the nonstationary global market trend with the stationary local price dynamics. This unified Bayesian framework properly accounts for all parameter uncertainties jointly, including uncertainty in the global trend. We describe the global trend model, modification to the MCMC, and associated housing results below. Further details on the global trend model selection, prior specification and posterior computations are in Supplement H of Ren, Fox and Bruce (2017).

7.1. *A nonstationary global trend model.* Similar to many economic time series, the overall housing market trend is nonstationary; this is clearly demonstrated in the estimated Seattle City global trend of Figure 10. The time series clustering model of Nieto-Barajas and Contreras-Cristán (2014) described in Section 3 models the nonstationary trend with a quadratic form. Such simple polynomial forms may not be flexible enough to capture long-term housing market trends such as boom, bust and recovery periods. To promote smoothness while allowing flexibility to capture the complex global market dynamics, we instead use natural cubic splines (NCS) [Smith (1979)] with monthly effects. In particular, the natural cubic splines interpolation process specifies n_B interior knots, which generates $N_B = n_B + 2$ basis functions including an intercept, piecewise cubic splines between knots and linear splines at the boundaries. More specifically, we propose the following model for the nonstationary global trend:

$$g_t = w_1 B_1(t) + \cdots + w_{N_B} B_{N_B}(t) + s_2 m_2(t) + \cdots + s_{12} m_{12}(t), \tag{7.1}$$

where B_j for $j = 1, \ldots, N_B$ are the basis functions and $m_j(t) = I(t = j)$ for $j = 2, \ldots, 12$ denote the monthly effects.

7.2. *Modification to the MCMC.* Recall the definition $\tilde{x}_{t,i} = g_t + x_{t,i}$. Based on an sample of the intrinsic price dynamics $x_{t,i}$ and hedonic effects $\beta_{i,1:H}$ at a given iteration of our MCMC described in Section 4, we can use our observation model of equation (3.2)—combined with the simple linear model for the hedonic effects in equation (3.4)—to define pseudo-observations:

$$r_{t,i,l} = \tilde{y}_{t,i,l} - x_{t,i,l} - \sum_{h=1}^{H} \beta_{i,h} U_{l,h} \tag{7.2}$$

such that

$$r_{t,i,l} = g_t + v_{t,i,l}, \qquad v_{t,i,l} \sim N(0, R_i), \tag{7.3}$$

that is, $r_{t,i,l}$ provides a noisy observation of the global trend since it represents the residual of each house sales observation (log price) after accounting for the local intrinsic dynamics and house hedonics. Equation (7.3) simply represents a standard regression setting with time as the predictor; that is, from data points $(t, r_{t,i,l})$, the NCS-based global trend can be fit straightforwardly if the number of knots is known. We use a model selection procedure based on Bayesian Information Criterion (BIC), as outlined in Supplement H of Ren, Fox and Bruce (2017), which suggests using 9 interior knots.

The modification to the overall MCMC is then as follows. Examining the MCMC overview at the beginning of Section 4, we simply condition on g_t and $\tilde{\mathbf{y}}$ in place of $\mathbf{y}$ in Steps 1–3. Those steps remain identical since $y_{t,i,l}$ is computed as $\tilde{y}_{t,i,l} - g_t$. We then add a step after Step 3 to sample g_t given everything else, as described above [i.e., fitting a NCS to pseudo-observations $(t, r_{t,i,l})$]. The final step of the MCMC (Step 4 previously), remains unchanged. Further details are in Supplement H of Ren, Fox and Bruce (2017).

7.3. *Housing data analysis with joint inference of global trend.* We now examine the impact of jointly estimating the global trend together with the other latent variables and model parameters. Figure 15 shows the posterior of the smoothed Bayesian global trend (after removing the monthly effects), along with the Case–Shiller index and ZHVI computed at the Seattle City level. The posterior mean of our global trend is more in alliance with ZHVI, which we believe provides a better estimate than the Case–Shiller index; ZHVI is calculated based on monthly-estimated house values for all houses in Seattle, whereas the Case–Shiller index is computed using only repeat sales data. The posterior global trend *with* monthly effects is shown in Figure 10. The predictions computed below include these monthly effects.

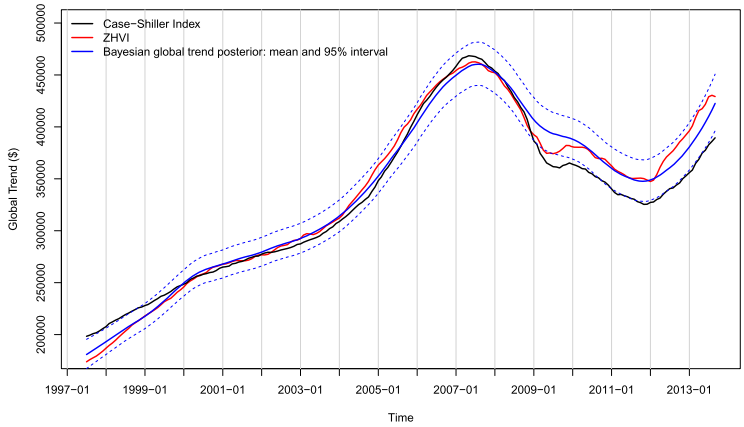

FIG. 15. *Comparison of the Bayesian global trend posterior with Case–Shiller index and ZHVI computed at the Seattle City level. For the Bayesian trend, the seasonal component depicted in Figure 10 is removed for direct comparison* (*since ZHVI is not available with a seasonal component*).

We compare the predictive performance of the full method that jointly estimates the global trend with the method of Section 6 that uses a pre-computed global trend. Table 7 shows that the predictive performance associated with the two models are quite close according to all metrics.

8. Discussion. We presented a method for constructing a housing index at fine-scale geographical units, with better space–time adjustment and specificity than existing approaches. In particular, the extreme scarcity of transactions at a fine spatiotemporal granularity poses a significant modeling challenge. Our proposed dynamical model utilizes a Bayesian nonparametric approach for flexible structure learning to correlate regions that share similar underlying price dynamics. This model leverages information from the region-specific time series within

TABLE 7
Predictive performance comparison of the Bayesian housing index using a precalculated global trend, as in Section 6, with the global trend joint estimation of this section

	Bayesian index		
	Pre-calculated global trend	**Global trend joint estimation**	**Change**
RMSE	122,026	122,083	0.05%
Mean APE	0.1633	0.1635	0.12%
Median APE	0.1231	0.1237	0.49%
90th APE	0.3422	0.3414	−0.23%
P10	0.4183	0.4198	0.35%

TABLE 8
Predictive performance of the Bayesian nonparametric housing index computed at the census tract lever versus the finer-granularity neighborhood level. See Supplement G.5 *of Ren, Fox and Bruce* (2017)

	Bayesian	
	Census tract index	**Neighborhood index**
RMSE	122,026	120,198
Mean APE	0.1633	0.1565
Median APE	0.1231	0.1165
90th APE	0.3422	0.3208
P10	0.4183	0.4392

the discovered clusters, providing a form of multiple shrinkage of individual trend estimates for each region.

We demonstrated that our methods provided a reliable monthly housing index at the census tract level. Furthermore, our methods can robustly scale to even finer spatial granularities. For example, using heuristically defined neighborhoods at a finer spatial granularity than census tracts [see Supplement G.5 of Ren, Fox and Bruce (2017)], our predictive performance can even *improve*. The results are summarized in Table 8. This is in contrast to the worsening performance of the Case–Shiller index with increasing spatial granularity, as demonstrated in Table 5.

Our clustering-based dynamical model avoids a reliance on repeated sales, providing an ability to track price changes in local housing markets. In contrast, constrained by few observations of multiple sales for the same house, classic repeat sales methods are usually only robustly estimated over larger regions, such as zip code or city, which may lack spatial specificity. Although sole reliance on repeated sales can be problematic for the reasons described above, one could imagine incorporating a similar idea within our model via a longitudinal trend for the same house in the model. Other extensions include considering longer memory processes with a higher order autoregressive model for the latent trend. We could also add side information, such as crime rate, road network information and school district ratings, to better inform the clusters of local areas. Finally, one could consider a prespecified geographic model combined with our cluster-induced heterogeneous spatial structure as a model of the residuals.

Acknowledgments. We would like to thank Stan Humphries, Yeng Bun, Bill Constantine, Dong Xiang and Chunyi Wang at Zillow for helpful discussions and guidance on the data.

SUPPLEMENTARY MATERIAL

Supplement to "Clustering correlated, sparse data streams to estimate a localized housing price index" (DOI: 10.1214/17-AOAS1019SUPP; .pdf). We

detail aspects of our MCMC sampler, including: (i) the required likelihood calculation via Kalman filtering variants and (ii) a parallel implementation of sampling the cluster memberships. We also include further synthetic data experiments and results from our Seattle City analysis, and specify the various settings used in our experiments. Finally, we provide additional details on our model selection, specification, and computations for the joint global trend analysis. A link to our code base and related housing data sources is included.

REFERENCES

BAILEY, M. J., MUTH, R. F. and NOURSE, H. O. (1963). A Regression Method for Real Estate Price Index Construction. *J. Amer. Statist. Assoc.* **58** 933–942.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

BRUNAUER, W., LANG, S. and UMLAUF, N. (2013). Modelling house prices using multilevel structured additive regression. *Stat. Model.* **13** 95–123. MR3179520

CASE, B. and QUIGLEY, J. M. (1991). The dynamics of real estate prices. *Rev. Econ. Stat.* **73** 50–58.

CASE, K. E. and SHILLER, R. J. (1987). Prices of single family homes since 1970: New indexes for four cities. *N. Engl. Econ. Rev.* 45–56.

CASE, K. E. and SHILLER, R. J. (1989). The efficiency of the market for single-family homes. *Amer. Econ. Rev.* **79** 125–137.

CLEVELAND, R. B., CLEVELAND, W. S., MCRAE, J. E. and TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on loess (with discussion). *J. Off. Stat.* **6** 3–73.

ENGLUND, P., QUIGLEY, J. M. and REDFEARN, C. L. (1999). The choice of methodology for computing housing price indexes: Comparisons of temporal aggregation and sample definition. *J. Real Estate Finance Econ.* **19** 91–112.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

GATZLAFF, D. H. and HAURIN, D. R. (1997). Sample selection bias and repeat-sales index estimates. *J. Real Estate Finance Econ.* **14** 33–50.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GOETZMANN, W. N. and PENG, L. (2002). The bias of the RSR estimator and the accuracy of some alternatives. *Real Estate Econ.* **30** 13–39.

IACOVIELLO, M. (2011). Housing wealth and consumption. Board of Governors of the Federal Reserve System, International Finance Discussion Papers 1027.

LIAO, T. W. (2005). Clustering of time series data—A survey. *Pattern Recognit.* **38** 1857–1874.

MACLAURIN, D. and ADAMS, R. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Uncertainty in Artificial Intelligence*.

MEESE, R. A. and WALLACE, N. E. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *J. Real Estate Finance Econ.* **14** 51–73.

MUNKRES, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Indust. Appl. Math.* **5** 32–38. MR0093429

NAGARAJA, C. H., BROWN, L. D. and ZHAO, L. H. (2011). An autoregressive approach to house price modeling. *Ann. Appl. Stat.* **5** 124–149. MR2810392

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

NIETO-BARAJAS, L. E. and CONTRERAS-CRISTÁN, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Anal.* **9** 147–169. MR3188303

PALLA, K., GHAHRAMANI, Z. and KNOWLES, D. (2012). A nonparametric variable clustering model. *Adv. Neural Inf. Process. Syst.* **25** 2987–2995.

PITMAN, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

REN, Y., FOX, E. B. and BRUCE, A. (2017). Supplement to "Clustering correlated, sparse data streams to estimate a localized housing price index." DOI:10.1214/17-AOAS1019SUPP.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SHILLER, R. (1991). Arithmetic repeat sales price estimators. *J. Housing Econ.* **1** 110–126.

SMITH, P. L. (1979). Splines as a useful and convenient statistical tool. *Amer. Statist.* **33** 57–62.

WILLIAMSON, S., DUBEY, A. and XING, E. (2013). Parallel Markov chain Monte Carlo for nonparametric mixture models. In *International Conference on Machine Learning* 98–106.

ZILLOW (2014). Zillow home value index: Methodology. http://www.zillow.com/research/zhvi-methodology-6032/.

Y. REN
E. B. FOX
A. BRUCE
DEPARTMENT OF STATISTICS
BOX 354322
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195
USA
E-MAIL: shirleyr@uw.edu
ebfox@stat.washington.edu
andrewb0@uw.edu