Bayesian Modeling of a High Resolution Housing Price Index

You Ren

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Emily B. Fox, Chair

Andrew Bruce

Vladimir Minin

Michael G. Yost

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bayesian Modeling of a High Resolution Housing Price Index

You Ren

Chair of the Supervisory Committee:
Professor Emily B. Fox
Statistics

Understanding how housing values evolve over time is important to consumers, real estate professionals, and policy makers. Existing methods for constructing housing indices are computed at a coarse spatial granularity, such as metropolitan regions. This coarse granularity does not have the representative power to encode the fine price dynamics apparent in local markets, such as neighborhoods and census tracts, and therefore leads to distorted price predictions. A challenge in moving to estimates at, for example, the census tract level is the sparsity of spatiotemporally localized house sales observations.

Our work addresses the data sparsity challenge by leveraging observations from multiple census tracts discovered to have correlated valuation dynamics. We propose a Bayesian nonparametric approach which 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, which includes all house sales record from 1997 to 2013. We further incorporate a non-stationary trend into our Bayesian framework to capture the global effect, jointly with the local dynamics.

Our further work seeks to define the local neighborhood structure itself, rather than using predefined census tract regions. Instead of working with Euclidean distance, we propose an optimiza-

tion based graph algorithm to discover neighborhoods of houses that have similar attributes and are closely connected by roads. Our discovered regions are at a finer scale than census tracts, and even in this case our methods described above produce a house index at this hyperlocal neighborhood level, with better predictive performance as compared to the index at the census tract level.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

It has been such an adventure to receive my training from Department of Statistics, University of Washington in the past six years. I am so grateful for the challenges and opportunities brought to me during the long journey. I must thank my advisor Emily Fox for all the guidance, support and understanding. Emily's broad knowledge and enthusiasm for research have been an inspiration to me. I have learnt a lot from her about the taste to research problems, the attitude to work and to life. No matter how challenging the circumstance is, Emily is always there for a sincere discussion and being truly considerate for her students. I would also like to thank Andrew Bruce who helps tremendously on my research by contributing lots of brilliant ideas and inspirational discussions that shaped this thesis in many ways. Andrew has been a great mentor and a passionate collaborator whose creativity and support make this work possible. I would also like to thank Vladimir Minin and Michael Yost for their service as my thesis committees. Their contribution is invaluable to this thesis.

I have also had the honor of working with Paul Sampson on various interesting problems, from aircrafts to air pollution. Paul introduced me to statistical consulting and taught me to think as an applied statistician. I am deeply grateful to learn from Samuel Po-Shing Wong and Leung Fu Cheung from Chinese University of Hong Kong, who first introduced me to Statistics and Probabilities, and took a leap of faith in me though I had little background training.

I am also thankful for working with Jay Jones and Susan Kaluzny for the trust and mentoring during my internship at Weyerhauser.

Lastly, I wish to thank my parents for the support and patience, and my best friend and husband Qi Shan for encouraging me to keep exploring along the way.

DEDICATION

To Qi and Harry

Chapter 1

INTRODUCTION

The housing market is a large part of the global economy. In the United States, roughly fifty percent of household wealth is in residential real estate, according to a Federal Reserve Study [30]. Between 15% and 17% of the U.S. gross domestic product is on housing and housing related services according to GDP statistics published by the U.S. Bureau of Economic Analysis. Understanding how the value of housing 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 non-discretionary spending. By contrast, valuation of residential real estate is intrinsically difficult due to the individual nature of houses. Since the composition of the houses sold changes from one time period to the next, the change in the reported prices does not necessarily reflect the overall change in value. 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 in [3]. The main idea is to use a pair of sales for the same house to model the price trend over time. Assuming the house remains in the same condition, the first sales price serves as a surrogate for the house *hedonics* (house-level covariates) and the difference in the subsequent sales price captures the change in value over that intra-sales period. This approach largely circumvents the problem caused by the change in composition of houses sold. A large body of literature extends the original repeat sales model with numerous modifications and improvements [10, 11, 22, 55, 25]. 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. Case and Shiller [10] report that, over a study period of 16 years, single sales make up as much as 93%-97% of total transactions for metropolitan areas such as Atlanta, Dallas, Chicago and San Francisco. As such, studies based on repeat sales data rely on only a fraction of all transactions and may not be a good representation of the entire house market. Englund et al. [16] and Meese and Wallace [39] detected a sampling selection bias in which the repeat sales properties are older, smaller and more modest than single-sale properties. Furthermore, small samples lead to less precise parameter estimation. To overcome this, Case and Quigley [9] propose a hybrid model that combines repeat sales with hedonic information to make use of all sales. Recently, Nagaraja et al. [42] propose an autoregressive repeat sales model that utilizes all sales data without the need for hedonic information. Their approach leads to an index estimated quarterly at the zip code level.

Existing repeat sales models, even those using all of the transactions, perform the best when fit to relatively large areas, such as metropolitan areas or cities. Despite the large number of house sales observations in aggregate, when considering fine spatial resolutions, such as neighborhoods or census tracts, we have a large p (number of regions) small n (number of spatiotemporally-localized sales) problem. For example, in our dataset described in Section 3.1, most census tracts (114 out of 140) have fewer than 5 sales per month on average (see Table 1.1). The sparsity of transactions makes it challenging to obtain stable parameter estimates for small regions, and repeat sales models lack stability and predictive accuracy. This is a significant limitation: the value of real estate is intrinsically local and coarse-scale estimates may mask or distort key phenomena.

Table 1.1: 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
Percentage of tracts	0.11	0.41	0.81	0.97	0.99

An alternative, bottom-up, approach to constructing a housing price index is to compute an

estimate of each individual house value and then aggregate the house-level estimates. Zillow pioneered this approach with the Zillow Home Value Index (ZHVI®) by taking the median of all house-level estimates (Zestimate®) within a given region [64]. The 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. While the ZHVI confers certain advantages, there are limitations with the method. The approach is empirical in nature, and as such, does not directly try to model the underlying spatiotemporal dynamics of house values. House-level estimates are based on a prediction model proprietary to Zillow that uses a variety of data from different sources. The most important data are recent transactions. Depending on the homogeneity of the homes in an area and the uniqueness of a particular home, a significant history of transactions may be needed for a reliable estimate. This is a problem because the prediction model needs to adjust for the time of sale in order to account for the change in home value over time. In other words, the accuracy of the house-level prediction model, and consequently the ZHVI, is dependent on how well it captures the spatiotemporal dynamics of house values.

The main contribution of this dissertation 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 into 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 (although the ideas scale to finer spatiotemporal resolutions). This latent process is informed by all 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 similar dynamics.

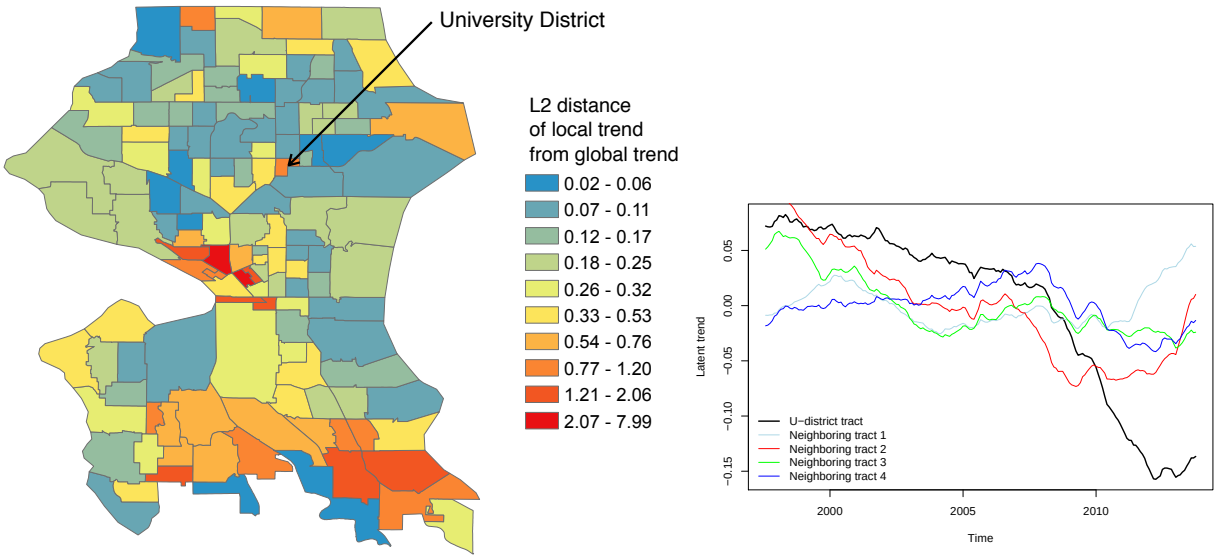
Unlike many spatiotemporal processes, modeling the correlation using Euclidean distance is not appropriate since spatially disjoint regions can be quite similar while neighboring census tracts can have significantly different value dynamics. For example, census tracts adjacent to water-

front, even if far apart, tend to share more in valuation dynamics than nearby census tracts that are not adjacent to waterfront. In our analysis of house sales in Seattle described in Section 3.8, we indeed find that certain census tracts vary dramatically from neighboring census tracts. Figure 1.1(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 in Figure 1.1(a) is the University District (U-District). Figure 1.1(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 building on the Dirichlet process—our formulation enables a flexible, data-driven method for discovering these clustered dynamics, including the number of clusters.

The approach taken offers several advantages over existing 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. We illustrate the impact of this multiple shrinkage in Figure 1.2, with a full analysis provided in Section 3.8. Likewise, the joint Bayesian framework considers all uncertainties together in the clustering, latent price inference and model parameter estimation.

1.1 Thesis Organization and Overview of Methods and Contributions

We provide an overview of the thesis organization, including contributions of each chapter, methods and results.



(a) (b)

Figure 1.1: (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 L_2 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 3.8.

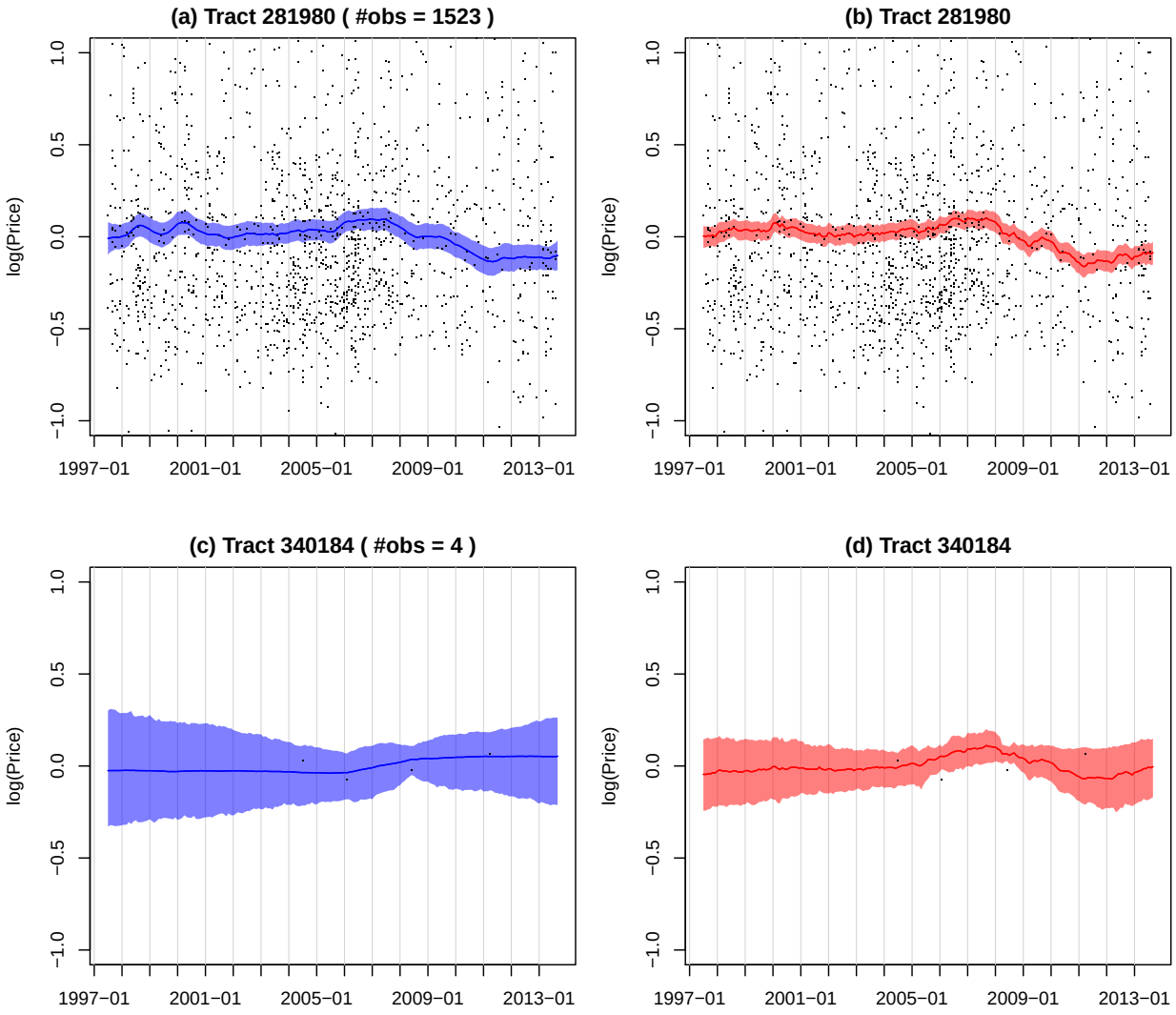


Figure 1.2: A demonstration of the effect of clustering: (a) and (b) show the posterior mean (*solid line*) and 95% intervals (*shaded gray*) for the latent price dynamics of a randomly sampled census tract with abundant observations (*dots*), whereas (c) and (d) examine a tract with sparse observations. Results are shown for models that either treat census tracts independently (*left*) or allow our Bayesian nonparametric clustering of tracts with similar dynamics (*right*) leading to narrower intervals, especially for tracts with few observations.

Chapter 2 Background We begin by reviewing statistical concepts and background that are used in this thesis. We first review the Markov and hidden Markov model as two important discrete-

time stochastic processes, usually with discrete state spaces. We then describe state space model which has the same graphical model as hidden Markov model, but with continuous state space. These statistical models serve as building blocks for our time series model for the housing index. We further discuss the Bayesian nonparametric method that is a flexible, data-driven approach for clustering. In particular, we review the Dirichlet process mixture model in details. For the Bayesian dynamical model we consider in this thesis, we rely on the Markov chain Monte Carlo techniques outlined in this chapter. We conclude the chapter by reviewing the spline models that will be used in modeling the non-stationary trend.

Chapter 3 Bayesian Dynamic Model for a Hyperlocal Housing Price Index We describe the Bayesian dynamical model for constructing a housing index at the census tract level, by overcoming data sparsity at such fine spatial granularity. Our model addresses the data sparsity challenge by leveraging observations from multiple regions discovered to have correlated valuation dynamics. We propose a Bayesian nonparametric approach which builds on the framework of latent factor models to enable a flexible, data-driven method for inferring the clustering of correlated regions. This model leverages information from the region-specific time series within a cluster, providing a form of multiple shrinkage of individual trend estimates for each region. We explore methods for scalability and parallelizability of computations. An analysis is provided on a large Seattle metropolitan housing dataset, which includes all house sales record from 1997 to 2013. Our main contribution includes providing a housing index at a resolution that was un-achievable before, enabling a flexible, data-driven clustering of correlated data streams and efficient computation.

Chapter 4 Housing Index Model Extensions We explore various modifications to the model and analyzing their effects, including joint estimation of the global market trend, test of the model sensitivity to the house covariates and possible extensions to our underlying dynamical model by considering long-memory processes.

Chapter 5 Defining the Neighborhood Regions This chapter seeks to study the problem of defining neighborhood structure for residential houses. To our best knowledge, this is the first attempt to learn neighborhoods rather than using pre-defined regions, such as census tract. We utilized the road network data to guide the neighborhood search over space. Our contribution includes a novel approach that encodes the spatial relationship through a graphical network of roads. We propose a cost minimization algorithm to optimize for within-neighborhood housing heterogeneity, with regularization on spatial boundaries of roads and neighborhood sizes. The discovered neighborhood structure in Seattle City shows consistent results with commonly recognized neighborhoods. We then apply our Bayesian dynamical model to produces a house index at this hyperlocal neighborhood level, which is a finer granularity than census tracts, yielding a better predictive performance as compared to the index at the pre-defined census tract level. This demonstrates that our Bayesian dynamical model is able to cope with extremely sparse scenarios.

Chapter 2

BACKGROUND

2.1 Markov and Hidden Markov Model

A Markov model or a Markov chain is a discrete-time sequence $\mathbf{x}_{1:T}$ such that the value at time t only depends on the value at the previous time $t-1$. In particular, it assumes the following

$$p(x_t|\mathbf{x}_{1:t-1}) = p(x_t|x_{t-1}). \tag{2.1}$$

The graphical model of a Markov chain is visualized in Figure 2.1. The joint distribution of a Markov chain is

$$p(\mathbf{x}_{1:T}) = p(x_1)\prod_{t=2}^{T} p(x_t|x_{t-1}). \tag{2.2}$$

If we further assume the transition function $p(x_t|x_{t-1})$ is independent of time, then we call the chain homogeneous or stationary. If x_t draws from a discrete distribution with finite state-space $\mathcal{S} = \{S_1, \cdots, S_n\}$, the transition probability $p(x_t|x_{t-1})$ can be written in a $n \times n$ matrix, called transition probability matrix.

A hidden Markov model (HMM) is a discrete-time Markov chain, consisting of hidden discrete-states x_t and observations generated from the hidden states according $p(y_t|x_t)$. The graphical

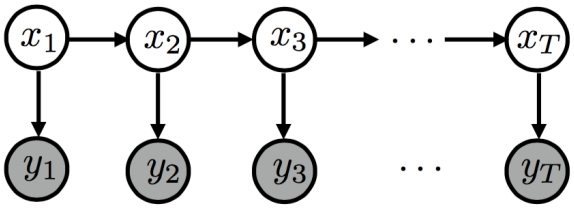


Figure 2.1: Graphical representation of a hidden Markov model (HMM) over T time steps. The dependencies in observations y_t are introduced by the latent states x_t.

model is illustrated in Figure 2.1. The sequence $\mathbf{x}$ is hidden and we observe data through the sequence $\mathbf{y}$. The key assumption for HMM is that given the hidden sequence $\mathbf{x}$ and the observation sequence $\mathbf{y}$, the value of y_t only depends on x_t. In particular,

$$p(y_t|\mathbf{x}_{1:T}, \mathbf{y}_{1:t-1,t+1:T}) = p(y_t|x_t). \tag{2.3}$$

The observation model $p(y_t|x_t)$ can be discrete or continuous distributions. The joint distribution of HMM is

$$p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1}) \prod_{t=1}^{T} p(y_t|x_t) \tag{2.4}$$

HMM establishes long-range dependence among observations, through the hidden states. HMM has wide applications in speech recognition [33, 34], activity recognition [59], genomics [52], protein sequence [15] and finance [4].

2.2 State Space Model

A state space model (SSM) is a discrete-time model like HMM, except that the hidden states are from continuous distributions. Its graphical model is the same one as HMM, shown in Figure 2.1. SSM consists of hidden continuous states $\mathbf{x}_t \in \mathbb{R}^n$, observations $\mathbf{y}_t \in \mathbb{R}^d$, an optional input signal $\mathbf{u}_t$ and model parameters. A commonly used and important SSM is a stationary linear-Gaussian state space model as follows

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t, \tag{2.5}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t + \mathbf{v}_t. \tag{2.6}$$

The system noise follows Gaussian $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$ and the observation noise follows $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$. The model parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ govern the dual processes. In this thesis, we assume uncontrolled process, i.e. $\mathbf{u} = \mathbf{0}$.

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{2.7}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{v}_t. \tag{2.8}$$

2.2.1 Kalman Filtering

The linear Gaussian assumption enables exact inference for the latent states $\mathbf{x}_{1:T}$ given observations and known parameters. The Kalman filter [35] is a recursive algorithm to infer the latent process for time t based on observations up to t. Kalman filtering performs the inference in an online fashion $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. In particular, it outputs the distribution of the filtered state $\mathbf{x}_t$ as follows.

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}) \tag{2.9}$$

Kalman filter recursively computes the mean and variance of the normal distribution in Eqn. (2.9), by taking a prediction step and an update step.

Working forward in time for each $t \in \{1, 2, \cdots, T\}$, the prediction step predicts the latent state at t given the past observations up to $t-1$

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1} \tag{2.10}$$

Based on the conditional independence $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and normality, Eqn. (2.10) becomes

$$\int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1} \tag{2.11}$$

$$= \int \mathcal{N}(\mathbf{x}_t; \mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}_t) \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1}\right) d\mathbf{x}_{t-1} \tag{2.12}$$

$$= \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right). \tag{2.13}$$

The marginal mean and variance can be obtained by the law of total expectation and the law of total variance as follows,

$$\boldsymbol{\mu}_{t|t-1} = E(\mathbf{x}_t) = E\left(E\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right)\right) \tag{2.14}$$

$$= E(\mathbf{A}\mathbf{x}_{t-1}) = \mathbf{A}\boldsymbol{\mu}_{t-1|t-1} \tag{2.15}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \text{Var}(\mathbf{x}_t) = E(\text{Var}(\mathbf{x}_t|\mathbf{x}_{t-1})) + \text{Var}(E(\mathbf{x}_t|\mathbf{x}_{t-1})) \tag{2.16}$$

$$= E(\mathbf{Q}_t) + \text{Var}(\mathbf{A}\mathbf{x}_{t-1}) \tag{2.17}$$

$$= \mathbf{Q}_t + \mathbf{A}_t \boldsymbol{\Sigma}_{t-1|t-1} \mathbf{A}_t^T \tag{2.18}$$

We then update the belief on the latent state $\mathbf{x}_t$ given the new observation $\mathbf{y}_t$. According to Bayes's rule, the posterior distribution can be derived as

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \mathbf{y}_t) &\propto p(\mathbf{x}_t|\mathbf{y}_{1:t-1})p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1}) && (2.19)\\ &= \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})\mathcal{N}(\mathbf{y}_t; C\mathbf{x}_t, R_t). && (2.20)\\ &\propto \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}). && (2.21) \end{aligned}$$

The derivation for the posterior mean and variance $(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$ is shown below. By conjugacy, we first obtain the posterior precision matrix

$$\boldsymbol{\Sigma}_{t|t}^{-1} = \boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{C}^T\mathbf{R}_t^{-1}\mathbf{C}. \tag{2.22}$$

The variance matrix is obtained by inversing the precision matrix. To simplify the result, we apply the matrix inversion lemma, also called Woodbury formula, to get

$$\begin{aligned} \boldsymbol{\Sigma}_{t|t} &= \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^T(\mathbf{R}_t + \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^T)^{-1}\mathbf{C}\boldsymbol{\Sigma}_{t|t-1} && (2.23)\\ &= (\mathbf{I} - \boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^T(\mathbf{R}_t + \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^T)^{-1}\mathbf{C})\boldsymbol{\Sigma}_{t|t-1}. && (2.24) \end{aligned}$$

To further simply the expression, we define

$$\begin{aligned} \mathbf{S}_t &= \mathbf{R}_t + \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^T && (2.25)\\ \mathbf{K}_t &= \boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^T\mathbf{S}_t^{-1} && (2.26) \end{aligned}$$

The latter is called Kalman gain matrix. Therefore, the variance in Eqn. (2.24) can be rewritten as

$$\boldsymbol{\Sigma}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\boldsymbol{\Sigma}_{t|t-1}. \tag{2.27}$$

We then obtain the posterior mean bu Bayes's rule and conjugacy,

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\Sigma}_{t|t}\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\mu}_{t|t-1} + \boldsymbol{\Sigma}_{t|t}\mathbf{C}\mathbf{R}_t^{-1}\mathbf{y}_t. \tag{2.28}$$

By utilizing the definition in Eqn. (2.26) , Eqn. (2.25) and the Woodbury's formula, the posterior mean can be simplified as

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{C}\boldsymbol{\mu}_{t|t-1}). \tag{2.29}$$

2.2.2 *Kalman Smoothing*

In contrast to the Kalman filter's inference in an online fashion $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, Kalman smoother considers an offline case $p(\mathbf{x}_t|\mathbf{y}_{1:T})$, conditioning on all observations and with the parameters known. In Kalman filter, we sweep the Markov chain from $t = 1$ to T and finally reaches $p(\mathbf{x}_T|\mathbf{y}_{1:T})$, whereas in Kalman smoother, we work backward from $t = T$ to $t = 1$. In particular, the output distribution

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) = \int p(\mathbf{x}_t|\mathbf{y}_{1:T}, \mathbf{x}_{t+1}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) d\mathbf{x}_{t+1}. \tag{2.30}$$

By conditional independence, we get

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}, \mathbf{x}_{t+1}) = p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{x}_{t+1}). \tag{2.31}$$

The conditional distribution of $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{x}_{t+1})$ can be obtained from the joint Gaussian distribution

$$\begin{pmatrix} \mathbf{x}_{t|t} \\ \mathbf{x}_{t+1|t} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{t|t} \\ \boldsymbol{\mu}_{t+1|t} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{t|t} & \boldsymbol{\Sigma}_{t|t}\mathbf{A}^T \\ \mathbf{A}\boldsymbol{\Sigma}_{t|t} & \boldsymbol{\Sigma}_{t+1|t} \end{pmatrix} \right). \tag{2.32}$$

We apply the law of total expectation and variance, similar to Eqn. (2.14) and Eqn. (2.16). The smoothed mean and variance are

$$\boldsymbol{\mu}_{t|T} = \boldsymbol{\mu}_{t|t} + \mathbf{J}_t(\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t}) \tag{2.33}$$

$$\boldsymbol{\Sigma}_{t|T} = \boldsymbol{\Sigma}_{t|t} + \mathbf{J}_t(\boldsymbol{\Sigma}_{t+1|T} - \boldsymbol{\Sigma}_{t+1|t})\mathbf{J}_t^T, \tag{2.34}$$

where $\mathbf{J}_t = \boldsymbol{\Sigma}_{t|t}\mathbf{A}^T\boldsymbol{\Sigma}_{t+1|t}^{-1}$. The Kalman smoother starts at $t = T$ and sweeps backward.

2.2.3 *Marginal Likelihood*

The marginal likelihood of observations with the latent states integrated out is

$$p(\mathbf{y}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \tag{2.35}$$

$$= \prod_{t=1}^{T} \mathcal{N}(\mathbf{y}_t; \mathbf{C}\boldsymbol{\mu}_{t|t-1}, \mathbf{S}_t). \tag{2.36}$$

The likelihood evaluation involves a Kalman filter algorithm to sweep along the chain from $t = 1$ to T.

2.2.4 Parameter Estimation via EM algorithm

We only observe the data sequence $\mathbf{y}$ in a state space model. To learn the parameters that drive both the hidden process and the observation process, we can use an Expectation-Maximization algorithm (EM) [56]. We provide a brief derivation and description of the EM algorithm here. Interested readers can refer to detailed derivations in Appendix A.

The unknown parameters in a state space model include $\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}$ for the doubly process. The joint log-likelihood of both $\mathbf{x}$ and $\mathbf{y}$ is

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \log p(\mathbf{x}, \mathbf{y}|\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}) && (2.37) \\ &= \log \left[\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{y}_t|\mathbf{x}_t) \right] && (2.38) \\ &= \sum_{t=1}^{T} \log p(\mathbf{x}_t|\mathbf{x}_{t-1}) + \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{x}_t). && (2.39) \end{aligned}$$

By substituting $\mathbf{x}_t|\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q})$ and $\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R})$, we get the explicit form of the joint log-likelihood as

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \sum_{t=0}^{T-1} \left(\frac{1}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)^T \mathbf{Q}^{-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t) \right) + && (2.40) \\ &\quad \sum_{t=0}^{T} \left(\frac{1}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t)^T \mathbf{R}^{-1} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t) \right) + \text{constant}. \end{aligned}$$

Since the hidden states are unknown, we marginalize out the latent states by taking the expectation of the joint log-likelihood as follows, known as the E-step in EM algorithm:

$$\begin{aligned} & E_{\mathbf{x}|\mathbf{y}} \left[l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) \right] && (2.41) \\ = \; & \frac{T}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{Q}^{-1} \sum_{t=0}^{T-1} \left[E(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T) - E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{x}_t\mathbf{x}_{t+1}^T) + \mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{A}^T \right] \right) + \\ & \frac{T}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{R}^{-1} \sum_{t=0}^{T} \left[\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t E(\mathbf{x}_t^T)\mathbf{C}^T - \mathbf{C}E(\mathbf{x}_t)\mathbf{y}_t^T + \mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{C}^T \right] \right) + \text{constant}. \end{aligned}$$

The maximum likelihood estimator for the parameters can be computed by solving the follow-

ing first derivative equations with respect to each parameter, known as the M-step in EM algorithm:

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial A} = \frac{1}{2}\mathbf{Q}^{-1}\left(\sum_{t=0}^{T-1}\left[2E(\mathbf{x}_{t+1}\mathbf{x}_t^T) - 2\mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\right]\right) = 0 \tag{2.42}$$

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial C} = \frac{1}{2}\mathbf{R}^{-1}\left(\sum_{t=0}^{T}\left[2\mathbf{y}_tE(\mathbf{x}_t^T) - 2\mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\right]\right) = 0 \tag{2.43}$$

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial Q^{-1}} \tag{2.44}$$

$$= \frac{T}{2}\mathbf{Q} - \frac{1}{2}\left(\sum_{t=0}^{T-1}\left[E(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T) - E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{x}_t\mathbf{x}_{t+1}^T) + \mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{A}^T\right]\right)^T = 0$$

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial R^{-1}} \tag{2.45}$$

$$= \frac{T+1}{2}\mathbf{R} - \frac{1}{2}\left(\sum_{t=0}^{T}\left[\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_tE(\mathbf{x}_t^T)\mathbf{C}^T - \mathbf{C}E(\mathbf{x}_t)\mathbf{y}_t^T + \mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{C}^T\right]\right)^T = 0. \tag{2.46}$$

Therefore, the estimators for maximizing the expected log-likelihood are

$$\hat{\mathbf{A}} = \left(\sum_{t=0}^{T-1}E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\right)\left(\sum_{t=0}^{T-1}E(\mathbf{x}_t\mathbf{x}_t^T)\right)^{-1} \tag{2.47}$$

$$\hat{\mathbf{C}} = \left(\sum_{t=0}^{T}\mathbf{y}_tE(\mathbf{x}_t^T)\right)\left(\sum_{t=0}^{T}E(\mathbf{x}_t\mathbf{x}_t^T)\right)^{-1} \tag{2.48}$$

$$\hat{\mathbf{Q}} = \frac{1}{T}\left(\sum_{t=0}^{T-1}\left[E(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T) - E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{x}_t\mathbf{x}_{t+1}^T) + \mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{A}^T\right]\right) \tag{2.49}$$

$$\hat{\mathbf{R}} = \frac{1}{T+1}\left(\sum_{t=0}^{T}\left[\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_tE(\mathbf{x}_t^T)\mathbf{C}^T - \mathbf{C}E(\mathbf{x}_t)\mathbf{y}_t^T + \mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{C}^T\right]\right). \tag{2.50}$$

Algorithm 1 describes detailed steps for the EM algorithm.

Initialize $\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}$;

while *not converged:* **do**

E-step:

Apply Kalman smoother to obtain the following statistics from the conditional distribution of hidden states given observations and parameters for $t = 1, \cdots, T$.

$$E\left(\mathbf{x}_t|\mathbf{y}_{1:T}\right), E\left(\mathbf{x}_t\mathbf{x}_t|\mathbf{y}_{1:T}\right), E\left(\mathbf{x}_{t+1}\mathbf{x}_t|\mathbf{y}_{1:T}\right).$$

M-step:

Update parameters $\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}$ to maximize the expected log-likelihood.

end

Algorithm 1: EM algorithm.

2.3 Bayesian Nonparametric Methods

Bayesian nonparametric methods provide a flexible framework that allows the model complexity to adapt to the size of the data. Bayesian nonparametric methods have been used in various applications including text documents, time series and images [20, 58]. Detailed reviews can be found in [40, 58, 61]. In particular, for our task of clustering, Bayesian nonparametric approach enables inference of the cluster structure, in terms of number of clusters and associated membership. We briefly describe one commonly used class of Bayesian nonparametric methods: the Dirichlet process (DP).

2.3.1 The Dirichlet Process

A DP [5, 18] is a distribution over countably infinite discrete probability measures. A draw $G \sim DP(\alpha, G_0)$, with concentration parameter α and base measure G_0, can be constructed as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, \quad \theta_k^* \sim G_0, \tag{2.51}$$

where the mixture weights π_k are sampled via a stick breaking construction [53]:

$$\pi_k = v_k \prod_{j=1}^{k-1}(1 - v_j), \quad v_k \sim \text{Beta}(1, \alpha). \tag{2.52}$$

We denote the stick breaking process as $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots)$. The DP prior produces clusters of $\theta_i \sim G$, $i = 1, \cdots, n$, 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 $z_i \sim \pi$ such that $z_i = k$ implies that θ_i takes the unique value θ_k^*. That is, $\theta_i = \theta_{z_i}^*$.

Integrating out the stick breaking measure $\boldsymbol{\pi}$, the predictive distribution of z_i given the memberships of other tracts $\mathbf{z}_{-i}$ is

$$P(z_i = k|\mathbf{z}_{-i}, \alpha) \propto \begin{cases} \frac{n_{-i,k}}{n-1+\alpha} & \text{for } k = 1, \cdots, K \\ \frac{\alpha}{n-1+\alpha} & \text{for } k = K+1, \end{cases} \tag{2.53}$$

where K indicates the number of unique values of z_i in $\mathbf{z}_{-i}$. That is, tract i may join one of the existing clusters with probability proportional to the size of the cluster, $n_{-i,k}$, or start a new cluster with probability proportional to α. The resulting sequence of partitions is referred to as the *Chinese Restaurant Process* (CRP) [48].

2.4 Markov Chain Monte Carlo

Markov Chain Monte Carlo is a key computing technique used in Bayesian inference. We first describe the background of Monte Carlo Integration in Section 2.4.1 and then move to discuss the Markov Chain Monte Carlo methods for high-dimensional data in Section 2.4.2.

2.4.1 Monte Carlo Integration

Monte Carlo methods can be viewed as a numerical integration problem. The main interest of the problem is taking expectations with respect to a distribution. Typically for Bayesian inference, we are interested in the mean of a given function $f(\boldsymbol{\theta})$ with respect to the posterior distribution for parameter $\boldsymbol{\theta}$ given observations $\mathbf{y}$,

$$E[f(\boldsymbol{\theta})|\mathbf{y}] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{2.54}$$

The integral in Eqn. (2.54) can be difficult to solve, especially if the parameter is high-dimensional. In such case, we resort to Monte Carlo techniques to solve such high dimensional integration by simulation. For a general case with any parameter distribution $p(\boldsymbol{\theta})$, If we simulate $\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(n)} \sim p(\boldsymbol{\theta})$ and $E[f(\boldsymbol{\theta})] < \infty$, then by the strong law of large number (SLLN) we get $\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}^{(i)})$ converges to $E[f(\boldsymbol{\theta})]$ almost surely as n goes to infinity. Therefore, we can approximate the target expectation with

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}^{(i)}) \approx E[f(\boldsymbol{\theta})]. \tag{2.55}$$

There are many techniques for simulating $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ if $\boldsymbol{\theta}$ is low dimension, including generalized inverse of the cumulative distribution function, accept-reject algorithm, importance sampling etc. More details can be found in [51].

2.4.2 Markov Chain Monte Carlo Methods

The classical Monte Carlo in Section 2.4.1 is hard to implement in high dimensional parameter space. If $\boldsymbol{\theta}$ is high dimensional or if we don't know how to simulate samples directly from $p(\boldsymbol{\theta})$, we turn to Markov Chain Monte Carlo (MCMC) methods. The goal of MCMC is the same as for classical Monte Carlo shown in Eqn. (2.54). The strategy of MCMC method is to construct an ergodic Markov chain $\{\boldsymbol{\theta}\}$ with stationary distribution $p(\boldsymbol{\theta})$, such that the resulting samples from the chain can be used to form the estimator in Eqn. (2.55), by the ergodic theorem [51].

We describe how to construct such a Markov chain. Suppose that the parameter space, $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, is a Cartesian product of smaller subspaces, $\boldsymbol{\Omega} = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_m$. Assume we can sample from full conditional distributions $\theta_i | \boldsymbol{\theta}_{-i}$. If we iteratively sample from these full conditional distributions, we will form a Markov chain with the target distribution $f(\boldsymbol{\theta})$, which is the sequential scan Gibbs sampling algorithm listed in Algorithm 2. See [51] for proof and further details.

Start with initial value $\boldsymbol{\theta}^{(0)}$;

for $t \leftarrow 0$ **to** N **do**

Sample $\theta_1^{(t+1)} \sim f_1\left(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \cdots, \theta_m^{(t)}\right)$

Sample $\theta_2^{(t+1)} \sim f_2\left(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \cdots, \theta_m^{(t)}\right)$

$\vdots$

Sample $\theta_m^{(t+1)} \sim f_p\left(\theta_p|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \cdots, \theta_m^{(t+1)}\right)$

end

Algorithm 2: Sequential scan Gibbs sampling.

2.5 Spline Models

A spline model is a linear model with a set of basis functions as predictors. It allows for great flexibility for modeling the function $f(x)$ to estimate the response variable y. One popular form of splines among piecewise polynomials is the cubic spline, which can be represented as

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{l=1}^{L} b_l (x - \xi_l)_+^3, \tag{2.56}$$

where $\{\xi_l\}$ are knots and $(x - \xi_l)_+$ denote truncated lines. The basis functions in Eqn. (2.56) include an intercept, linear, quadratic and cubic terms and truncated cubic basis terms. The parameterization in Eqn. (2.56) also provides continuous first and second derivatives at all knots to produce a smooth curve.

Spline models can produce erratic behavior beyond the boundary knots. Motivated by this, a natural spline enforces linearity beyond the boundary, that is

$$f(x) = a_1 + a_2 x \quad \text{if } x \leq \xi_1, \tag{2.57}$$

$$f(x) = a_3 + a_4 x \quad \text{if } x \geq \xi_L. \tag{2.58}$$

Natural cubic splines are proven to be optimal for minimizing the penalized least squares for data

points $(x_i, y_i), i = 1, \cdots, n$ as follows

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx, \tag{2.59}$$

where the second derivative $f''(x)$ measures the roughness of the curve and λ is the weight of the penalty on roughness. The estimator $f(\cdot)$ that minimizes the penalized least squares error in Eqn. (2.59) is the natural cubic spline with knots at the unique data points. The proof can be found in [27].

For implementation, B-spline basis functions are commonly used to represent a natural cubic spline model for numerical stability, since B-spline basis functions are non-zero over a limited range. More details are reviewed in [60].

Chapter 3

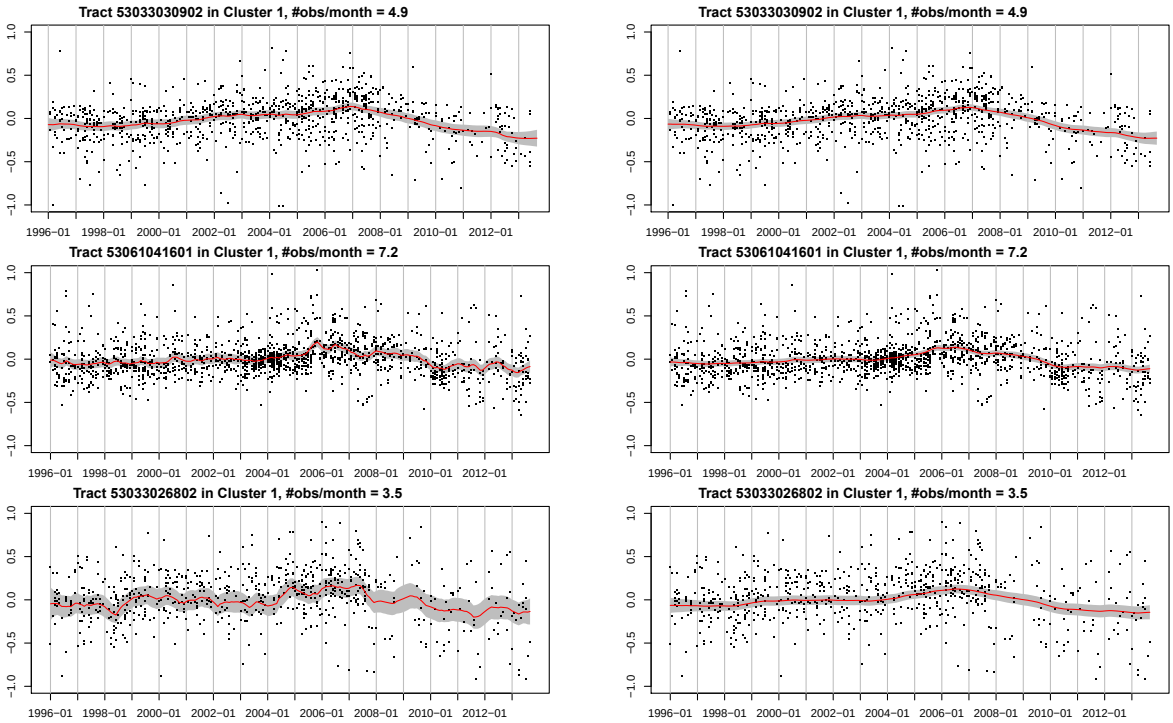
BAYESIAN DYNAMIC MODEL FOR A HYPERLOCAL HOUSING PRICE INDEX

In this Chapter, we introduce a Bayesian dynamic model for constructing housing index at a fine spatial granularity, such as census tract level. We addresses the data sparsity challenge by leveraging observations from multiple census tracts discovered to have correlated price dynamics. We propose a Bayesian nonparametric approach which builds on a latent factor model to enable a flexible, data-driven method for inferring the clustering of correlated census tracts.

This Chapter is organized as follows. Section 3.1 introduces the house transaction data used in our analysis. Section 3.2 describes the dynamical model for each census tract individually, and then the correlation structure introduced to couple the tract dynamics within a large geographic region. Section 3.3 explains the prior distributions for each component in the Bayesian model. Section 3.4 provides a model overview and Section 3.5 proves an outline of the posterior sampling steps. Section 3.6 discusses some of the computational challenges and a strategy to implement the algorithm in parallel. A simulation study is provided in Section 3.7 and a detailed analysis on our Seattle housing dataset is in Section 3.8.

3.1 House Transaction Data

Our house sales data consist of 124,480 transactions in 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`. Naively, the number of bedrooms



(a) (b)

Figure 3.1: An illustration of relating time series: (a) univariate Kalman smoother applied independently to the time series of each census tract, (b) multivariate Kalman smooth applied jointly to the tracts in the same cluster inferred using hierarchical clustering.

might be considered a strong predictor of the sales price. Indeed, there is a positive correlation between the number of bedrooms and sales price; however, this can be attributed to the association of this hedonic with the total finished square feet of a home. With the inclusion of square feet in the model, the number of bedrooms is not a significant variable. This is common facet of many house price regression models.

To assess the importance of considering related regions jointly, we performed the following data analysis. First, we examined the state space model of Eqs. (3.3)-(3.4) independently across regions i (whereas in Section 3.2 the focus is on joint modeling of regions). In that model, the latent state

sequence represents the underlying price evolution of a given region—our desired index—and the observations are the individual house sales. To infer the latent state sequence jointly with the model parameters, we use a Kalman smoother embedded in an expectation maximization (EM) algorithm. For this analysis and the remainder of the paper, our spatial granularity of interest is a census tract. We compare the performance of this independent model to one that jointly analyzes related tracts, where relatedness is determined by a hierarchical clustering approach. The hierarchical clustering is based on L_2 distance between the independently Kalman smoothed estimates of the latent state sequence. After performing the hierarchical clustering and cutting the tree at a certain level, we consider a multivariate latent state model as in Eq. (3.3) where all tracts i falling in the same cluster have correlated innovations, $\epsilon_{t,i}$. That is, $\boldsymbol{\epsilon}_t^{(k)} \sim N(0, \boldsymbol{\Sigma}^{(k)})$ with a full covariance matrix $\boldsymbol{\Sigma}^{(k)}$, where $\boldsymbol{\epsilon}_t^{(k)}$ is the vector of $\epsilon_{t,i}$ for tracts i in cluster k. The observation model remains as in Eq. (3.4). We then applied a Kalman-smoother-within-EM algorithm to the resulting multivariate state space model.

Unsurprisingly, without sharing observations from similar tracts, the baseline independent approach does not perform well when the observations are sparse, as shown in Figure 3.1(a). In contrast, by pooling observations from other tracts, the hierarchical clustering-based latent price dynamics are smoother and with narrower intervals, as shown in Figure 3.1(b). 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 one 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 $\boldsymbol{\Sigma}^{(k)}$. In Figure 3.1(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 3.2.

3.2 A Model for Relating Multiple Time Series

Recall our goal of inferring a housing valuation index for a small geographic region, e.g. census tract. We are faced with a large number of geographic regions and a relatively small number of observations for each region. Our modeling strategy is to discover price dynamics shared between these 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 the processes that share similar price dynamics. Throughout, we will assume that our geographic unit of interest is a census tract.

3.2.1 Per-Series Dynamics

We model the dynamics of the house sales prices within a census tract via a state space model. Each census tract i may have multiple house sale observations $\tilde{y}_{t,i,l}$ at time 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):

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

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

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

For the sake of simplicity, since our focus is on small geographic regions, we assume that the global trend g_t is known or pre-calculated based on all transactions in the market. Computing a global trend is relatively straightforward since we have sufficient data in aggregate. Instead, we focus on modeling the deviance of the latent dynamics of census tracts from the global market trend. This deviance can be defined as $x_{t,i} \equiv \tilde{x}_{t,i} - g_t$. To further simplify the model, we assume

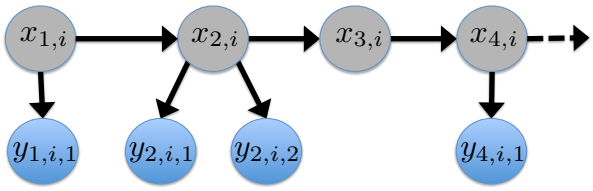


Figure 3.2: An illustration of the state space model of Eqs. (3.3)-(3.4) for census tract i's data stream. The observed sales prices (after removing the global market trend) are denoted by $y_{t,i,l}$ and the (detrended) intrinsic price of census tract i by $x_{t,i}$.

the house feature function $f_i(\cdot)$ is composed of linear basis functions. The simplified model is

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

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

where $y_{t,i,l} \equiv \tilde{y}_{t,i,l} - g_t$ represents the deviance of each house sales price at time t from the global market price at that time. We refer to the latent $x_{t,i}$ order 1 autoregressive process (AR(1)) in Eq. (3.3) as the *intrinsic price dynamics* for each census tract. Since we are modeling deviances from the global trend, the choice of a stationary process is reasonable. We call the series of observations from one census tract a *data stream*. Eqs. (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 3.2.

3.2.2 Relating Multiple Data Streams

There are clearly temporal trends to house values, and these trends may vary significantly and sometimes abruptly across geographic locations. We want to share information between related tracts which exhibit similar temporal trends, and aim to discover these groups of tracts from the observed data streams. 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 infer-

ence perspective, clustering census tracts increases power and precision in parameter estimation by pooling the observations from grouped data streams.

We now seek to define the mechanism by which data streams relate, and then use this to cluster the series. If house prices in one neighborhood increase, prices in related neighborhoods are also likely to increase. This type of sharing of price dynamics can be modeled by correlating the *innovation* terms $\epsilon_{t,i}$ in the intrinsic price dynamics of Eq. (3.3). We then cluster all series with correlated innovations and assume independence between the dynamics of those falling into separate clusters. More specifically, let $z_i = k$ denote that census tract i is in cluster k and $\boldsymbol{\epsilon}_t^{(k)}$ be the vector of innovations $\epsilon_{t,i}$ for census tracts in cluster k. Instead of treating the $\epsilon_{t,i}$ independently across i, the intrinsic price dynamics $x_{t,i}$ within cluster k can be correlated by considering $\boldsymbol{\epsilon}_t^{(k)} \sim \mathcal{N}(0, \Sigma_k)$ for Σ_k non-diagonal. We assume $\boldsymbol{\epsilon}_t^{(k)}$ is independent of $\boldsymbol{\epsilon}_t^{(j)}$ for all $j \neq k$. Stacking up all $\boldsymbol{\epsilon}_t^{(k)}$, $k = 1, \cdots, K$, into a large $\boldsymbol{\epsilon}_t$ vector of length p (the number of census tracts), our model is equivalent to $\boldsymbol{\epsilon}_t \sim N(0, \Sigma)$ for Σ block diagonal with blocks Σ_k.

Jointly clustering census tracts *and* correlating the dynamics within a given cluster is a challenging task; it is equivalent to inferring the block structure of Σ, which entails discovering both an ordering on the census tracts and the dimensions of the blocks Σ_k. Since we do not assume that the number of clusters (number of blocks in Σ) is known, this adds an additional challenge.

To generatively define block diagonal covariance matrices with unknown block sizes, we leverage **latent factor models**. We start by assuming that there are K clusters with known membership, and then revisit the idea of inferring the memberships and number of clusters in Section 3.3. In particular, consider:

$$\epsilon_{t,i} = \lambda_{iz_i} \eta^*_{t,z_i} + \tilde{\epsilon}_{t,i}, \quad \tilde{\epsilon}_{t,i} \sim N(0, \sigma_0^2), \quad \eta^*_{t,k} \sim N(0,1). \tag{3.5}$$

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 $\tilde{\epsilon}_{t,i}$ is idiosyncratic noise drawn independently over time and tracts. We model $\lambda_{ik} \sim N(\mu_\lambda, \sigma_\lambda^2)$. We can then write $\boldsymbol{\epsilon}_t = (\Lambda \cdot Z)\boldsymbol{\eta}_t^* + \tilde{\boldsymbol{\epsilon}}_t$, where Λ is a $p \times K$ Gaussian matrix, Z is an indicator matrix with $Z_{ik} = 1[z_i = k]$, $\boldsymbol{\eta}_t^* \sim \mathcal{N}_K(0, I)$, and $\tilde{\boldsymbol{\epsilon}}_t \sim \mathcal{N}_p(0, \sigma_0^2 I)$. Here, $A \cdot B$ represents the element-wise product. Conditioned on the factor

loading matrices Λ and Z, the covariance for $\boldsymbol{\epsilon}_t$ is $\Sigma = (\Lambda \cdot Z)(\Lambda \cdot Z)^T + \sigma_0^2 I$. Equivalently,

$$\text{cov}(\epsilon_{t,i}, \epsilon_{t,i'} | \Lambda, Z) = \begin{cases} \lambda_{ik}\lambda_{i'k} + \sigma_0^2 \delta(i, i') & z_i = z_{i'} = k, \forall k \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

From Eq. (3.6), the conditional covariance for $\boldsymbol{\epsilon}_t$ is a block-diagonal matrix defined by the clusterings specified by z_i. That is, data streams within the same cluster will have correlated dynamics, and those in different clusters will evolve independently.

To infer the clustering of region-specific data streams, we propose a Bayesian nonparametric approach using a Dirichlet Process (DP) prior that leads to an adaptive, data-driven clustering, allowing for an unknown number of blocks (clusters) in the covariance. This model is related to that of [46], but specified for the time series domain. The details of our prior specification are in Section 3.3.

3.3 Prior Specification

In this section, we describe the prior specifications for our various model parameters.

3.3.1 Cluster Membership

We provide background on the DP in Section 2.3.1. Here we describe how we utilize this prior in our dynamical model to cluster tracts with correlated intrinsic price dynamics in the presence of an unknown number of clusters.

Clustering of Regions by Latent Dynamic Factors

In our housing application, we place a DP prior on the parameter by which we wish to cluster the census tract intrinsic dynamics. As detailed in Section 3.2, we relate the data streams within a cluster via correlated dynamics induced by a latent factor model with a cluster-specific latent factor process $\eta^*_{1:T,k}$. As such, to specify a Bayesian nonparametric clustering model we take $\eta_{1:T,i} \sim G$ with $G \sim \text{DP}(\alpha, G_0)$, where $\eta_{1:T,i}$ is the latent factor process for census tract i. In our indicator variable representation, we define mixture weights $\pi \sim \text{GEM}(\alpha)$, cluster-specific

parameters $\eta^*_{1:T,k} \sim G_0$, and cluster indicators $z_i \sim \pi$ such that $\eta_{1:T,i} = \eta^*_{1:T,z_i}$. That is, $\eta_{1:T,i}$ serves the role of θ_i and $\eta^*_{1:T,k}$ equates with θ^*_k in the generic Dirichlet process mixture model of Section 2.3.1. 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 $t = 1, \ldots, T$, $k = 1, 2, \ldots$.

3.3.2 Latent Autoregressive Process Parameters

The latent autoregressive (AR) process in Eq. (3.3) has an autoregressive parameter a_i, a factor loading λ_{ik}, and the variance of the idiosyncratic noise σ_0^2. We place conjugate priors on these parameters, respectively:

$$a_i \sim \mathcal{N}(\mu_a, \sigma_a^2), \quad i = 1, \ldots, p \tag{3.7}$$

$$\lambda_{ik} \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2), \quad i = 1, \cdots, p, \ k = 1, 2, \ldots \tag{3.8}$$

$$\sigma_0^2 \sim IG(\alpha_{\epsilon 0}, \beta_{\epsilon 0}). \tag{3.9}$$

The hyperparameters $\mu_a, \sigma_a^2, \mu_\lambda, \sigma_\lambda^2$ are given priors as well. Details of these hyperpriors and settings of the hyperparameters $\alpha_{\epsilon 0}, \beta_{\epsilon 0}$ are provided in Supplement E.1.

3.3.3 Emission Parameters

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

$$\beta_{i,h} \sim \mathcal{N}(\mu_h, \sigma_h^2) \quad i = 1, \cdots, p, \ h = 1, \cdots, H, \tag{3.10}$$

$$R_i \sim IG(\alpha_{R0}, \beta_{R0}) \quad i = 1, \cdots, p. \tag{3.11}$$

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.

3.4 Model Overview

Our model assumes that the observed house sales prices center about the intrinsic price of the associated census tract and transaction month, after accounting for hedonic effects. The intrinsic price

for each census tract follows an AR(1) marginally. The DP provides a flexible prior for nonparametric clustering of the intrinsic price dynamics associated with each census tract based on our latent factor model, which induces correlation of price dynamics within a cluster. Figure 3.3 shows the graphical model representation. The overall model specification for the Bayesian nonparametric house sale dynamic model is summarized as:

1. Draw Dirichlet process realization $G \sim DP(\alpha, G_0)$:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, \qquad \text{where } \theta_k^* = \boldsymbol{\eta}_{1:T,k}^* \tag{3.12}$$

2. For the data stream associated with each census tract i from 1 to p:

   (a) Draw cluster membership $z_i|\boldsymbol{\pi} \sim \boldsymbol{\pi}$

   (b) Draw factor loadings $\lambda_{ik} \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2)$

   (c) For each timestep t from 1 to T:

      i. Draw the state sequence $x_{t,i}|x_{t-1,i}, z_i \sim \mathcal{N}(a_i x_{t-1,i} + \lambda_{iz_i}\eta_{t,z_i}^*,\ \sigma_0^2)$

      ii. Draw an observation $y_{t,i,l}|x_{t,i} \sim \mathcal{N}\left(x_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,h},\ R_i\right)$

3.5 MCMC Posterior Computations

Our posterior computations are based on a Gibbs sampler, with steps detailed below. Scaling this sampling strategy to our large housing dataset is discussed in Section 3.6.

Letting $\boldsymbol{\psi} = \{\mathbf{a} = \{a_i\}, \boldsymbol{\lambda} = \{\lambda_{ik}\}, \mathbf{R} = \{R_i\}, \boldsymbol{\beta} = \{\beta_{i,h}\}, \sigma_0^2\}$ and $\boldsymbol{\psi}^{(k)}$ the associated subset of parameters corresponding to the k-th cluster based on assignments $\mathbf{z} = \{z_i\}$, the Gibbs sampler is outlined as follows:

1. Sample $z_i = k|\mathbf{z}_{-i}, \alpha, \mathbf{y}, \boldsymbol{\psi}$. Note we marginalize the stick-breaking random measure $\boldsymbol{\pi}$, the latent housing valuation processes $\mathbf{x}^{(k)}$, and the cluster latent factor processes $\boldsymbol{\eta}^{*(k)}$.

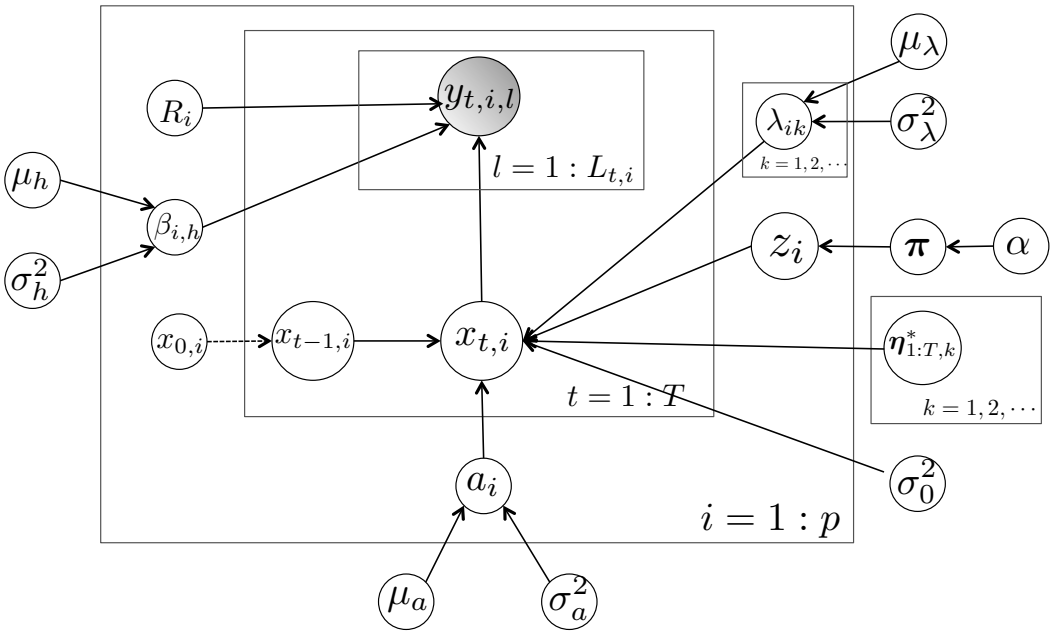


Figure 3.3: Graphical model representation of our Bayesian nonparametric house sales dynamic model summarized in Section 3.4. Boxes indicate replication of random variables and shaded nodes the observations. Note that $x_{1:T}$ forms a length T Markov chain realization; our box here is an abuse of notation used for compactness.

2. Impute $\mathbf{x}$ and $\boldsymbol{\eta}^*$ as auxiliary variables. Specifically, block sample $\mathbf{x}$, $\boldsymbol{\eta}^*$ as $\mathbf{x}^{(k)}|\mathbf{z}, \mathbf{y}^{(k)}, \boldsymbol{\psi}^{(k)}$ and $\boldsymbol{\eta}^*|\mathbf{z}, \mathbf{x}, \boldsymbol{\psi}$.

3. Sample $\boldsymbol{\psi}^{(k)}|\mathbf{z}, \mathbf{y}^{(k)}, \mathbf{x}^{(k)}, \boldsymbol{\eta}^{*(k)}$

4. Discard $\mathbf{x}$ and $\boldsymbol{\eta}^*$ to sample hyperparameters conditional on $\boldsymbol{\psi}, \mathbf{z}$.

3.5.1 Sampling the cluster membership

We sample the cluster indicators z_i conditional on model parameters and house sales transactions. We analytically marginalize out the infinite set of mixture weights $\boldsymbol{\pi}$, latent factor process $\boldsymbol{\eta}^* = \{\eta^*_{1:T,k}\}$ and the intrinsic dynamics $\mathbf{x} = \{x_{1:T,i}\}$. Specifically, the full conditional of indicator z_i for census tract i is:

$$\begin{aligned} &P\left(z_i = k \left| \mathbf{z}_{-i}, \mathbf{y}_{1:T}, \{a_j\}^{(k)}, a_i, \{\lambda_{jk}\}, \sigma_0^2, \{R_j\}^{(k)}, R_i, \{\beta_{j,h}\}^{(k)}, \{\beta_{i,h}\}, \alpha\right.\right) \\ &\quad \propto P\left(z_i = k \left| \mathbf{z}_{-i}, \alpha\right.\right) P\left(\mathbf{y}_{1:T,i} \left| z_i = k, \mathbf{z}_{-i}, \mathbf{y}^{(k)}_{1:T,-i}, \{a_j\}^{(k)}, \Sigma^{(k)}, \{R_j\}^{(k)}, \{\beta_{j,h}\}^{(k)}\right.\right). \end{aligned} \quad (3.13)$$

The first factor is the prior belief of cluster membership for tract i conditional on memberships of all other tracts, which results from the CRP prior of Eq. (2.53) (and the use of exchangeability). 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, $\mathbf{y}^{(k)}_{1:T,-i}$, and the covariance between intrinsic dynamics in the cluster, $\Sigma^{(k)}$, specified via Eq. (3.6). The other model parameters for cluster k include: the AR coefficients $\{a_j\}$, observational variances $\{R_j\}$, and hedonic effects $\{\beta_{j,h}\}$ for all tracts j in cluster k (including i when conditioning upon $z_i = k$). We denote these restricted sets via $\{\cdot\}^{(k)}$.

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 $\mathbf{x}^{(k)}_{1:T}$. 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.

For the special case of census tract i creating a new cluster, i.e. $z_i = K+1$, the prior belief follows the CRP prior of Eq. (2.53). The likelihood becomes simply $P\left(\mathbf{y}_{1:T,i} \left| \mathbf{z}, a_i, \Sigma^{(K+1)}, R_i, \beta_{i,h}\right.\right)$, where $\Sigma^{(K+1)} = \lambda^2_{i,K+1} + \sigma_0^2$, as specified in Eq. (3.6) and having sampled $\lambda_{i,K+1} \sim \mathcal{N}(\mu_\lambda, \sigma^2_\lambda)$ for

all tracts, but marginalizing $\eta^*_{1:T,K+1}$. This represents a variant of Neal's Algorithm 8 for sampling from DP models [43].

3.5.2 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 $\mathbf{x}^{(k)}_{1:T}$ 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.

We then sample $\boldsymbol{\eta}^*$ given $\mathbf{x}$. Recall that the intrinsic price dynamics for multiple tracts in the same cluster k are correlated through the common latent factor process $\eta^*_{1:T,k}$ for the AR(1) innovations $\boldsymbol{\epsilon}^{(k)}$, as in Eq. (3.5). By conjugacy, we sample the cluster-specific latent factor $\eta^*_{t,k}$ for time period $t = 1, \cdots, T$ and K existing clusters as follows:

$$\boldsymbol{\eta}^*_t \,|\boldsymbol{\lambda}, \mathbf{z}, \mathbf{x}, \mathbf{a}, \sigma_0^2 \sim \mathcal{N}_K \left\{ \begin{array}{c} V \frac{1}{\sigma_0^2}(\Lambda \cdot Z)^T(\mathbf{x_t} - A\mathbf{x_{t-1}}), \\ V = \left[I_K + \frac{1}{\sigma_0^2}(\Lambda \cdot Z)^T(\Lambda \cdot Z)\right]^{-1} \end{array} \right\}. \tag{3.14}$$

The derivation is provided in Supplement C.2.

3.5.3 Sampling factor loadings λ

We sample the loadings λ_{ik} of the loadings matrix $\Lambda_{p\times K}$ as

$$\lambda_{ik}|\mathbf{x}, \mathbf{a}, \boldsymbol{\eta}^*, \mathbf{z}, \sigma_0^2 \sim \mathcal{N}(\mu^*_{ik}, v^*_{ik}), \tag{3.15}$$

where

$$(\mu^*_{ik}, v^*_{ik}) = \left\{ \begin{array}{ll} \mu_\lambda, \sigma^2_\lambda & \text{if } Z_{ik} = 0 \\ v\left(\frac{\mu_\lambda}{\sigma^2_\lambda} + \frac{1}{\sigma_0^2}\sum_{t=1}^T \epsilon_{t,i}\eta^*_{t,k}\right), v = \left(\frac{1}{\sigma^2_\lambda} + \frac{1}{\sigma_0^2}\sum_{t=1}^T (\eta^*_{t,k})^2\right)^{-1} & \text{if } Z_{ik} = 1 \end{array} \right. .$$

Here, we recall the definition of the membership matrix Z from Section 3.2.2. Note that $\epsilon_{t,i} = x_{t,i} - a_i x_{t-1,i}$ and $\sum_{t=1}^T \epsilon_{t,i}\eta^*_{t,k}$ can be written as the inner product $\boldsymbol{\epsilon}_i^T \boldsymbol{\eta}^{*(k)}$. The derivation is given in Supplement C.3.

3.5.4 Sampling AR parameters a_i

Using conjugacy results of the normal distribution, and conditioning upon a cluster assignment $z_i = k$, we sample the tract-specific AR coefficient a_i for $i = 1, \cdots, p$ as

$$a_i \left| z_i = k, \mathbf{x}_i, \boldsymbol{\eta}^{*(k)}, \lambda_{ik}, \sigma_0^2, \mu_a, \sigma_a^2 \right. \sim \mathcal{N}\left(V\left[\frac{\mu_a}{\sigma_a^2} + \sum_{t=1}^{T} \left(\frac{x_{t-1,i}^2}{\sigma_0^2} \cdot \frac{x_{t,i} - \lambda_{ik}\eta_{t,k}^*}{x_{t-1,i}} \right) \right], V = \left(\frac{1}{\sigma_a^2} + \sum_{t=1}^{T} \frac{x_{t-1,i}^2}{\sigma_0^2} \right)^{-1} \right).$$

The derivation is provided in Supplement C.4.

3.5.5 Sampling emission parameters R, β, σ_0^2

By conjugacy, we can sample the observation variance R_i for $i = 1, \cdots, p$ as

$$R_i \left| \mathbf{x}_{1:T,i}, \mathbf{y}_{1:T,i}, \alpha_{R0}, \beta_{R0} \right. \sim \text{IG}\left(\alpha_{R0} + \frac{1}{2} m_i,\ \beta_{R0} + \frac{1}{2} \sum_{t=1}^{T} \sum_{l=1}^{Lt} (y_{t,i,l} - x_{t,i})^2 \right),$$

where m_i is the number of transactions in census tract i. The values of the hyperparameters α_{R0}, β_{R0} are provided in Supplement E.2.

We sample the covariate effect $\beta_{i,h}$ for $i = 1, \cdots, p$ and $h = 1, \cdots, H$ as

$$\beta_{i,h} | \mu_h, \sigma_h^2, R_i, \mathbf{x}_{1:T,i}, \mathbf{y}_{1:T,i} \sim N\left\{ \begin{array}{c} v\left[\frac{\mu_h}{\sigma_h^2} + \frac{1}{R_i} \sum_{t=1}^{T} \sum_{l=1}^{L_t} U_{l,h} \left(y_{t,i,l} - x_{t,i} - \sum_{s \neq h} \beta_{i,s} U_{l,s} \right) \right], \\ v = \left(\frac{1}{\sigma_h^2} + \frac{1}{R_i} \sum_{t=1}^{T} \sum_{l=1}^{L_t} U_{l,h}^2 \right)^{-1} \end{array} \right\}.$$

Finally, the variance parameter σ_0^2 has full conditional

$$\sigma_0^2 | \boldsymbol{\lambda}, \boldsymbol{\eta}^*, \mathbf{a}, \mathbf{z}, \mathbf{x}, \alpha_{\epsilon 0}, \beta_{\epsilon 0} \sim \text{IG}\left(\alpha_{\epsilon 0} + \frac{Tp}{2}, \beta_{\epsilon 0} + \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (x_{t,i} - a_i x_{t-1,i} - \lambda_{ik} \eta_{t,k}^*)^2 \right).$$

The details can be found in Supplement C.5.

3.5.6 Sampling hyperparameters

The hyperparameters $\mu_\lambda, \sigma^2_\lambda, \mu_a, \sigma^2_a$ and μ_h, σ^2_h for hedonics $h = 1, \ldots, H$ can be sampled as follows:

$$\mu_\lambda \left| \mathbf{z}, \boldsymbol{\lambda}, \sigma^2_\lambda, \mu_{\lambda 0}, \sigma^2_{\lambda 0} \sim N \left[\; v \left(\tfrac{\mu_{\lambda 0}}{\sigma^2_{\lambda 0}} + \tfrac{1}{\sigma^2_\lambda} \textstyle\sum_{k=1}^K \sum_{i:z_i=k} \lambda_{ik} \right), v = \left(\tfrac{1}{\sigma^2_{\lambda 0}} + \tfrac{p}{\sigma^2_\lambda} \right)^{-1} \; \right] \right. \tag{3.16}$$

$$\sigma^2_\lambda \left| \mathbf{z}, \boldsymbol{\lambda}, \mu_\lambda, \alpha_{\lambda 0}, \beta_{\lambda 0} \sim \text{IG} \left(\alpha_{\lambda 0} + \frac{p}{2}, \beta_{\lambda 0} + \frac{1}{2} \sum_{k=1}^K \sum_{i:z_i=k} (\lambda_{ik} - \mu_\lambda)^2 \right) \right.$$

$$\mu_a \left| \{a_i\}, \sigma^2_a, \mu_{a0}, \sigma^2_{a0} \sim N \left[v \left(\frac{\mu_{a0}}{\sigma^2_{a0}} + \frac{1}{\sigma^2_a} \sum_{i=1}^p a_i \right), v = \left(\frac{1}{\sigma^2_{a0}} + \frac{p}{\sigma^2_a} \right)^{-1} \right] \right. \tag{3.17}$$

$$\sigma^2_a \left| \{a_i\}, \mu_a, \alpha_{a0}, \beta_{a0} \sim IG \left[\alpha_{a0} + \frac{p}{2}, \beta_{a0} + \frac{1}{2} \sum_{i=1}^p (a_i - \mu_a)^2 \right] \right.$$

$$\mu_h | \beta_{1:p,h}, \sigma^2_h, \mu_{h0}, \sigma^2_{h0} \sim N \left[v \left(\frac{\mu_{h0}}{\sigma^2_{h0}} + \frac{1}{\sigma^2_h} \sum_{i=1}^p \beta_{i,h} \right), v = \left(\frac{1}{\sigma^2_{r0}} + \frac{p}{\sigma^2_r} \right)^{-1} \right] \tag{3.18}$$

$$\sigma^2_h | \beta_{1:p,h}, \mu_h, \alpha_{h0}, \beta_{h0} \sim \text{IG} \left[\alpha_{h0} + \frac{p}{2}, \beta_{h0} + \frac{1}{2} \sum_{i=1}^p (\beta_{i,h} - \mu_h)^2 \right].$$

3.5.7 Sampling the DP hyperparameter

We assume a hyperprior for the DP concentration parameter $\alpha \sim \text{Gamma}(\alpha_\alpha, \beta_\alpha)$ and follow the sampling procedure suggested by [17]. Details are provided in Supplement C.6.

3.6 Computational challenges and strategies

Although marginalizing $\boldsymbol{\pi}, \mathbf{x}$, and $\boldsymbol{\eta}$—i.e., considering a *collapsed* sampler— reduces the dimension of the posterior we explore in our sampling, the marginalization of $\boldsymbol{\pi}$ induces dependencies between the z_i. As such, we must rely on the CRP-based sequential sampling described in Section 3.5.1. 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 $n^{(k)} >> p^{(k)}$.

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 [62] for parallel collapsed MCMC sampling in DP mixture models (DPMM). A similar approach was proposed in [38]. The conventional DPMM assumes that observations x_i with emission distribution $F()$ are drawn as

$$\begin{aligned} G &\sim \text{DP}(\alpha, G_0), \\ \theta_i \mid G &\sim G, \\ x_i \mid \theta_i &\sim F(\theta_i). \end{aligned} \tag{3.19}$$

In order to do exact but parallel MCMC sampling for the DPMM on some P processors, Williamson et al. [62] proposed the following auxiliary variable representation:

$$\begin{aligned} G_j &\sim \text{DP}(\alpha/P, G_0), \\ \boldsymbol{\phi} &\sim \text{Dirichlet}(\alpha/P, \cdots, \alpha/P), \\ \gamma_i \mid \phi &\sim \text{Multinomial}(\boldsymbol{\phi}), \\ \theta_i \mid G, \gamma_i &\sim G_{\gamma_i}, \\ x_i \mid \theta_i &\sim F(\theta_i). \end{aligned} \tag{3.20}$$

The auxiliary variable γ_i assigns data point i to processor γ_i. Theorem 1 in [62] proves that for $\boldsymbol{\phi}$ and G_j defined as in Eq. (3.20), $G := \sum_j \phi_j G_j \sim \text{DP}\left(\sum_j \alpha/P, \frac{\sum_j (\alpha/P) G_0}{\sum_j \alpha/P}\right) = \text{DP}(\alpha, G_0)$. Therefore, the marginal distribution 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 housing price dynamic model, 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 in our model are described in Supplement D.

Beyond parallelizing the sampler, we additionally 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 $\left(\bar{\psi}_{t,i}, L_{t,i}\right)$ 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.

3.7 Model Validation by Simulation

3.7.1 Settings

We first validate our model using simulated data with aspects set to match our real data analysis of Section 3.8. 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 i according to Eqs. (3.3) and (3.5) (see Figure 3.4). Note that the tracts within each cluster have similar price dynamics, as intended by our model. Second, we generated the observed sales prices, $y_{t,i,l}$, according to Eq. (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-

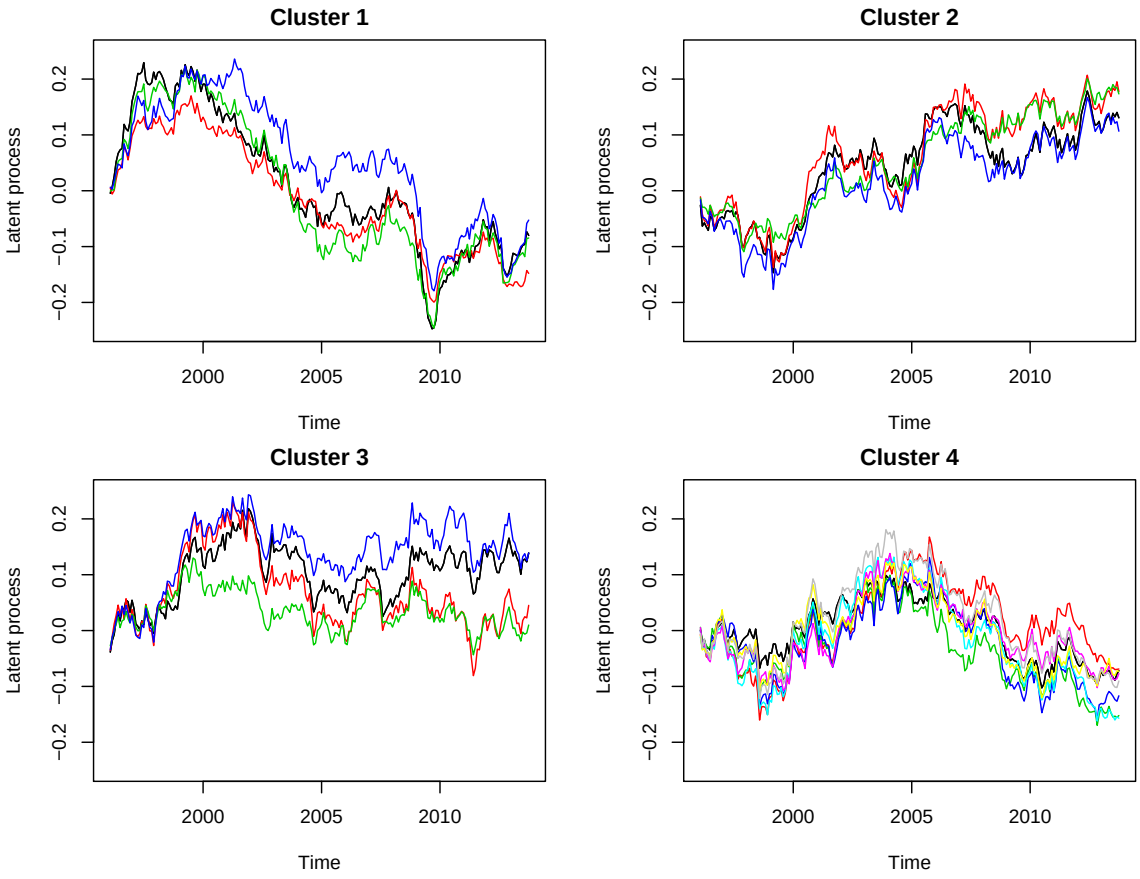


Figure 3.4: Simulated latent price processes for 20 census tracts from 4 clusters. Traces within each plot correspond to specific census tracts in each cluster.

data frequency of observations and house characteristics. The resulting generated sales prices are shown in Figure 3.5.

3.7.2 Results

We ran the MCMC sampler for 1200 iterations on the simulated data. Figure 3.6 shows the normalized Hamming distance between the estimated and true cluster assignments after an optimal mapping between the sets of labels [41], demonstrating successful recovery of the underlying clus-

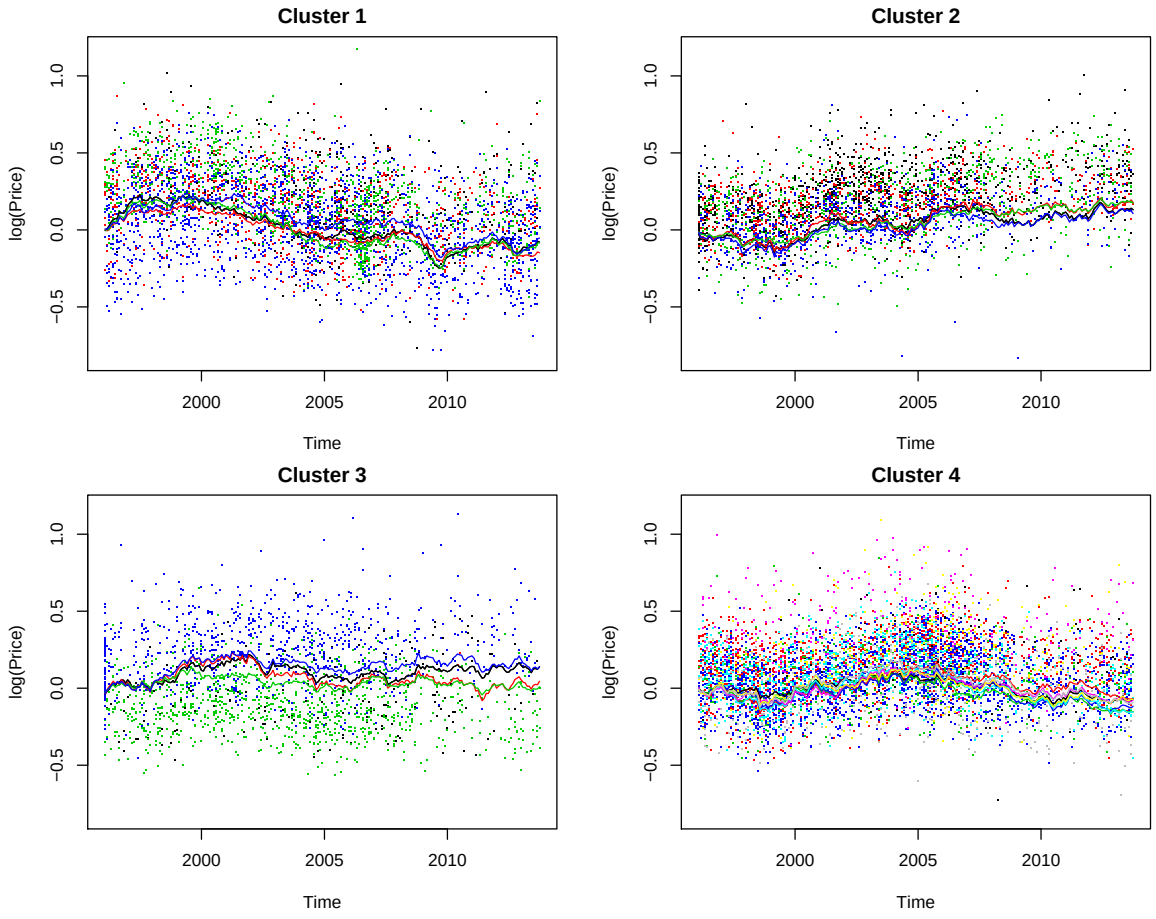


Figure 3.5: Simulated latent process (*solid lines*) and sales prices (*dots*) for the 20 clustered census tracts for each of the 4 ground truth clusters.

ters. We see that our sampler converges very rapidly.

Given sparse observations per month at the census tract level, Figure 3.7 demonstrates that our posterior estimate of the latent processes nicely tracks the true latent dynamics for each census tract. As a baseline comparison, we considered applying a Kalman smoother independently on each census tract. Unsurprisingly, without sharing observations from similar tracts, the baseline approach fails when the observations are sparse. For other census tracts, please refer to Supplement F.

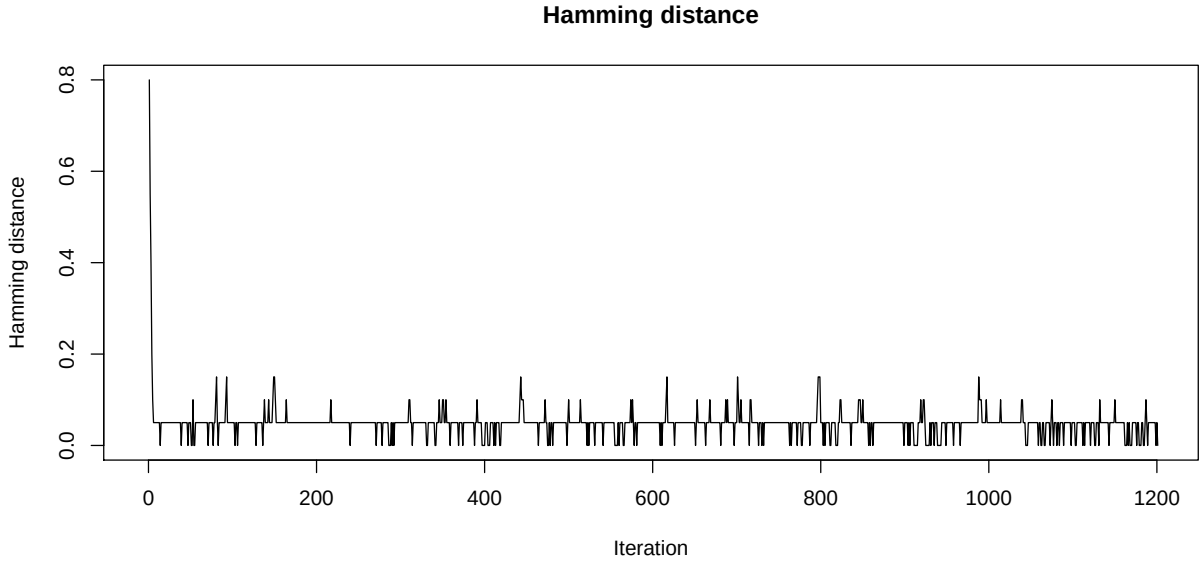


Figure 3.6: Hamming distance between posterior samples of cluster indicators and true cluster memberships (after an optimal mapping) as a function of Gibbs iteration.

To evaluate the importance of the DP clustering beyond the benefits provided by our hierarchical Bayesian dynamic model, we compare results by enabling / disabling clustering in our proposed model. For the latter, we fixed each census tract to form its own cluster and simply did not resample the cluster indicators in our MCMC. Figure 3.8 shows the test set RMSE for predicting the latent trend $\mathbf{x}$ as a function of the number of observations in the census tract. For tracts with fewer observations, the clustering method provides substantial improvement in prediction error. As expected, when observations are abundant, the improvement diminishes.

We also experimented with other simulation scenarios, summarized in Table 3.1. When the latent factor processes have relatively large factor loadings (large μ_λ) leading to large noise variance on the latent price dynamics, the improvement in predicting latent trends $\mathbf{x}$ are very significant compared to the model without clustering. However, even under such scenarios, the improvement in predicting the observations $y_{i,t,l}$ 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

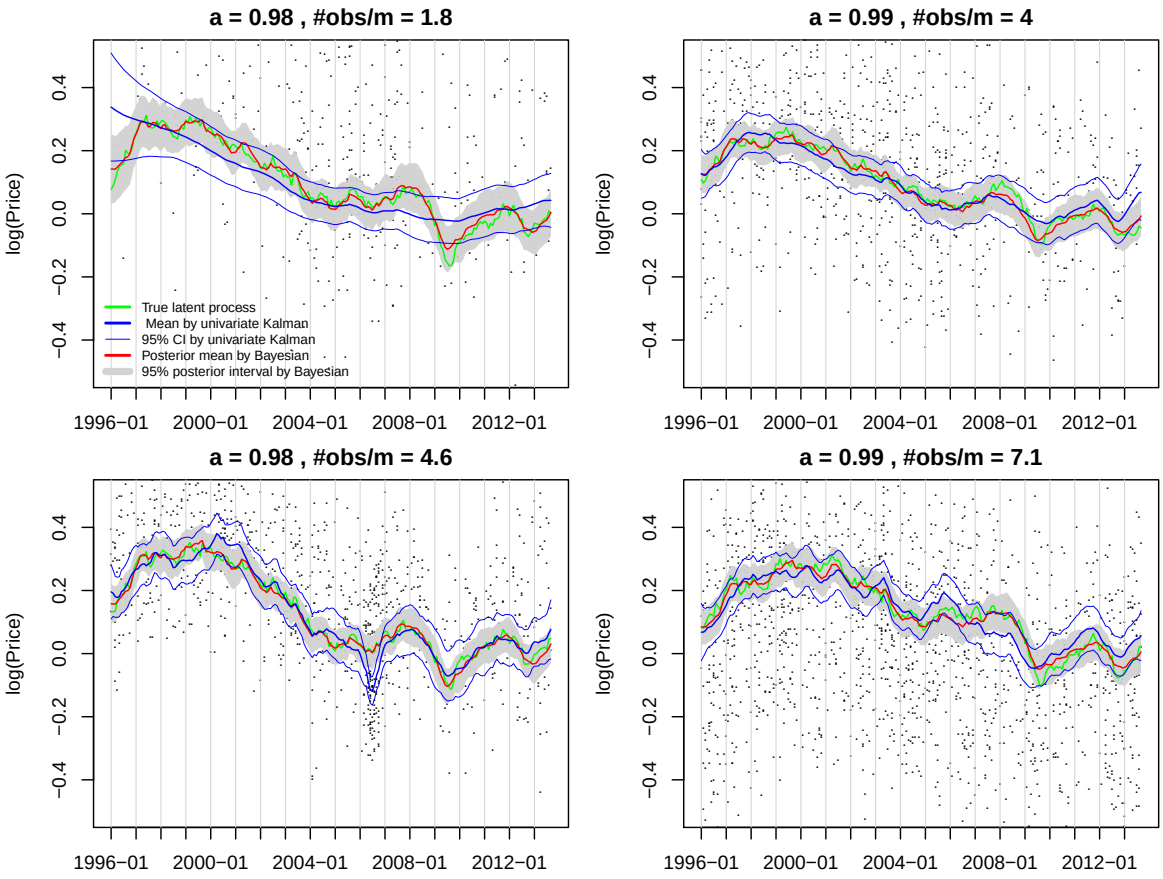


Figure 3.7: Performance of estimating the latent price processes, $x_{t,i}$, shown in green for the 4 census tracts in Cluster 1. The posterior mean and 95% posterior intervals for our proposed non-parametric clustering-based model are shown in red and shaded gray, respectively. The blue lines correspond to the independent Kalman smoother baseline approach.

interested in the intrinsic price dynamics $\mathbf{x}$ themselves, which form our fine-resolution index.

3.8 Housing Data Analysis

We now turn to our housing data analysis based on the City of Seattle data described in Section 3.1. Recall our goal of forming a census-tract level index. For simplicity, we have assumed a separately estimated global trend (g_t in Eq. (3.1)), which captures the city-wide price dynamics, though it

Table 3.1: Three simulation scenarios and results on out-of-sample prediction of latent trends $x_{1:T,i}$ and house prices $y_{i,t,l}$. We compare our proposed Bayesian model both with and without the DP-based nonparametric clustering component.

		No clustering	Clustering	Improvement
$\mu_a = 0.99$, $\mu_\lambda = 0.015$	RMSE in x	0.0234	0.0191	18%
	RMSE in y	0.1192	0.1186	0.5%
$\mu_a = 0.99$, $\mu_\lambda = 0.15$	RMSE in x	0.0737	0.0335	55%
	RMSE in y	0.3375	0.3211	4.9%
$\mu_a = 0.60$, $\mu_\lambda = 0.15$	RMSE in x	0.0786	0.0313	60%
	RMSE in y	0.1335	0.1219	8.7%

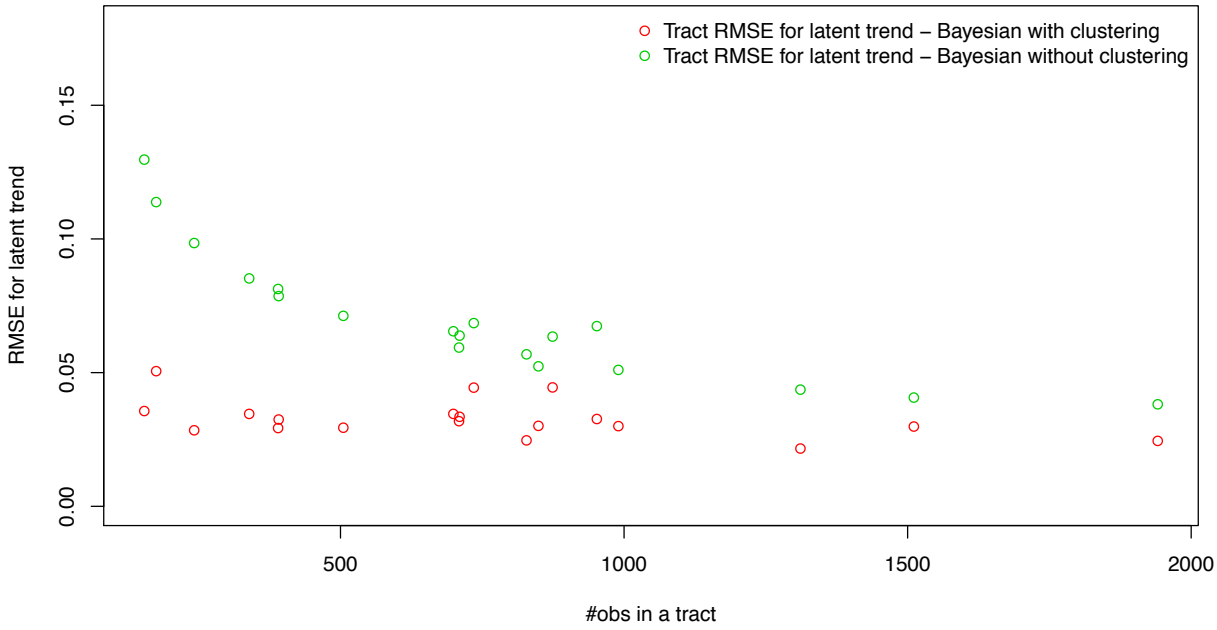


Figure 3.8: Prediction error (RMSE) in latent trend by tract of varying number of observations.

would be straightforward to incorporate joint inference of g_t in our MCMC. For our experiments in this section, we base this global trend on an estimate formed as follows. We first consider a non-tract-specific regression akin to Eq. (3.4) in order to remove the hedonic effects: $y_{t,i,l} = \alpha_0 + \sum_t \alpha_t I(t) + \sum_{h=1}^{H} \beta_h U_{l,h} + v_{t,l}$, where α_t captures the monthly effect and β_h the hedonic effects on the global trend. The noise $v_{t,l}$ is independent across time and sales. Note that in aggregate, we have roughly 640 observations per month on average. After removing the hedonic effects, we then apply the seasonal decomposition approach of [13] to decompose the estimated global trend into a trend component, seasonal component, and noise; we discard the noise term. The resulting global trend (Figure 3.9) has a small but significant seasonal effect. This 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.

To asses 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. The convergence diagnostics are provided in Section 3.8.2.

Figure 3.10 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 3.11. Cluster 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. Cluster 11 and 13 are mostly low-income areas with less expensive housing where the housing recovery has been 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 shows the cluster average index in raw price scale.

For this MAP sample, the University District (U-District) census tract highlighted in Chapter 1 gets assigned to Cluster 3—the largest cluster—driven by "the rich get richer" property of the CRP

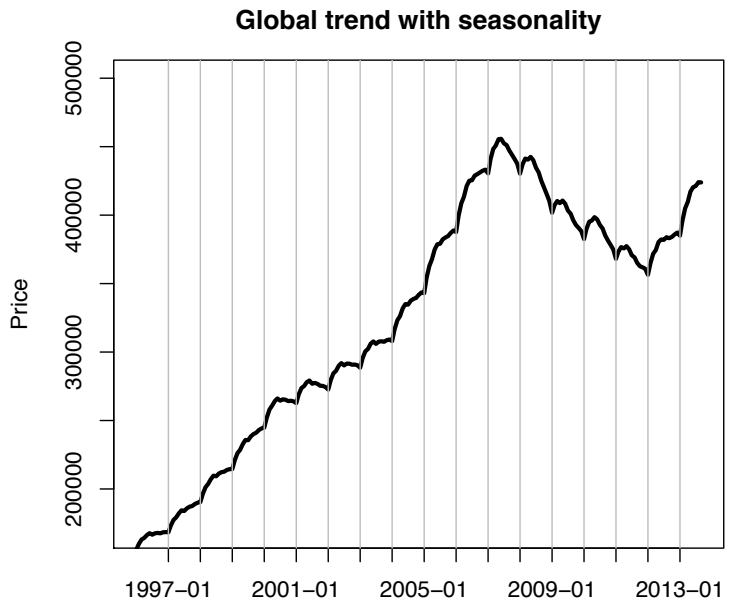


Figure 3.9: Estimated global trend using the seasonality decomposition approach of [13], after adjusting for hedonic effects.

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 similar price dynamic patterns.

3.8.1 Comparison with other methods

We compared our Bayesian nonparametric approach with the Case-Shiller housing index [10] described in Chapter 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 latent process $\mathbf{x}$ in our model, and then fit a regression model with tract-specific hedonic effects as in Eq. (3.4). The estimated

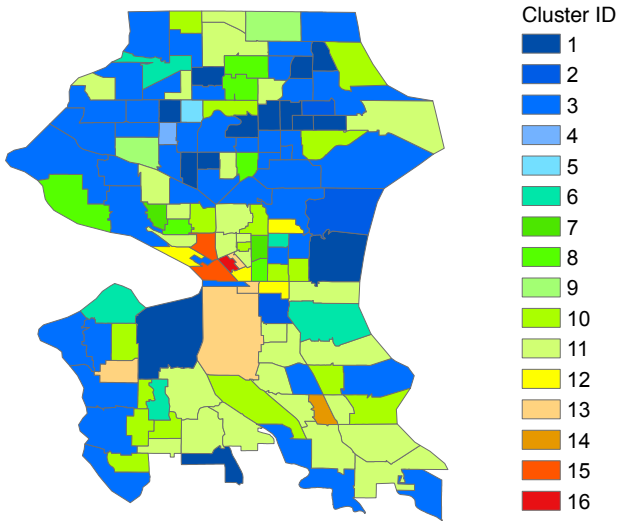


Figure 3.10: Map of clusters under the MAP sample. The cluster labels and associated map colors are selected to indicate the level of deviance of the cluster's average (across tracts) latent trend from the global trend. Blue (1) represents a small deviance while red (16) represents the largest.

hedonic effects together with Case-Shiller index are then used to predict the house prices. Due to the scarcity of repeat sales observations 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 3.2, 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 approximated by our MCMC posterior samples:

$$P(y^*_{t,i}|\mathbf{Y}) = \int_\theta P(y^*_{t,i}|\theta)P(\theta|\mathbf{Y})d\theta \approx \sum_{m=1}^{M} p(y^*_{t,i} \mid \theta^{(m)}), \tag{3.21}$$

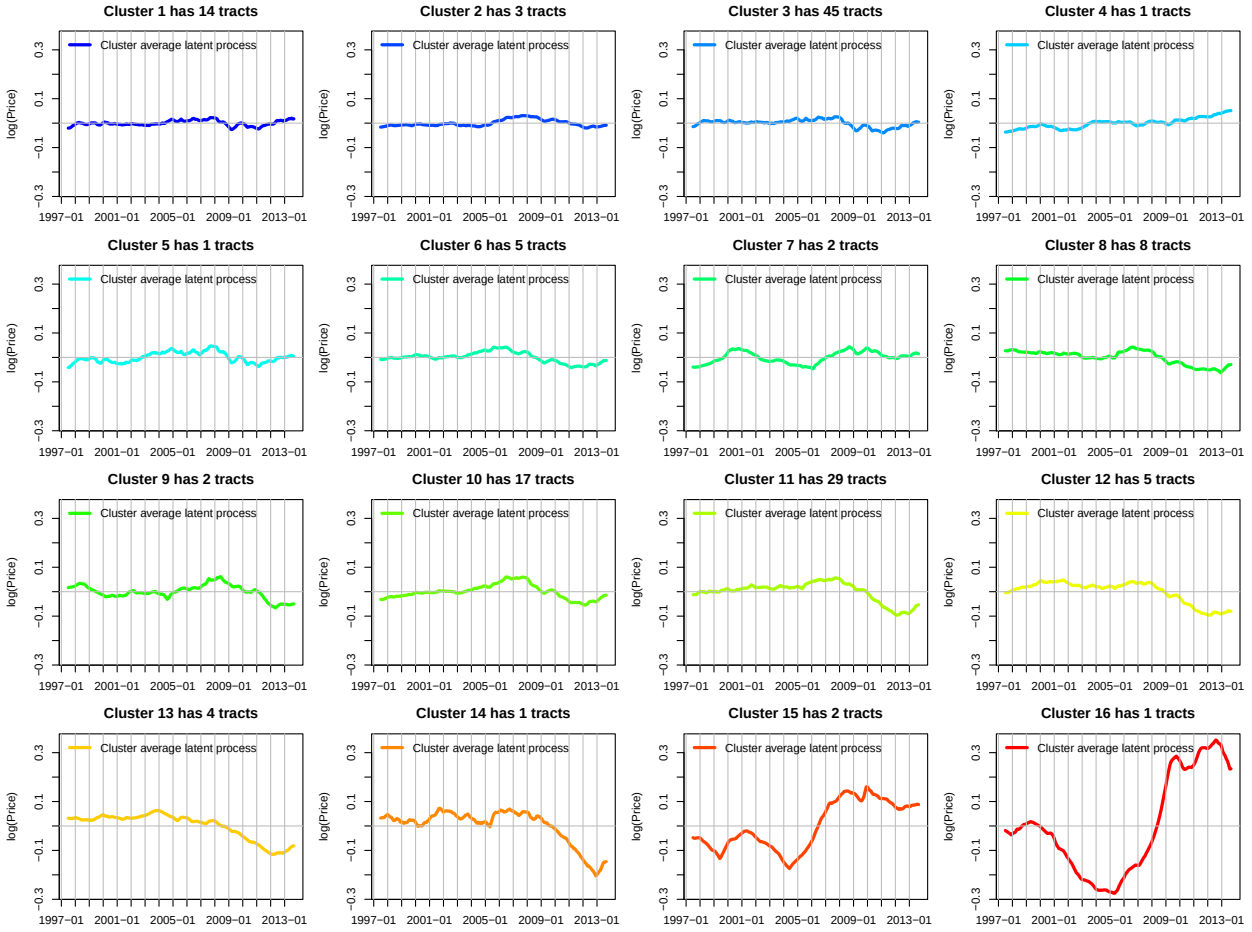


Figure 3.11: Under the MAP sample, cluster-average intrinsic price dynamics computed by averaging $\mathbf{x}_{1:T,i}$ over all i with $z_i = k$ for $k = 1, \ldots, 16$. The color scheme is the same as in Figure 3.10.

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 Eq. (3.4). We then use the mean of these posterior predictive samples as prediction for any house in the test set.

For all of our comparisons, we used the same training and test split. In Table 3.3, we summarize the out-of-sample predictive performance with five metrics: root mean squared error (RMSE) in price, mean / median / 90% quantile of absolute percentage error (Mean APE, Median APE,

Table 3.2: For our predictive performance comparison summarized in Table 3.3, 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 City	Case-Shiller Zip Code	Case-Shiller Census Tract	Bayesian Census Tract
# tracts using	11	121	8	140
# observations using	1,294	26,576	3,248	31,118

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 break the analysis down by deviation of the inferred latent trend from the global trend. For the top 5% tracts with most dramatic local price dynamics (measured in L_2 distance of posterior mean latent trend over time), we see even more dramatic 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 a neighborhood structure, we see in Figure 1.1 that certain regions (e.g. the U-District) do not get shrunk to trends in 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 3.3. It is not surprising to see the most significant improvements being 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

Table 3.3: 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), 90th percentile absolute percentage error (90th APE) and proportion within 10% error (P10).

	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 (1,111 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%

global trend.

Table 3.4 lists the improvement in predictive performance of our Bayesian tract index over using the Case-Shiller index computed at a city or zip code level. The most significant improvement is for houses in tracts with fewer sales (lower 5% tracts). 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

Table 3.4: Predictive performance improvement of our Bayesian tract index over Case-Shiller City and Zip code indices for tracts of different sales frequency, using various measures: mean absolute percentage error (Mean APE) and 90th percentile absolute percentage error (90th APE).

	Improvement over Case-Shiller City index	Improvement over Case-Shiller Zip Code index
Top 5% tracts with most sales (3,569 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%

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 various modeling components on our overall performance. We start by comparing the performance of our approach with simpler dynamical models. In particular, we compare against models that treat each census tract independently. Both use the per-tract dynamics specified in Eqs. (3.3)-(3.4), though one of our comparisons omits the hedonics term. In both cases, the intrinsic price dynamics and associated model parameters are inferred independently for each census tract using a Kalman smoother embedded in an expectation maximization (EM) procedure. The results are summarized in Table 3.5. (Note that the last column of Table 3.5

Table 3.5: Predictive performance comparison on variants of the proposed Bayesian nonparametric model using the same metrics as in Table 3.3.

	Univariate Kalman Smoother w/o hedonics	Univariate Kalman Smoother w/ 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

coincides with that of Table 3.3, and is repeated for readability.) We see reduced predictive performance at each stage of breaking down our Bayesian dynamical model. Additionally, as motivated by the results of Table 3.1, 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.

We now turn to the central focus of the Chapter 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) as a proxy. As mentioned in Chapter 1, the index is formed by taking the median of Zillow house-level estimates of value and provides a stable empirical estimate at fine-scale regions. 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. 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 the a zip code.

Figure 3.12 shows that the Bayesian index with (*red line*) and without (*green line*) the DP-based

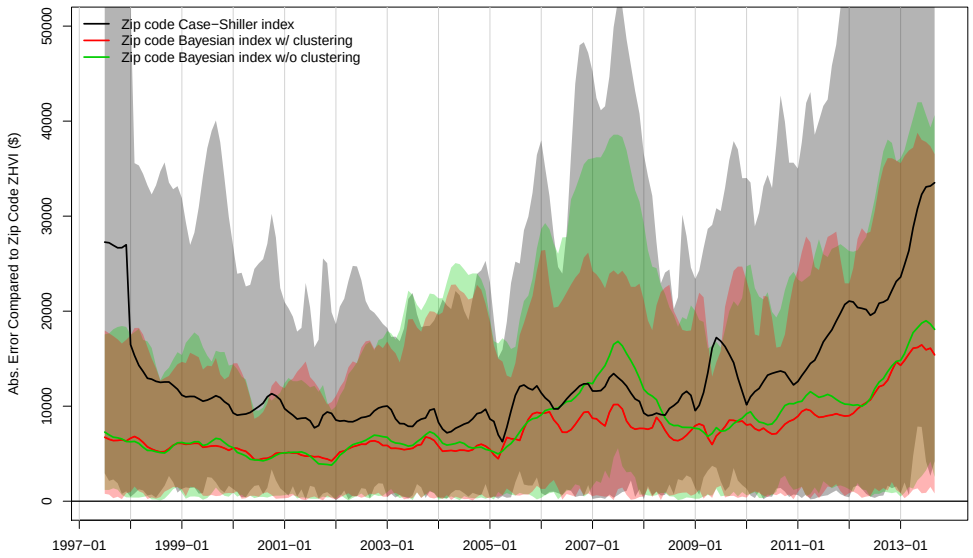


Figure 3.12: Treating the Zillow Home Value Index (ZHVI) as a surrogate ground truth, errors of various index methods relative to ZHVI at the zip code level. Examining performance across zip codes, the mean absolute error (*red line*) and 90% interval (*shaded red*) of our proposed Bayesian index is compared to that of the Bayesian index without the DP-based nonparametric clustering component (*green* and *shaded green*) and the Case-Shiller zip code index (*black* and *shaded gray*). The performance of the Bayesian methods are based on posterior mean estimates.

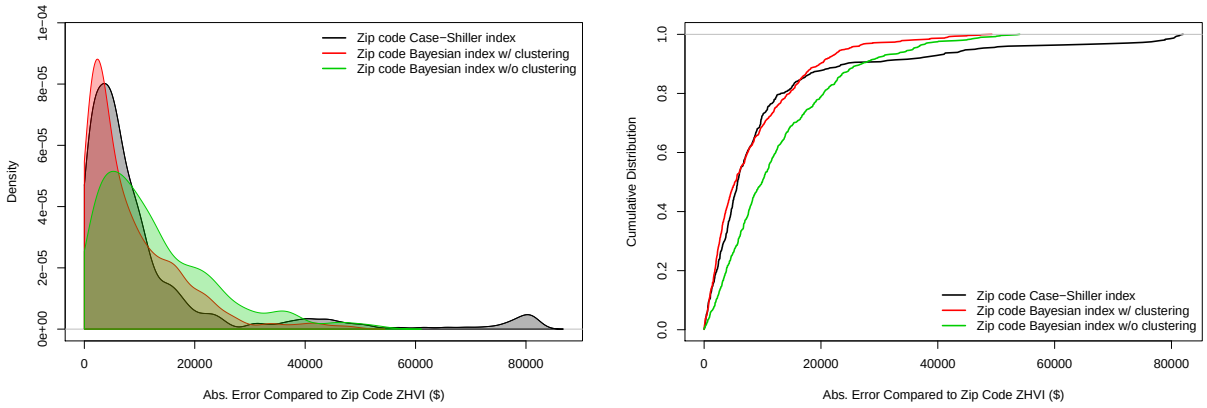
nonparametric clustering component have significantly different performance during the 2006-2007 period, and to a lesser extent in 2010-2011. In 2006-2007, the Seattle housing market was in a boom period with high sales and volatility (see Figure G.3 in Supplement G.1). After the bust, the housing market started to stabilize in 2010-2011. The market boom and subsequent stabilization were manifested in the different housing sectors in disparate ways. The DP-based clustering, especially in the highly volatile year of 2007, is more closely aligned with the ZHVI, since it is better able to capture the dynamics of the change in value for different housing sectors. This is because the non-clustering Bayesian hierarchical model shrinks the census tracts with few observations towards a global mean, whereas our clustering model allows atypical census tracts to be shrunk towards a more informed structure, such as the one shown in Figure 3.11.

Figure 3.12 also compares the zip code Case-Shiller index (*black line*), which is significantly more different from the ZHVI than the proposed Bayesian index (*red line*) during all times. Without any kind of sharing information and shrinkage across different regions, the Case-Shiller index has the widest interval among the three methods. The beginning and the end of the study periods are extremely challenging for Case-Shiller index, because of having fewer repeated sales available at the boundary of the study period. In the middle of the series, the difference between Case-Shiller and the ZHVI is especially large during the highly volatile period of 2007.

Figure 3.13 shows that Case-Shiller (*black*) has a long-tailed distribution of absolute error relative to ZHVI in contrast to the shrinkage provided by the other two Bayesian methods, with the clustering approach clearly the best. In particular, looking the cumulative distribution of Figure 3.13(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, the Bayesian non-clustering model also has fewer low-error zip codes relative to the Case-Shiller baseline. In contrast, our proposed Bayesian nonparametric clustering index has as many low-error zip codes as Case-Shiller, tracking this baseline index in the low-error range, but also has many fewer high-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.

3.8.2 Convergence

Convergence of the MCMC chains is assessed using the potential scale reduction factor (Gelman Rubin statistics) [23], calculated with the "coda" package in R. The Gelman Rubin statistics in Table 3.6 are all less than the threshold of 1.1, which is an indication of good convergence. Visual examination of the trace plots also show good mixing and convergence. Some selected trace plots are provided in Appendix G.



(a) (b)

Figure 3.13: A more detailed examination of the distribution of errors in Figure 3.12 during 2007. (a) Estimated density and (b) associated cumulative distribution of the absolute error.

Table 3.6: Convergence diagnostics using the potential scale reduction factor (Gelman Rubin statistic) by Gelman & Rubin [23]. The reported statistics are smaller than 1.1, which is an indication of good convergence of the MCMC chains.

Parameter	Gelman Rubin Statistic
σ_0	1.050
α	1.005
μ_a	1.032
σ_a	1.010
μ_λ	1.049
σ_λ	1.033
$a_i, i = 1, \cdots, p$	≤ 1.062
$\lambda_i, i = 1, \cdots, p$	≤ 1.095

3.9 Discussion

We presented a method for constructing a housing index at fine-scale geographical units, with improved space-time adjustment and specificity than existing approaches. In particular, the extreme

sparsity of transactions at a fine spatio-temporal granularity poses a significant modeling challenge. The 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 a cluster, providing a form of multiple shrinkage of individual trend estimates for each region. Our main contribution includes providing a housing index at a resolution that was un-achievable before, enabling a flexible, data-driven clustering of correlated data streams and efficient computation through parallel MCMC.

We model dependence among tracts through a block diagonal covariance matrix. The intuition is that we seek to capture the dependence of simultaneous price movement of regions which might be driven by same underlying factors such as social economic status, urban development plans, etc. There are other model choices for building dependence, for instance, a full covariance matrix, a low-rank covariance matrix, or a conditional independencies through sparsity of the precision matrix. In the following, we will discuss each of these alternatives. The full covariance matrix is a superset of a block diagonal covariance matrix. However, a full covariance without imposed sparsity looses the interpretable relationship among tracts, in the sense of clustering and independence across clusters. Additionally, the curse of dimensionality might cause unrobust estimates of the full covariance. A low-rank approximation to the full covariance can solve the robustness issue. However, working with a full covariance or a low-rank covariance invokes costly computation since the likelihood evaluation involves a Kalman-filter-like algorithm whose complexity grows cubicly with the state dimension. In contrast, our proposed block diagonal covariance matrix applies the likelihood evaluation independently to each cluster, therefore leads to less computation by working with a lower state dimension. Another computationally efficient way to introduce independencies is using the graphical model through a sparse precision matrix. It encodes the conditional independence between tracts, which is not the type of dependence structure we seek to capture as simultaneous price movement driven by the same source of exogenous factors.

Our clustering-based dynamical model address the limitation on the reliance of repeated sales. Our model provides the ability of tracking 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 spacial 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 jointly estimating the non-stationary global trend, and 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 pre-specified geographic model combined with our cluster-induced heterogeneous spatial structure as a model of the residuals.

Chapter 4

HOUSING INDEX MODEL EXTENSIONS

Having developed the Bayesian dynamical model for hyperlocal housing index in Chapter 3, we now turn to exploring various modifications to the model and analyzing their effects. The first is to jointly estimate the global market trend together with the local dynamics as one unified model in Section 4.1. This provides a fully Bayesian approach in contrast to the two-stage approach presented in Chapter 3. The second is to test the model sensitivity to the house covariates in Section 4.2. In Chapter 3, we included finished square feet, square feet of lot size and number of bathrooms. A natural question is what impact each of these has on the inferred index, and whether it might be beneficial to add additional covariates. Lastly, we explore possible extensions to our underlying dynamical model. In Chapter 3, we assumed a first-order latent autoregressive process for the intrinsic house price dynamics. However, our inferred autoregressive coefficients were very close to 1 (the stationary/non-stationary boundary), indicating a longer memory process. We discuss ways of incorporating long-memory processes into our framework in Section 4.3.

In all of these cases, we see that the model of Chapter 3 represented a reasonable first cut at a hyper-local housing index, but that refinements to the model are possible.

4.1 Modeling the Global Trend

In Chapter 3, we assumed a fixed, pre-calculated global trend, and extracted it from the data as a pre-processing step. In this section, we propose to jointly model the non-stationary global market trend and the stationary local price dynamics. The unified Bayesian framework can properly address the parameter uncertainties jointly and coherently, including the uncertainties in the global trend.

Similar to many economic time series, the housing market trend is non-stationary, based on

the estimated global trend on the data of Seattle City in Chapter 3. In general, a non-stationary time series can be modeled as a deterministic trend or a stochastic trend. The time series clustering model in [45] models the non-stationary trend with a quadratic form. Such simple polynomial forms may not be flexible enough to fit and capture the long term trend in housing market such as boom, bust and recovery. Therefore we choose a natural cubic spline model to capture the complex global market dynamics.

Alternatively, one could consider a stochastic trend, such as the autoregressive integrated moving average (ARIMA) model. The ARIMA model is a popular model choice for economic time series [47]. Under an ARIMA(p, d, q) model, the observed series is non-stationary process, but its d-th differenced series follows a stationary ARMA$(p - d, q)$ process. The differencing parameter $d = 1$ and $d = 2$ are frequently used for economic time seres, implying that the local behavior of the series is independent of its level (for $d = 1$) and of its level and slope (for $d = 2$) [8]. We focus on natural cubic splines to model the global trend, and leave examination of ARIMA processes for future work.

4.1.1 A Model for a Non-stationary Global Trend

Following the same notations and definitions used in Chapter 3, we recall our dynamic model. Each census tract i can have multiple house sale observations $\tilde{y}_{t,i,l}$, where t denotes the transaction time and l is the index of a house. 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 footage, number of bedrooms):

$$\tilde{x}_{t,i} = g_t + a_i(\tilde{x}_{t-1,i} - g_{t-1}) + \epsilon_{t,i} \quad \epsilon_{t,i} \sim \mathcal{N}(0, \sigma_i^2) \tag{4.1}$$

$$\tilde{y}_{t,i,l} = \tilde{x}_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,h} + v_{t,i,l} \quad v_{t,i,l} \sim \mathcal{N}(0, R_i). \tag{4.2}$$

Our discrete-time model is indexed monthly, which represents the frequency at which house sales are recorded. The term g_t is the global market trend that captures overall, non-stationary behavior of the time series.

The local dynamics $x_{t,i}$ modeled in Chapter 3 represent the deviance of the latent dynamics of census tracts from the global market trend: $x_{t,i} \equiv \tilde{x}_{t,i} - g_t$. Equivalently,

$$\tilde{x}_{t,i} = g_t + x_{t,i}. \tag{4.3}$$

Substituting Eq. (4.3) into Eq. (4.2), we get

$$\tilde{y}_{t,i,l} \quad = \quad g_t + x_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,h} + v_{t,i,l}. \tag{4.4}$$

By re-arranging the terms, we get

$$\tilde{y}_{t,i,l} - x_{t,i} - \sum_{h=1}^{H} \beta_{i,h} U_{l,h} \quad = \quad g_t + v_{t,i,l}. \tag{4.5}$$

Let $r_{t,i,l} = \tilde{y}_{t,i,l} - x_{t,i} - \sum_{h=1}^{H} \beta_{i,h} U_{l,h}$ be the residual of each sales price after accounting for the local time trend and house hedonics. Eq. (4.5) has the following simple form

$$r_{t,i,l} \quad = \quad g_t + v_{t,i,l}. \tag{4.6}$$

Recalling our MCMC in Chapter 3, at a given iteration, we obtain a sample of $x_{t,i}, \beta_{i,1:H}$ that can be used to compute the pseudo-observations $r_{t,i,l}$. With the pseudo-observations, Eq. (4.6) simply represents a standard regression setting with time as the predictor. We now seek a model for the global trend g_t. To promote smoothness while allowing flexibility, we choose to use natural cubic splines (NCS) [57] 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, piece-wise cubic splines between knots and linear splines at the boundaries. Background of natural cubic splines is provided in Chapter 2. More specifically, we propose the following model for the non-stationary global trend

$$g_t = w_1 B_1(t) + \cdots + w_{N_B} B_{N_B}(t) + s_2 m_2(t) + \cdots + s_{12} m_{12}(t), \tag{4.7}$$

where $B_j, j = 1, \cdots, N_B$ are the basis functions and $m_j(t) = I(t = j), j = 2, \cdots, 12$ denotes the j-th monthly effect. From data points $(t, r_{t,i,l})$, the global trend can be fit straightforwardly, if the number of knots is known. Section 4.1.2 describes the exploratory analysis to determine the number of knots.

4.1.2 Exploratory Analysis for the Global Trend

We experimented on the residual data $(t, r_{t,i,l})$ from the posterior samples of the model in Chapter 3. Figure 4.1 shows the fit of the seasonal global trend model in Eq. (4.7) with different number of knots for NCS, where it clearly exhibits the monthly effects within each year. Figure 4.2 shows the estimated smoothed global trend after removing the monthly effects, in order to visualize the fit of NCS basis functions with different number of knots alone. With too few knots, for example three, the curve is more constrained and the fit is unsatisfactory. The model selection procedure suggests using 9 interior knots for the model in Eq. (4.7), according to the Bayesian Information Criterion (BIC), as shown in Figure 4.3. The 9 interior knots generate 10 spline basis functions and one intercept basis. Therefore, we will be using $N_B = 11$.

4.1.3 Prior Specification for Global Trend

In this section, we focus on the non-stationary smooth trend over the long term horizon, as modeled by natural cubic splines (NCS), assuming the monthly effects are pre-estimated and removed. The priors for the coefficients of NCS are specified as

$$w_j \sim \mathcal{N}(0, \sigma^2_{w_j}), \quad j = 1, \cdots, N_B. \tag{4.8}$$

The hyper priors are specified as

$$\sigma^2_{w_j} \sim \text{IG}(\alpha_{w0}, \beta_{w0}), \quad j = 1, \cdots, N_B. \tag{4.9}$$

4.1.4 MCMC Posterior Computation with Global Trend

Our posterior computation is implemented via a Gibbs sampling scheme building directly on the sampler derived in Chapter 3. Define $\mathbf{g} = \{g_t\}, \mathbf{w} = \{w_j\}$ and $\mathbf{s} = \{s_j\}$. The global trend $\mathbf{g}$ is a function of the parameters $\mathbf{w}$ and $\mathbf{s}$. Let $\boldsymbol{\psi} = \{\mathbf{a} = \{a_i\}, \boldsymbol{\lambda} = \{\lambda_{ik}\}, \mathbf{R} = \{R_i\}, \boldsymbol{\beta} = \{\beta_{i,h}\}, \sigma_0^2\}$ and $\boldsymbol{\psi}^{(k)}$ be the associated subset of parameters corresponding to the k-th cluster based on assignments $\mathbf{z} = \{z_i\}$.

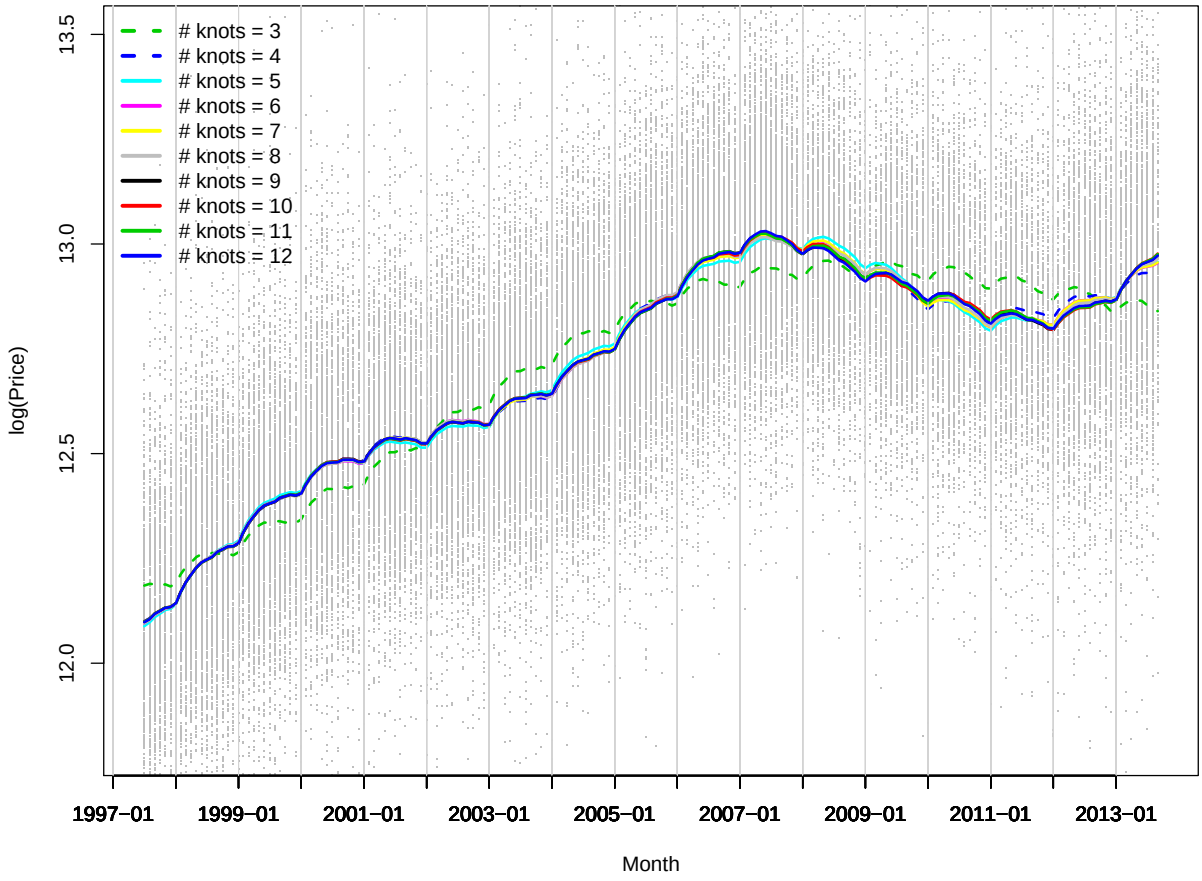


Figure 4.1: The fit of the seasonal global trend model in Eq. (4.7) with different number of knots for NCS. It clearly exhibits the monthly effects within each year.

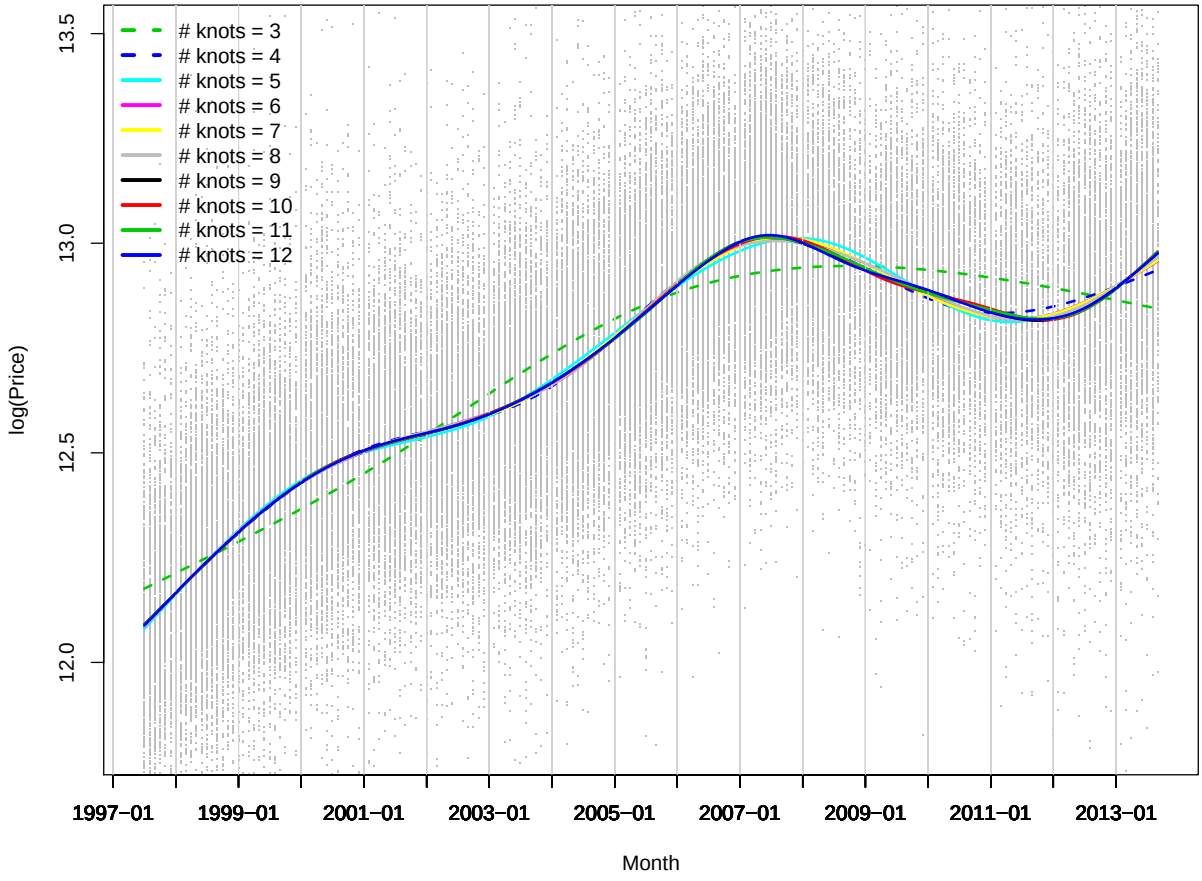


Figure 4.2: The estimated smooth global trend after removing the monthly effects, for visualizing the fit of NCS basis functions with different number of knots.

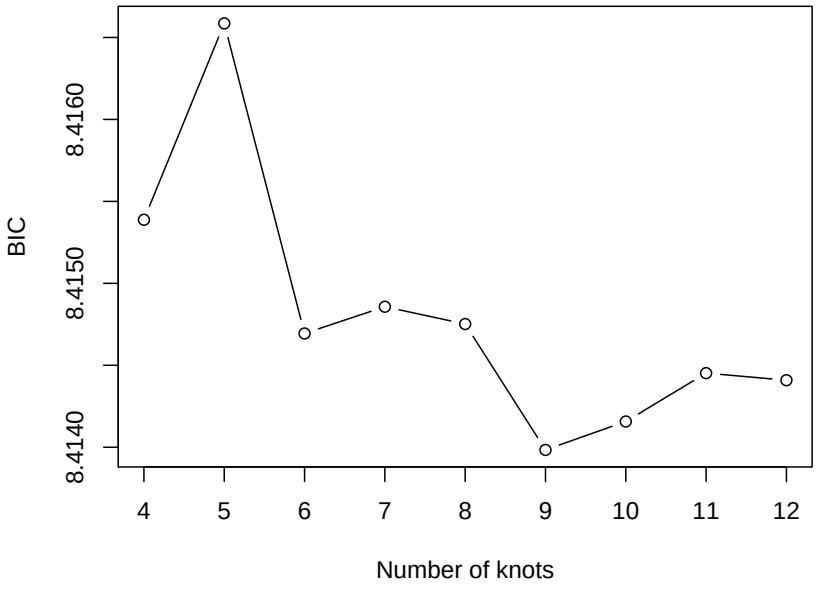


Figure 4.3: Model selection of the number of knots by BIC, in the natural cubic spline model with monthly effects for the global trend.

Based on the local latent dynamics $\mathbf{x}$ and house hedonic effects $\boldsymbol{\beta}$, we can form the residuals $\mathbf{r} = \{r_{t,i,l}\}$, according to $r_{t,i,l} = \tilde{y}_{t,i,l} - x_{t,i} - \sum_{h=1}^{H} \beta_{i,h} U_{l,h}$. Given these pseudo observations $\mathbf{r}$, we then sample the conditional posterior of the global trend $\mathbf{g}$. Our Gibbs sampler is outlined as follows:

1. Sample $z_i = k|\mathbf{z}_{-i}, \alpha, \tilde{\mathbf{y}}, \boldsymbol{\psi}, \mathbf{g}$. Note 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 $\mathbf{x}^{(k)}|\mathbf{z}, \tilde{\mathbf{y}}^{(k)}, \boldsymbol{\psi}^{(k)}, \mathbf{g}$ and $\boldsymbol{\eta}^*|\mathbf{z}, \mathbf{x}, \boldsymbol{\psi}$.

3. Sample $\boldsymbol{\psi}^{(k)}|\mathbf{z}, \tilde{\mathbf{y}}^{(k)}, \mathbf{x}^{(k)}, \boldsymbol{\eta}^{*(k)}, \mathbf{g}$

4. Obtain a sample of $\mathbf{g}$ by sampling $\mathbf{w}|\tilde{\mathbf{y}}, \mathbf{x}, \boldsymbol{\psi}, \mathbf{s}$.

5. Discard $\mathbf{x}$ and $\boldsymbol{\eta}^*$ to sample hyperparameters conditional on $\boldsymbol{\psi}, \mathbf{z}$.

4.1.5 Sampling the coefficients of Natural Cubic Splines in the global trend

By the model specification of the global trend in Eq. (4.6) and Eq. (4.7),

$$r_{t,i,l} = \mathbf{w}_{-j}^T \mathbf{B}_{-j}(t) + w_j B_j(t) + \mathbf{s}^T \mathbf{m}(t) + v_{t,i,l}, \quad v_{t,i,l} \sim \mathcal{N}(0, R_i). \tag{4.10}$$

Therefore,

$$r_{t,i,l} \sim \mathcal{N}\left(\mathbf{w}_{-j}^T \mathbf{B}_{-j}(t) + w_j B_j(t) + \mathbf{s}^T \mathbf{m}(t), R_i\right). \tag{4.11}$$

By rearranging the terms, we get

$$\frac{r_{t,i,l} - \mathbf{w}_{-j}^T \mathbf{B}_{-j}(t) - \mathbf{s}^T \mathbf{m}(t)}{B_j(t)} \sim \mathcal{N}\left(w_j, \frac{R_i}{B_j^2(t)}\right), \quad \text{i.i.d. across } l. \tag{4.12}$$

By conjugacy, the posterior distribution for the coefficient of the NCS in the global trend is

$$
\begin{aligned}
p(w_j|\mathbf{r},\mathbf{s},\mathbf{R},\sigma^2_{w_j}) &\propto \mathcal{N}(w_j|0,\sigma^2_{w_j})\prod_{t,i,l}\mathcal{N}\left(\frac{r_{t,i,l}-\mathbf{w}^T_{-j}\mathbf{B}_{-j}(t)-\mathbf{s}^T\mathbf{m}(t)}{B_j(t)}\,\middle|\,w_j,\frac{R_i}{B^2_j(t)}\right)\\
&\propto \mathcal{N}\left\{\begin{array}{c} v\left[\sum_{i=1}^p \frac{\sum_{t,l}B_j(t_{t,i,l})\cdot\left(r_{t,i,l}-\mathbf{w}^T_{-j}\mathbf{B}_{-j}(t)-\mathbf{s}^T\mathbf{m}(t)\right)}{R_i}\right], \\ v=\left[\frac{1}{\sigma^2_{w_j}}+\sum_{i=1}^p\frac{\sum_{t,l}B^2_j(t_{t,i,l})}{R_i}\right]^{-1}\end{array}\right\}. \qquad (4.13)
\end{aligned}
$$

4.1.6 Housing Data Analysis with Incorporating Global Trend

We now examine the impact of the joint global trend estimation. Figure 4.4 shows the posterior of the Bayesian smoothed global trend, along with the Case-Shiller Index and Zillow Home Value Index (ZHVI). The posterior mean of our global trend is more in alliance with ZHVI, which we believe is a better estimate than Case-Shiller Index, since ZHVI is calculated based on all homes while Case-Shiller Index is computed with homes of repeated sales. The result matches the analysis of Chapter 3 (See Figure G.2).

To evaluate the prediction performance, we compare the baseline method, described in Chapter 3, of a pre-calculated global trend to the extended model with global trend joint estimation. Table 4.1 shows that the predictive performance metrics of the two models are quite close. The extended model has slightly worse prediction in the metrics of RMSE, mean and median APE, but has slightly better performance in the metrics of 90-th APE and proportion within 10% of error. The latter can be attributed to appropriately including the variance of global trend in the joint model. We argue that the improvement on the 90-th APE and P10 metrics, which focus on the tail part of the prediction error distribution (predictions with the largest errors), shows that our proposed index model is a meaningful extension over the baseline as it improves the worst-case scenarios in the prediction model. In fact, improving the worst-case has been an important area of study for the house price prediction application.

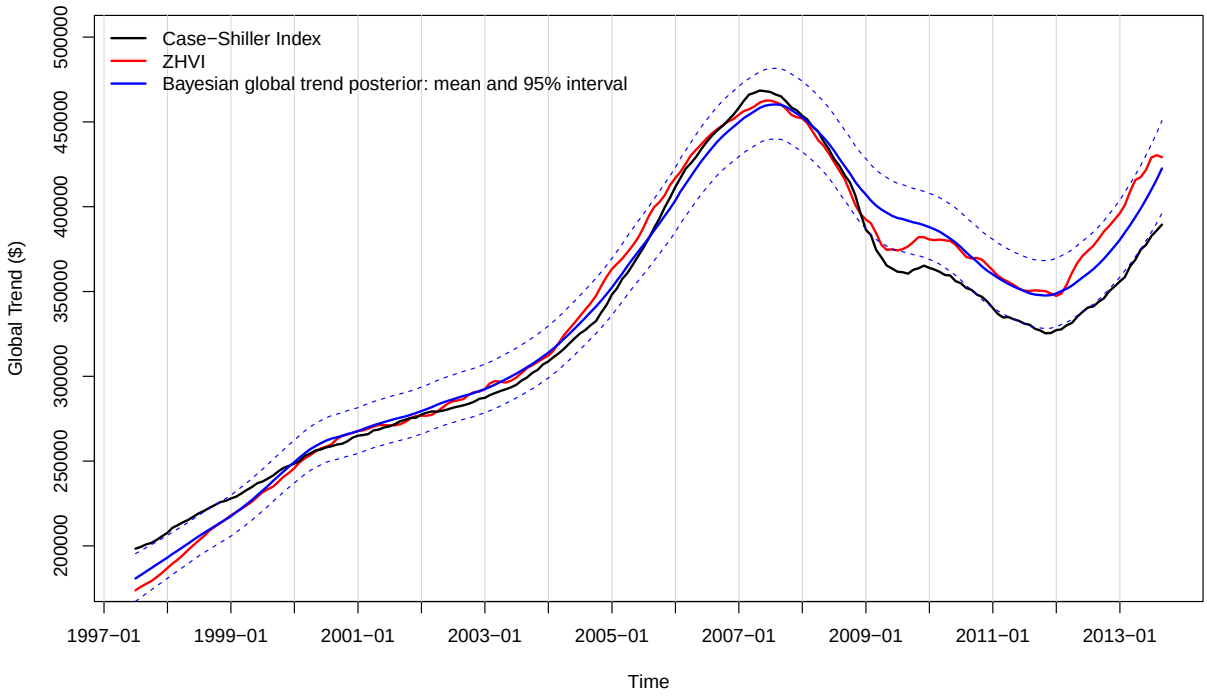


Figure 4.4: Comparison of the Bayesian global trend posterior with Case-Shiller Index and Zillow Home Value Index (ZHVI)

Table 4.1: Predictive performance comparison on Bayesian housing index with pre-calculated global trend as in Chapter 3 and with the global trend joint estimation of this Chapter.

	Bayesian Index with pre-calculated Global Trend	Bayesian Index with Global Trend joint Estimation
RMSE	122,026	122,083
Mean APE	0.1633	0.1635
Median APE	0.1231	0.1237
90th APE	0.3422	0.3414
P10	0.4183	0.4198

4.2 Model Sensitivity to Hedonics

The experiment of Chapter 3 used three covariates: `finished sqft`, `lot size`, and `number of bathrooms`. We think house hedonics play an important role in constructing a housing index from sales data. By specifying the appropriate hedonics effects, the model can do a better job at taking out the source of variance and is able to focus on tracking the price dynamics. To test the sensitivity of the resulting local index to the hedonics selected, we remove one covariate at a time from the baseline model of Chapter 3. The rest of the model is kept unchanged. Table 4.2 shows the predictive performance for models with different covariate combinations. Unsurprisingly, the predictive performance gets worse when we remove hedonics from the baseline model. However, the degree of performance downgrade varies. We see that the finished sqft is the most sensitive factor, as the RMSE increases from $122,026$ to $165,835$ and mean Absolute Percentage Error (APE) increases from 0.1633 to 0.2084. When we remove the lot size hedonic, the prediction gets slightly worse, RMSE from $122,026$ to $127,616$, and mean APE from 0.1633 to 0.1736. The model is least sensitive to the number of bathrooms. The RMSE changes slightly from $122,026$ to $123,471$ and mean APE changes from 0.1633 to 0.1662. Another interesting setting is to add more covariates to the model, such as age, number of stories, and number of bedrooms. We obtain better prediction performance with the model of six covariates by all metrics. The RMSE is reduced

Table 4.2: Predictive performance for Bayesian nonparametric models with different covariates. Removing one covariate at a time from the baseline model worsens the prediction performance as seen in the model variants 1 to 3. Adding three more covariates to the baseline model (variant 4) yields better prediction.

Covariates	Baseline	Variant 1	Variant 2	Variant 3	Variant 4
sqft,	✓	.	✓	✓	✓
lot size	✓	✓		✓	✓
# bathroom	✓	✓	✓		✓
age, #stories, #bedrooms					✓
RMSE	122,026	165,835	127,616	123,471	**119,557**
Mean APE	0.1633	0.2084	0.1736	0.1662	**0.1506**
Median APE	0.1231	0.1552	0.1319	0.1253	**0.1141**
90th APE	0.3422	0.4373	0.3619	0.3469	**0.3168**
P10	0.4183	0.3406	0.3945	0.4145	**0.4502**

from $122,026$ to $119,557$, and mean APE from 0.1633 to 0.1506. By accounting for the source of variance more appropriately, the model is able to track the price dynamics better and yield more accurate predictions.

4.3 Model Extension: AR(m) for the Index Process

The proposed model in Chapter 3 assumes an AR(1) process for the intrinsic price dynamics, as in Eq. (3.3). The posterior computation for the model reveals that the autoregressive coefficients $\{a_i\}$ for the AR(1) process are close to 1, with posterior mean $\mu_a = 0.9986$. This suggests a long memory process for the intrinsic price. A possible extension for capturing longer memory is to model the latent price dynamics as an ARFIMA (autoregressive fractionally integrated moving average) procees. ARFIMA is first introduced in [28] and widely used in Financial time series and hydrology. It generalizes the classic ARIMA(p, d, q) (autoregressive integrated moving average) model by allowing the differencing parameter d to be fractional, where p and q are the autoregres-

sive and moving average parameters. The long memory property of ARFIMA is characterized by the autocorrelation decaying at a hyperbolic rate, in contrast to decaying faster at an exponential rate with ARIMA model. The paper [28] also shows that a stationary ARFIMA(p, d, q) can be represented as an autoregressive process with infinite degrees, AR(∞). To compute the coefficients in AR(∞) process, [12] considers an approximation by truncating up to a lag m. Therefore in our housing index model, a simple extension for capturing long memory is to model the latent price dynamics as an AR(m) process. For the choice of the truncation lag m, Bondon and Palma [7] show that the quality of the AR(m) truncation to ARFIMA(p, d, q) depends asymptotically on the differencing parameter d. Grassi and Magistris [26] find that in practice the convergence is good, even with small values of m for different choices of d. Based on the simulated data, in order to get unbiased estimation, they conclude to use the truncation lag $m = 30$ for time periods less than 500 and $m = 50$ for time periods greater than 500.

The extended model becomes

$$y_{t,i,l} = x_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,h} + v_{t,i,l}, \quad v_{t,i,l} \sim \mathcal{N}(0, R_i) \tag{4.14}$$

$$x_{t,i} = a_i^{(1)} x_{t-1,i} + a_i^{(2)} x_{t-2,i} + \cdots + a_i^{(m)} x_{t-m,i} + \epsilon_{t,i} \tag{4.15}$$

The extended model can be represented as a state space model via

$$y_{t,i,l} = \mathbf{C}\mathbf{q}_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,h} + v_{t,i,l}, \quad v_{t,i,l} \sim \mathcal{N}(0, R_i) \tag{4.16}$$

$$\mathbf{q}_{t,i} = \mathbf{A}\mathbf{q}_{t-1,i} + \boldsymbol{\tau}, \tag{4.17}$$

where

$$\mathbf{q}_{t,i} = \begin{pmatrix} x_{t,i} \\ x_{t-1,i} \\ \vdots \\ x_{t-m+1,i} \end{pmatrix}, \mathbf{C} = \begin{pmatrix} 1, & \mathbf{0}_{m-1}^T \end{pmatrix}, \mathbf{A} = \begin{pmatrix} a_i^{(1)} & a_i^{(2)} & \cdots & a_i^{(m)} \\ & & & \\ & \mathbf{I}_{m-1} & & \mathbf{0}_{m-1} \\ & & & \end{pmatrix}, \boldsymbol{\tau} = \begin{pmatrix} \epsilon_{t,i} \\ \\ \mathbf{0}_{m-1} \\ \\ \end{pmatrix} \tag{4.18}$$

Here, $\mathbf{0}_{m-1}$ is a vector of dimension $(m-1) \times 1$ with all values 0, T stands for the vector or matrix transpose, and $\mathbf{I}_{m-1}$ is the identity matrix of dimension $(m-1) \times (m-1)$. Instead of working with a

collection of one-dimensional states as in Chapter 3, the extended model with AR(m) process has a state vector of dimension m, specified as $\mathbf{q}_{t,i}$. Besides the change of the intrinsic value process, the other model assumptions remain the same as proposed in Chapter 3. Correlation among multiple data streams is assumed through the innovation $\epsilon_{t,i}$ by using the latent factor model in Eq. (3.5), which leads to the covariance in Eq. (3.6). It will be a natural extension from our proposed model and we will leave the implementation of the model with AR(m) process to future work.

4.4 Discussion

The main contribution of this chapter is extending the Bayesian dynamical model in Chapter 3 to make it a more coherent and thorough study on housing price index. We extended the model in three aspects respectively: the global trend, hedonics effects and intrinsic price dynamics. First, we incorporated a non-stationary trend into our Bayesian framework to capture the global effect, jointly with the local dynamics. The resulting model yields very similar prediction performance to the model with a pre-calculated global price trend. The joint model has slightly better prediction for those hard-to-predictor houses at the tail of the error distribution. Second, we tested the sensitivity of hedonic effects. By properly adjusting for housing hedonics, the model is able to focus on clustering the local price dynamics better. Third, we explored the possibility of modeling the intrinsic prices as a long memory process.

As future work, each of our analyses covered in this Chapter could be extended. For the global trend, we could consider a stochastic global trend instead of a deterministic trend in our joint Bayesian dynamical model, such as the ARIMA model that is commonly used for non-stationary economic time series. For the sensitivity analysis, we could also include interactions in the hedonics. Finally, beyond just implementing the long-memory version of our model, there are possible complications that might arise computationally from having a higher-dimensional latent state.

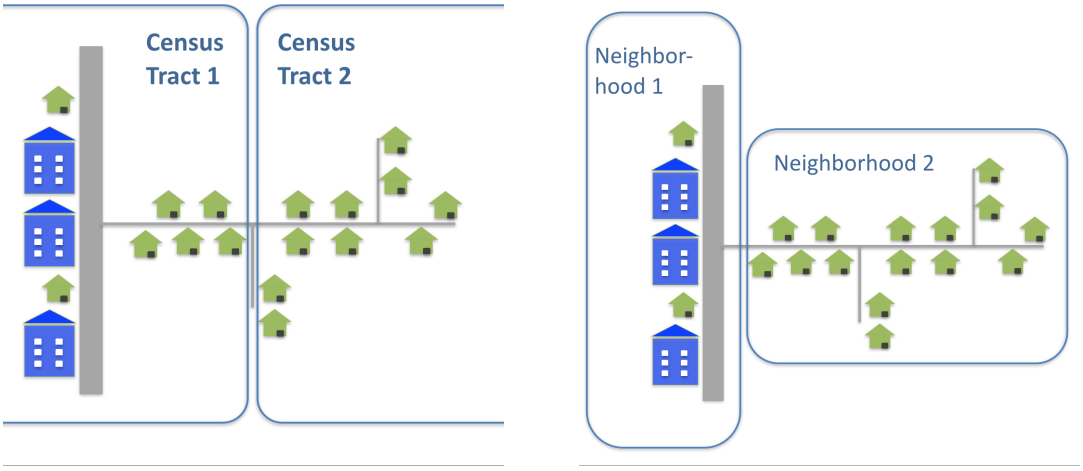
Chapter 5

DEFINING THE NEIGHBORHOOD REGIONS

We used census tracts as the finest unit of spatial granularity for the Bayesian model to compute a housing price index. The primary purpose of census tracts is to provide a permanent and stable fine scale geographic unit for statistical analysis. While census tract boundaries generally follow natural or legally based features, they are not constructed to ensure that the housing units and parcels within the tract are homogeneous. As a result, census tracts can be composed of a diverse set of housing types, from condos to townhouses to single family residences on large parcels. This is potentially problematic for the Bayesian clustering model since a census tract may exhibit different temporal price patterns due to the heterogenous nature of the housing units. For example, if a census tract is evenly split between condos and high-end single family residences, it will not be possible to classify the tract with an appropriate cluster. Figure 5.1 provides an illustrative example of regions defined by census tracts and by neighborhoods, where the arbitrary census tracts include mixed road types and heterogeneous houses and a desired neighborhood should be spatially aware of road and housing characteristics. To avoid the inherent problem in using a somewhat arbitrarily defined geographic unit, such as a census tract, we explore ways to define neighborhood regions that are homogeneous and are a naturally identifiable region.

5.1 Introduction to Neighborhood Clustering using Road Network

People typically look for specific neighborhoods while shopping for homes. Houses within a neighborhood usually have similar hedonics and are closely connected by roads. Inspired by this idea, we perform clustering to discover these neighborhoods of closely connected houses with homogeneous characteristics. Most existing methods cluster houses by their features, e.g. square footage, number of bedrooms. The challenge is to encode complex spatial proximity into the clustering



(a) (b)

Figure 5.1: An illustrative example of using pre-defined census tracts versus defining neighborhoods. The thickness of grey lines represent the traffic flow: thicker lines denote arterial roads and thin lines denote neighborhood roads. Different housing types are also shown as condo buildings (*blue*) and single-family-homes (*green*). (a) Regions defined by census tracts may arbitrarily include mixed road types and mixed housing types together, as in census tract 1. (b) Regions defined by neighborhoods should be adaptive to different road types and housing types, to form more homogeneous geographical units.

process. For example, houses separated by a short Euclidean distance might have a long driving distance, or be separated by a major highway. The road network is an essential aspect in defining a neighborhood. To get from point A to point B, you must traverse on a road. Homogeneous subdivisions are typically built around small roads with limited access points to major roads. Major highways can act as significant barriers that delineate neighborhoods. In this chapter, we improve the existing featured-based house clustering schemes by exploring the road connectivity, such that the resultant clusters or neighborhoods not only have similar house features, but are also closely connected by roads.

We solve this problem by constructing a street segment graph (see Figure 5.8), where each node represents an individual street segment, and the edge distance encodes their road connectivity. To be more specific, an edge between two nodes in the graph denotes that two street segments are connected by a road intersection. From the NavTech data for this study, a street segment typically does not cover an entire street. Empirical data shows an average street segment has 11 houses. Since street segments are so small, for the purposes of defining neighborhoods, we assume houses on the same street segment share the same neighborhood. Instead of clustering individual houses, we cluster street segments. Our goal is to search for similar street segments that are closely connected, to define a neighborhood. The "similarity" is defined by the features associated with the houses on the street segments, such as median house value (Zestimate is a good estimate of the current value of a house), percentage of single-family-houses, median bathroom count, median sqft, median lot size etc.

5.2 Literature Review

Existing methods on studying neighborhoods for housing usually use proxy spatial units, such as census tracts. Li and Brown [37] discover the effects of neighborhoods on housing prices, using census tracts as a proxy for neighborhoods. The studies in [32] and [31] gather the data source from the national version of the American Housing Survey (NAHS), where 630 houses were randomly selected to be neighborhood kernels. A kernel and its up to ten nearest neighboring houses are referred to as a neighborhood cluster. However, there are few clusters in a given region, which

limits their neighborhood model to be based on census tracts. To the best of our knowledge, this chapter describes the first work on defining the housing neighborhood structure itself.

Defining housing neighborhoods is relevant to tasks as defining communities in networks or graph. In the literatures of community detection, a community in a network is typically formed by members that are densely inter-connected whereas nodes from different communities are sparsely connected. The connection or edge usually encodes degrees of similarity between two nodes. Community detection seeks to find groups that have similarity among nodes within groups. It is worth noting that the number of communities and their sizes are usually not known and they are part of the results by the community detection algorithm. Defining community structures in networks is being studied in many areas other than housing, and with various approaches. Girvan and Newman [24] studied community structure in social and biological networks, such as a collaboration network and a food web. The main idea is using centrality indices to find community boundaries. Cut minimization method is used by [36] to assign the electronic circus to the circus boards for minimizing the connection between boards. The same approach is also used by Flake et al. [19] to studied the community structure in website links. Hierarchical methods are adopted by [24, 49] to successively divide networks into communities with a dis-similarity measure. A spectral method based on a modularity matrix is proposed in [44] to divide networks into multiple communities. Recently, a simple method based on label propagation is proposed in [50]. The algorithm initializes each node with its own label, where each label represents a community. At every iteration, each node updates its label by the label that a maximum number of neighbors have. A consensus label can be quickly propagated within a densely-connected community. This algorithm demonstrates simplicity and computing efficiency compared to other methods. The application of this algorithm to Zachary's karate club friendship network [63] and the US college football network [24] shows consistent results with the actual communities present in these datasets. Attracted by its simplicity and scalability, we follow the same idea of localized community detection by examining each node and its neighbors in developing our cost minimization algorithm for housing neighborhood clustering. A detailed algorithm is described in Section 5.3.

5.3 A Cost Minimization Algorithm for Neighborhood Clustering

In order to capture the natural characteristics of a neighborhood, we use a cost minimization approach that enforces homogeneity of street segments in terms of house characteristics within clusters, follows natural boundaries as defined by the road network and targets a specified cluster size.

Given a set of observations of n street segments $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, each observation $\mathbf{x}_i$ is a d-dimensional vector of features (median Zestimate, percentage of single-family-house, median bathroom count, median sqft, and median lot size on a street segment). The goal of the clustering is to partition these n observations into K clusters $\mathbf{C} = \{C_1, C_2, \cdots, C_K\}$, such that the value of the following cost function of the graph reaches a minimum.

$$\text{Cost}(\mathcal{G}) = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 + \sum_{k=1}^{K} f\left(|C_k|\right) + \sum_{k=1}^{K} \sum_{(s,d) \in \mathcal{G}, (s,d) \in C_k} g\left(S_s, S_d\right). \tag{5.1}$$

Here $\boldsymbol{\mu}_k$ is the center of points in set C_k, and $f()$ is the penalty function on the cluster size $|C_k|$. The cluster size $|C_k|$ for the set C_k is $|C_k| = \sum_{i \in C_k} m^{(i)}$, where $m^{(i)}$ is number of properties on street segment i. Therefore $|C_k|$ is the total number of houses for cluster k. The penalty function enforces the cluster size to be close to a user-specified size range, therefore the number of clusters K is implied by the total number of houses and the targeted size of a cluster. The penalty function $g()$ is to penalize different road types of two connected street segments s and d that are assigned in the same cluster C_k. The road network space is represented by the graph $\mathcal{G} = (V, \boldsymbol{\xi})$, where V is the node set, and $\boldsymbol{\xi}$ is the edge set. A node $s \in V$ denotes a street segment and an edge $e \in \boldsymbol{\xi}$ means an road intersection between two nodes (street segments).

Detailed steps of our cost minimization algorithm for neighborhood clustering is described in Algorithm 3.

Initialize each node to its own cluster.;

while *Cost*($\mathcal{G}$) *not converged and iteration* < 300*:* **do**

Step 1. Perform "Gather + Apply" to all nodes $c \in V$**, in parallel**;

Step 1.1. Gather

For each node c, we calculate and gather costs for c joining the cluster of its connected neighbors. That is, for each node c, we compute Cost(s, c) for $s \in$ Neighbor(c) with cost function

$$\text{Cost}(s, c) = \|\mathbf{x}_c - \boldsymbol{\mu}_{k(s)}\|^2 + f\left(|C_{k(s)}| + m^{(c)}\right) + g\left(S_s, S_c\right). \tag{5.2}$$

Let $k(s)$ be the cluster assignment of node s.

Step 1.2. Apply

Apply a cluster label to the node c that minimizes cost(s, c) for all neighbor s of c.

$$k(c) = k(s^*), \tag{5.3}$$

$$s^* = \arg\min_s \text{Cost}(s, c), \text{ for } s \in \text{Neighbor}(c) \tag{5.4}$$

Step 2. Update

Update the new cluster statistics. That is, update cluster mean $\boldsymbol{\mu}_k$ and cluster size $|C_k|$, for all k.

Step 3. Evaluate

Evaluate the overall cost of the graph $\mathcal{G}$ as in Eq. (5.1).

end

Algorithm 3: Cost minimization algorithm for neighborhood clustering.

We implement the algorithm in GraphLab Create [14] to leverage its parallel architecture to

perform efficient computation with graph data. In specific, we use the "triple apply" toolkit in Graphlab Create that applies a given function to each triple of (source node, edge, target node) in a graph. In our housing graph, we apply the cost function Cost(s, c) in Eq. (5.2) to every triple (s, c, e) where node $c \in V, s \in$ Neighbor(c), and the edge $e \in \boldsymbol{\xi}$. Graphlab Create is able to parse the nodes and edges into multiple parts and perform computation on the divided parts in parallel, on a single machine with multi-cores and shared memory. Therefore, the computational efficiency has been significantly improved on a big graph. In our data, it is able to compute on a graph of 30,547 nodes and 137,784 edges within seconds per iteration.

5.4 Neighborhood Clustering

5.4.1 Dataset

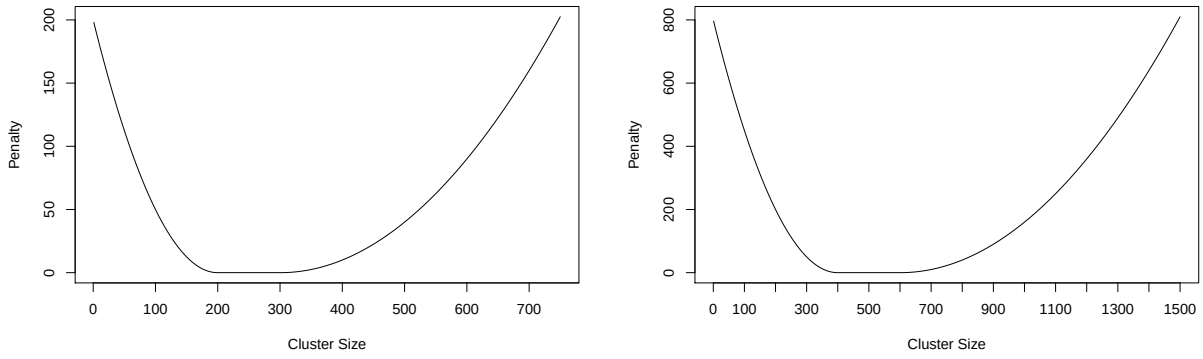
We evaluate our neighborhood clustering approach on a Seattle housing dataset. The dataset consists of snapshots of all houses in the City of Seattle in Summer 2014. There are in total 202,628 houses, spreading over 30,547 street segments. As we described in Section 5.1, we construct a graph where each street segment is a node. There is an edge between two nodes if the two corresponding street segments are connected by a street intersection. Our street segment graph has 30,547 nodes and 137,784 edges.

5.4.2 Regularizing cluster size

We regularize the cluster size with a function $f()$ as follows:

$$f(|C_k|) = 5\lambda_f \left[\min(|C_k| - C_l^*, 0)\right]^2 + \lambda_f \left[\max(|C_k| - C_u^*, 0)\right]^2 \tag{5.5}$$

where λ_f controls the weight of the penalty, and C_l^*, C_u^* are the lower and upper bounds for the penalty to be 0. The intuition of this regularization is that we want to discourage clusters with too many or too few members. $f()$ requires to specify two parameters C_l^* and C_u^* to define an appropriate neighborhood size. We have experimented with two settings, $C_l^* = 200, C_u^* = 300$ and $C_l^* = 400, C_u^* = 600$. Figure 5.2 shows $f()$ under these two settings. It has zero penalty within the



(a) (b)

Figure 5.2: Penalty function $f()$ on the cluster size. (a) $C_l^* = 200, C_u^* = 300$, (b) $C_l^* = 400, C_u^* = 600$.

desired cluster range, and the penalty gradually increases when the cluster size deviates from the desired range. In our implementation, the weight parameter λ_f is specified by users and its value reflects the emphasis on regularizing the cluster size (large λ_f) or on the feature similarity within a cluster (small λ_f). In our Seattle housing graph, we set $\lambda_f = 0.001$.

5.4.3 Road types

An important feature of our street segment graph is that we incorporate road types such as freeway, arterial and neighborhood roads. A road is characterized by its speed limit. Table 5.1 shows a complete list of speed categories form 1 to 8 and their responding speed limit ranges. Some speed categories include few roads and it is not a material difference between their speed limits, therefore we group them together to form a high-level representation as road types. In our data, the majority of street segments are in speed category 6, which corresponds to 19-31 mph. We call these *local roads*. Roads in speed category 7-8 have the speed limit less than 19 mph, which are typically *neighborhood roads* in residential areas. Roads in speed category 4-5 have speed range 32 - 56 mph, which are usually *arterial* roads. *Freeways* are in speed category 1-3, with speed limit above

Table 5.1: Speed category, corresponding speed limit range and road type.

Speed Category	km / hour	mile / hour	Road Type, S_v
1	> 130	> 81	freeway
2	101 - 130	63 - 81	freeway
3	91 - 100	57 - 62	freeway
4	71 - 90	44 - 56	arterial
5	51 - 70	32 - 43	arterial
6	31 - 50	19 - 31	local road
7	11 - 30	7 - 19	neighborhood road
8	< 11	< 7	neighborhood road

57 mph. This road type representation is based on Seattle road network data. The user can naturally define other sets of road types to represent the characteristics of roads in other cities.

We choose the road cost function $g()$ to penalize changing road type: neighborhoods should try to crawl a local neighborhood road network before crossing arterial roads, especially not freeways. Conversely, some arterial roads, which often have multifamily units and higher density housing, may naturally define a neighborhood distinct from adjacent neighborhoods that are defined by neighborhood roads.

The penalty function $g(S_a, S_b)$ on road type has the following form for any connected road segments a and b,

$$g(S_a, S_b) = \begin{cases} 0 & \text{if } S_a = S_b \\ \lambda_g & \text{if } (S_a = \text{local road}, S_b = \text{neighborhood road}) \textbf{ or} \\ & (S_a = \text{neighborhood road}, S_b = \text{local road}) \\ 5\lambda_g & \text{otherwise.} \end{cases} \tag{5.6}$$

The function $g()$ has zero penalty for two connected roads of the same road type to share a cluster. It pays some penalty λ_g to assign a local road and a neighborhood road to the same cluster. The maximum penalty $5\lambda_g$ is paid if the two roads have different road types other than one local and

one neighborhood roads. Examples for paying the maximum penalty are to assign a freeway and a local road to the same cluster, or clustering an arterial and a neighborhood road together. The weight parameter λ_g in the penalty function $g()$ is set by the user to balance the goodness of fit in terms of feature similarity within a cluster and the recognition of disparate road types. The relative values of λ_f and λ_g should also reflect the user's preference on regularizing cluster size and enforcing awareness of road types. In our implementation on Seattle housing graph, we use $\lambda_f = 0.001$ and $\lambda_g = 10$. These parameters are first set to match the magnitude of the squared in-cluster distance in Eq. (5.2), and then can be further tuned based on the clustering results obtained and the user's prior knowledge of local neighborhoods, road types and cluster sizes.

5.4.4 Cost minimization over iterations

Figure 5.3 shows the cost minimization over the iterations on our Seattle housing graph for the settings of $C_l^* = 400, C_u^* = 600$. Figure 5.4 shows the resulting clusters with the convex hull of each cluster shown as boundaries in Figure 5.5. Figure 5.6 shows a close-up view of Figure 5.5 near Sand Point Way and University District. The resultant clusters are shaped by the house features, road types and connectivity. The clustering results are interpretable and consistent with real neighborhoods. In Figure 5.6, the high-end neighborhoods of Windermere and Laurelhurst are clearly segmented from the lower Sandpoint Way neighborhood that features higher density housing. The Sandpoint Country Club neighborhood, a community built around a golf course, is well delineated from neighboring homes despite similarity in the composition of homes.

We also experimented with using $C_l^* = 200, C_u^* = 300$. Figures 5.7 - 5.9 parallel those of Figures 5.3 - 5.5 using $C_l^* = 400, C_u^* = 600$. As expected, we see smaller cluster sizes. To select which cluster size to use, we picked the resulting clusters that better match perceived notions of Seattle neighborhoods. And also the choice of cluster size depends on that how we use the clusters. In our case, for computational considerations of the Bayesian model, we use the larger cluster size to determine neighborhoods.

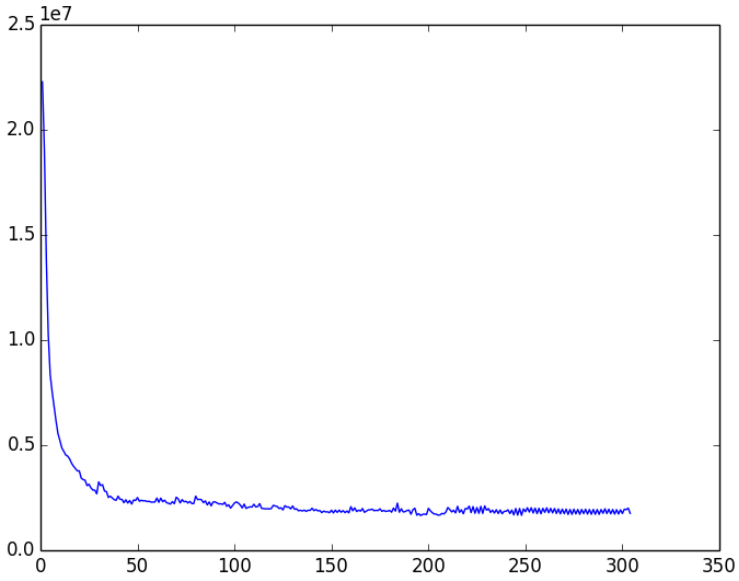


Figure 5.3: Seattle housing graph cost minimization over iterations with $C_l^* = 400, C_u^* = 600, \lambda_f = 0.001, \lambda_g = 10$.

5.5 Neighborhood Price Index

To evaluate the cluster results, we run our Bayesian nonparametric model on the discovered neighborhoods, and calculate the prediction performance for the resulting neighborhood index. In our data, Seattle has 140 census tracts whereas our cost minimization algorithm yields 259 neighborhoods (under the setting of desired cluster range $C_l^* = 400, C_u^* = 600$). Hence, our estimated neighborhoods are at a finer granularity than the census tracts. Even at this hyperlocal level, the Bayesian nonparametric model is able to provide a robust housing price index. This is demonstrated by the prediction performance summarized in Table 5.2, where we see performance exceeding that at the census tract level. By representing hyperlocal dynamics, the neighborhood index outperforms census tract index in all metrics, importantly coping with the even greater degree of sparsity at this fine level.

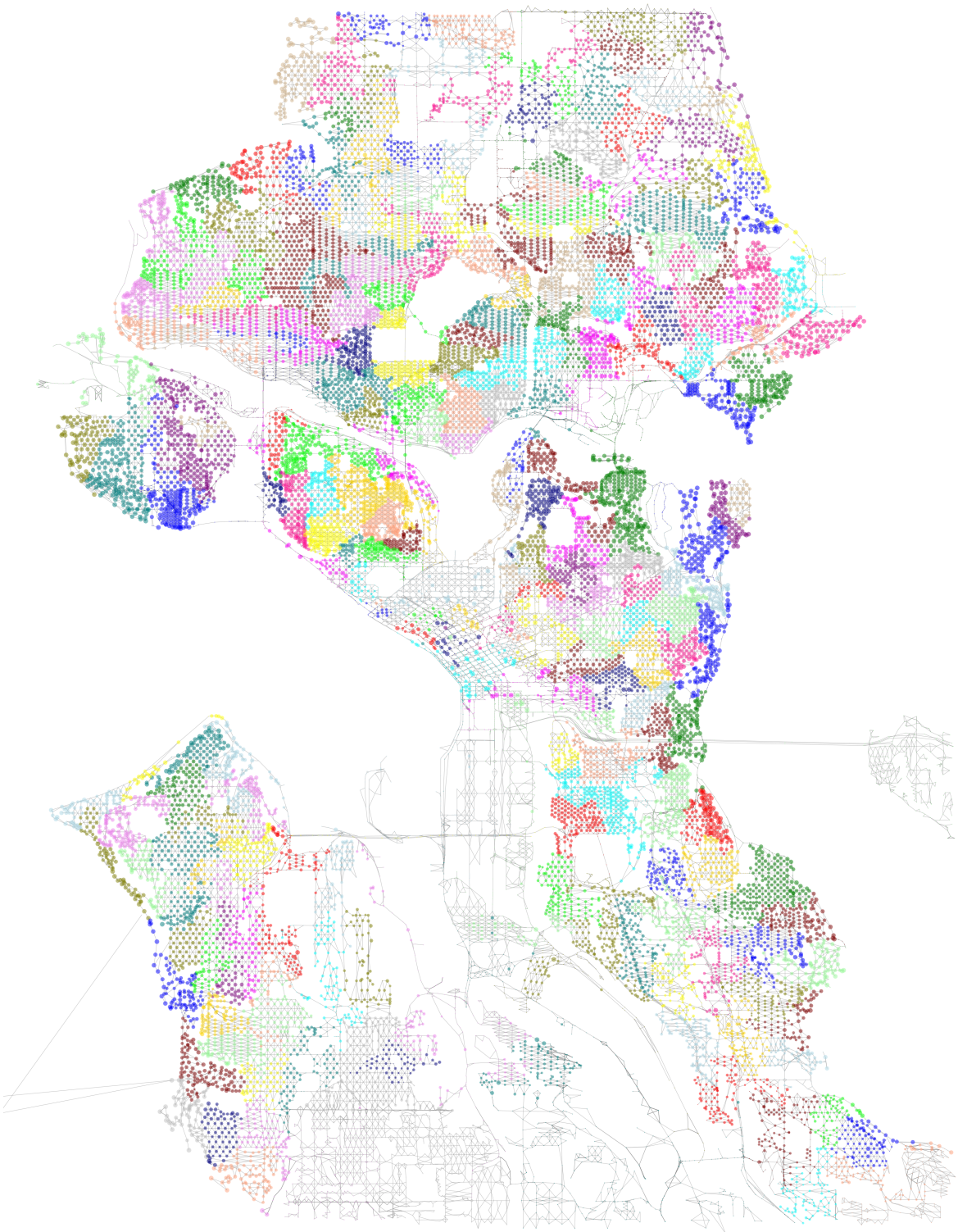


Figure 5.4: Resulting City of Seattle inferred neighborhoods (clusters), $C_l^* = 400, C_u^* = 600$.

Figure 5.5: Associated cluster convex hulls for clustering of Figure 5.4, $C_l^* = 400, C_u^* = 600$.

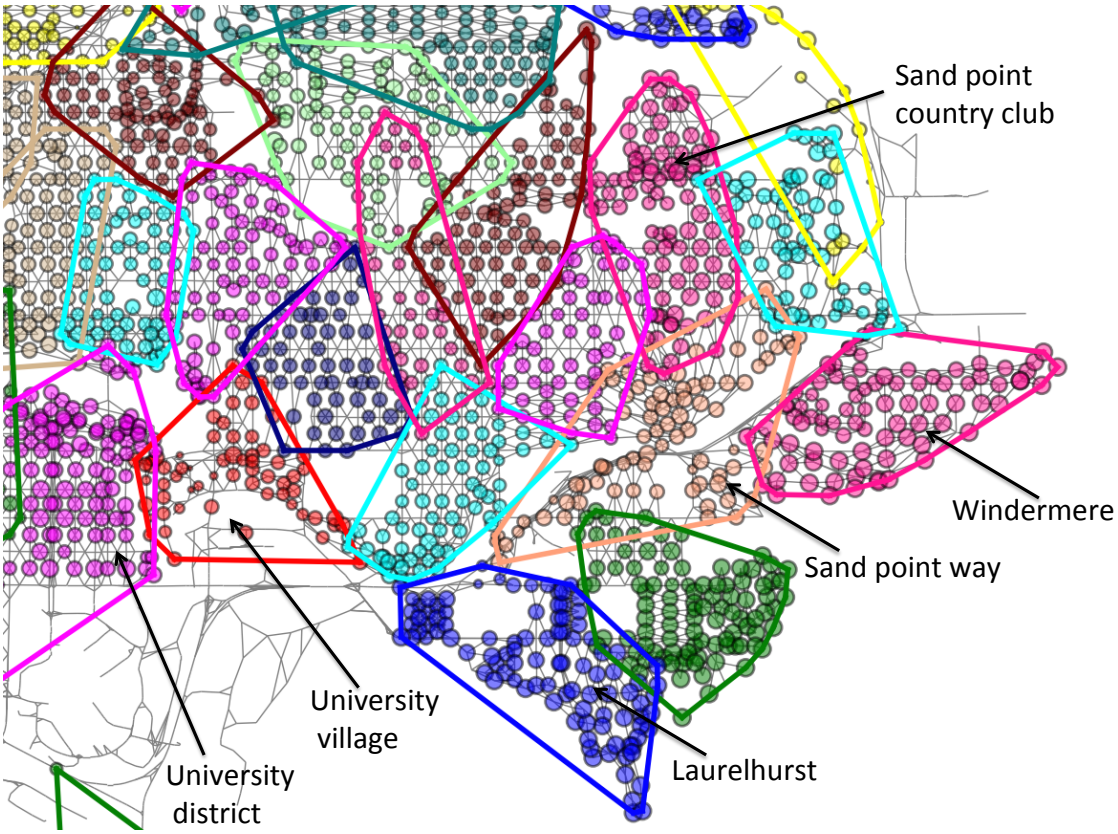


Figure 5.6: Close-up view of Figure 5.5. The discovered clusters are shaped by house features, road types and connectivity. Our algorithm shows consistent clustering with real neighborhoods near Sand point way and University District.

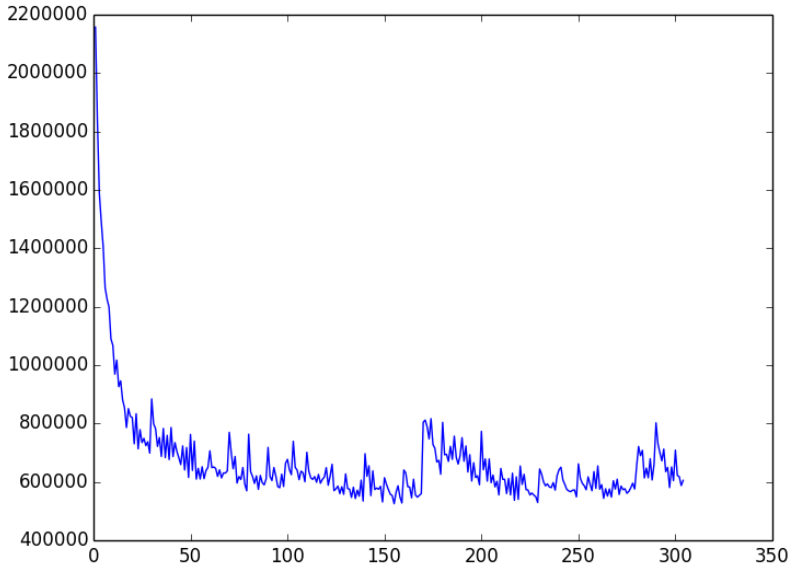


Figure 5.7: Seattle housing graph cost minimization over iterations with $C_l^* = 200, C_u^* = 300, \lambda_f = 0.001, \lambda_g = 10$.

Table 5.2: Predictive performance comparison of the Bayesian nonparametric housing index approach of Chapter 3 and 5, based on neighborhoods defined either by census tracts or the graph-based clustering method of this Chapter. Note that the graph clusters are at a finer granularity than the census tracts.

	Bayesian Census Tract Index	Bayesian 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

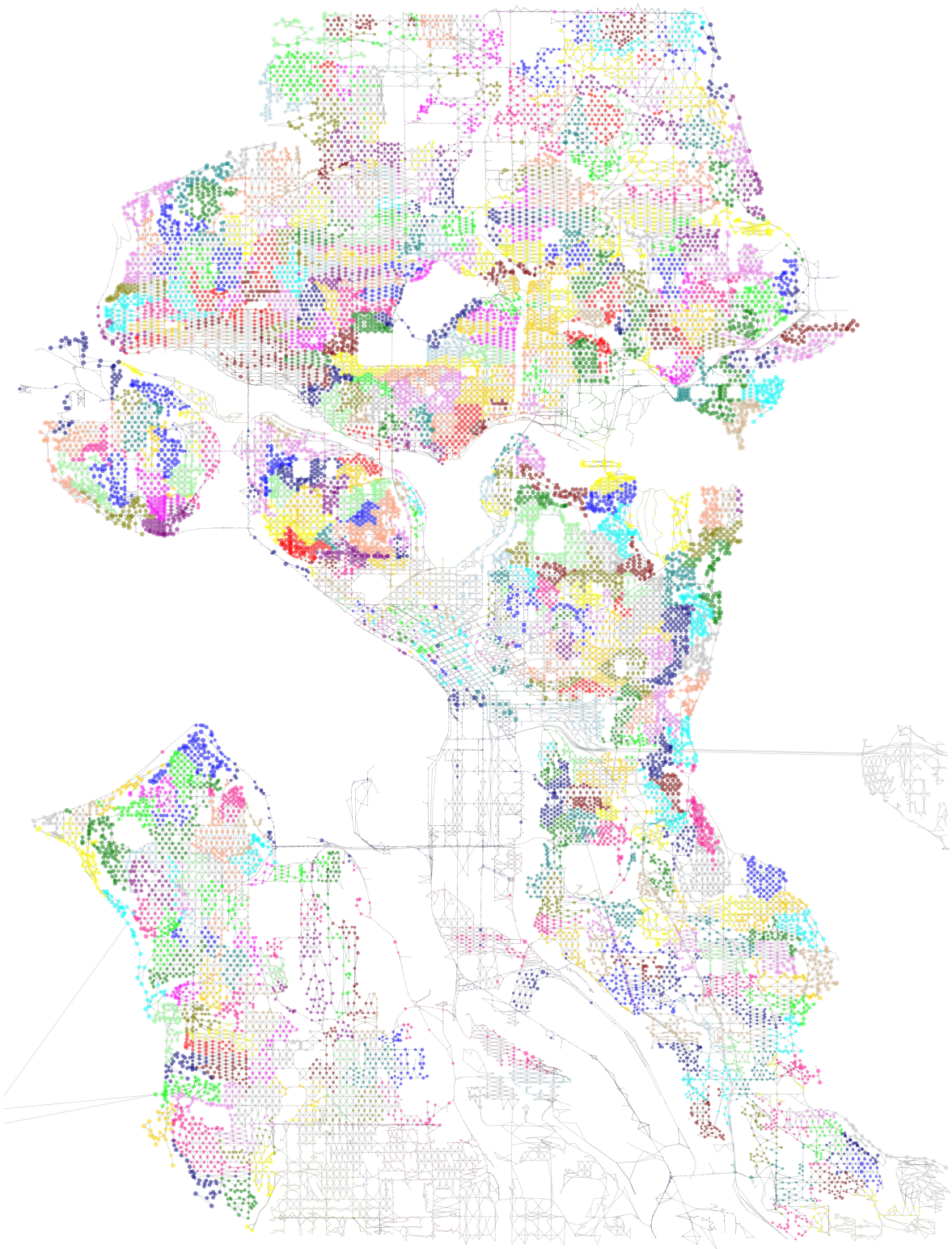


Figure 5.8: Resulting City of Seattle inferred neighborhoods (clusters), $C_l^* = 200, C_u^* = 300$.

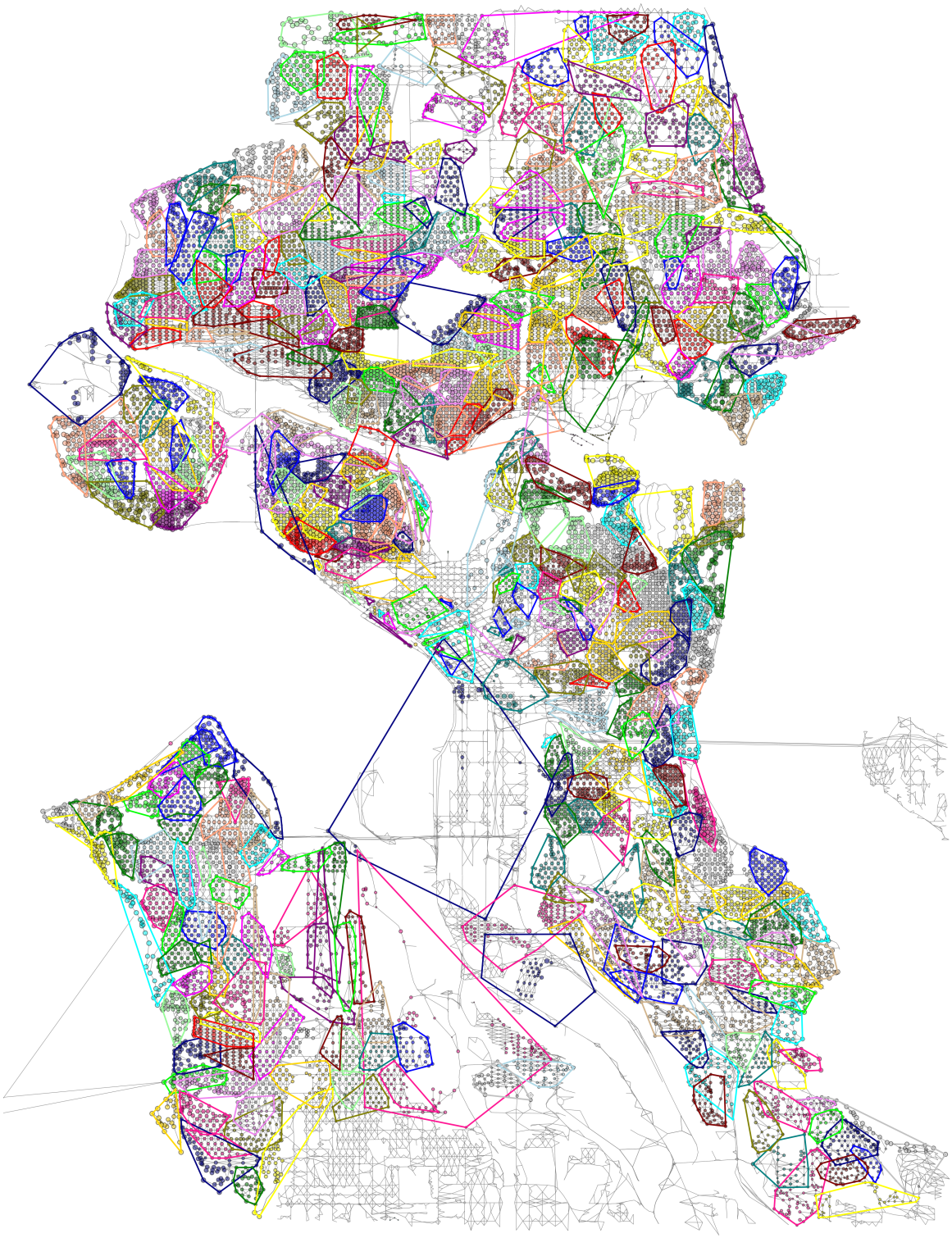


Figure 5.9: Associated cluster convex hulls for clustering of Figure 5.8, $C_l^* = 200, C_u^* = 300$.

5.6 Discussion

We study the problem of defining neighborhood structure for residential houses. To our best knowledge, this is the first attempt to learn neighborhood rather than using pre-defined jurisdiction regions, such as census tract, as a proxy. Based on the idea that neighborhoods should be spatially aware of housing and road characteristics, we utilized the road network data to guide the neighborhood search over space. In particular, our contribution includes a novel approach that encodes the spatial relationship through a graphical network where the node denotes a street segment along with housing and road features, and the edge denotes connectivity between two street segments. A cost minimization algorithm is designed to optimize for within-neighborhood housing heterogeneity, follow natural boundaries defined by road network and regularize neighborhood sizes. The discovered neighborhood structure in Seattle City shows consistent results with commonly recognized neighborhoods. We apply the Bayesian dynamical model on the discovered neighborhoods and achieve better prediction performance than the same Bayesian model computed on pre-defined census tracts. This demonstrates the neighborhoods are more homogeneous geographical units than census tracts.

Our proposed cost minimization algorithm empirically performs well on the Seattle housing graph. However, the algorithm has some limitations that warrant significant further studies. A major limitation of the algorithm is, at each iteration, the Gather + Apply step is performed on only direct neighbors. This leads to several issues: the algorithm is not guaranteed to discover a global optimum, it may not converge and the clusters are not necessarily spatially contiguous. The convergence issue is illustrated by Figures 5.3 and 5.7 since the cost function is not monotonically decreasing. Figure 5.10 shows how clusters may be separated apart. To alleviate the latter issue, in this thesis, we added a connected component algorithm every 10th iteration to re-assign the separated clusters with different cluster IDs. However, this reassignment increases the value of overall cost on the graph. By expanding the search area for the Gather + Apply step, we can alleviate these shortcomings at the expense of higher computational overhead. Another improvement to the algorithm is, instead of specifying a penalty function, we might want to learn the parameters of the

penalty functions.

It is also worthwhile to explore other approaches, particularly those used for image segmentation in computer vision. Algorithms like normalized cut [54], probabilistic aggregation [2] and affinity propagation [21] can be naturally extended to address the house clustering problem on network data. In particular, the affinity propagation method takes a similarity measure between two nodes as an input and evaluate all pairs of nodes. For our road network graph, the similarity of any two street segments can be a combination of housing feature similarity and spacial proximity. The latter can depend on edge distance between two nodes in the road graph. By working with the full matrix of similarity measures between any pair of nodes, the affinity propagation algorithm expands the node search beyond just adjacent nodes, which is the case in our cost minimization algorithm.

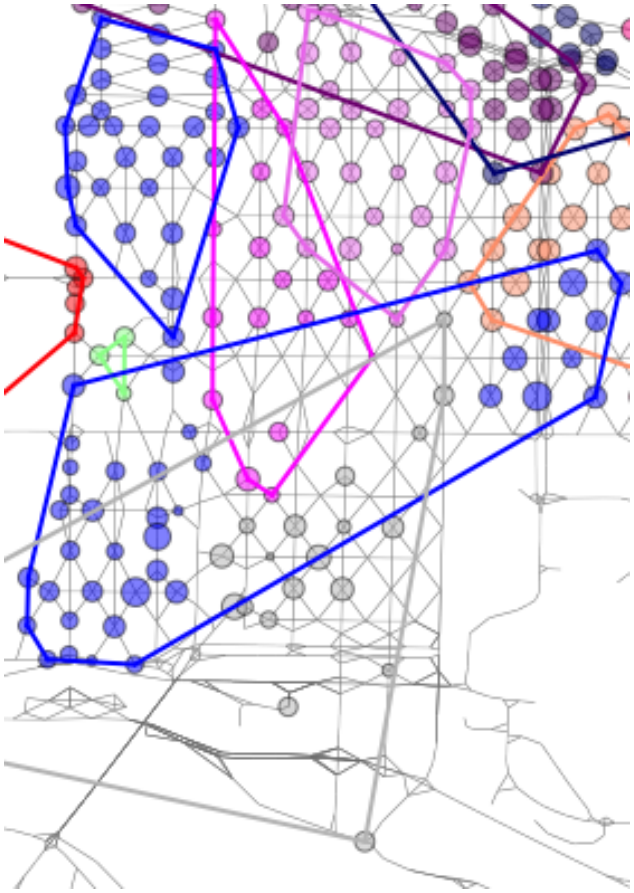


Figure 5.10: An illustration of a cluster being separated. The horizontal cluster in blue is set apart by the cluster in grey and another cluster in pink. This is caused by the nodes that were previously in blue cluster and then change to be in the grey or pink cluster. Over a few iterations, the blue cluster becomes separated by those changing nodes.

Chapter 6

CONTRIBUTIONS AND RECOMMENDATIONS

6.1 Summary of Methods and Contributions

This thesis presents a method for constructing a housing index at fine-scale geographical units, with improved space-time adjustment and specificity than existing approaches. In particular, the extreme sparsity of transactions at a fine spatio-temporal granularity poses a significant modeling challenge. This work addresses the data sparsity challenge by leveraging observations from multiple regions discovered to have correlated valuation dynamics. We propose a Bayesian nonparametric approach which builds on the framework of latent factor models to enable a flexible, data-driven method for inferring the clustering of correlated regions. This model leverages information from the region-specific time series within a cluster, providing a form of multiple shrinkage of individual trend estimates for each region. We explore methods for scalability and parallelizability of computations by parallel MCMC, yielding a housing valuation index at the level of census tract rather than zip code, and on a monthly basis rather than quarterly. An analysis is provided on a large Seattle metropolitan housing dataset, which includes all house sales record from 1997 to 2013. Our main contribution includes providing a housing index at a resolution that was un-achievable before, enabling a flexible, data-driven clustering of correlated data streams and efficient computation.

We further explore various modifications to the model and analyzing their effects, including jointly estimation of the global market trend, test of the model sensitivity to the house covariates and possible extensions to our underlying dynamical model by considering long-memory processes.

Finally this thesis seeks to study the problem of defining neighborhood structure for residential houses. To the best of our knowledge, this is the first attempt to learn neighborhoods rather than using pre-defined jurisdiction regions, such as census tract, as a proxy. Based on the idea

that neighborhoods should be spatially aware of housing and road characteristics, we utilized the road network data to guide the neighborhood search over space. In particular, our contribution includes a novel approach that encodes the spatial relationship through a graphical network. A cost minimization algorithm is designed to optimize for within-neighborhood housing heterogeneity, follow natural boundaries defined by road network and regularize neighborhood sizes. The discovered neighborhood structure in Seattle City shows consistent results with commonly recognized neighborhoods. Our discovered regions are at a finer scale than census tracts, and even in this case our Bayesian dynamical model produces a house index at this hyperlocal neighborhood level, with better predictive performance as compared to the index at the pre-defined census tract level. This demonstrates the neighborhoods are more homogeneous geographical units than census tracts.

6.2 Suggestions for Future Research

We conclude by discussing the possible future research directions to extend this work on Bayesian dynamical model for housing index and neighborhood clustering algorithm.

6.2.1 Future Work for Bayesian Housing Price Index

Vector Autoregression with Full Coefficient Matrix Various modeling approaches are possible to address the data sparsity challenge in the hyperlocal regions. This thesis assumes a diagonal autoregressive coefficient matrix and correlates regions through the innovation term of the latent process. One may consider introducing dependencies between regions through the autoregressive coefficients. In particular, the latent price dynamics of multiple regions can be modeled as a vector autoregressive process with full autoregressive coefficient matrix. In that case, the data generating process implies that the price in one region today may impact the price in another region tomorrow.

Extensions on Clustering Extensions can be made on the clustering model. In this thesis, we assumed a fixed cluster membership for each region. Interesting extensions include considering a changing cluster structure. Specifically, one region's membership may change over time periods, therefore the dependence structure among regions may evolve over time. Another possible future

work is to consider a mixed membership model where each census tract might be related to a subset of other census tracts. Furthermore, 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.

Stochastic Global Trend For the non-stationary global trend, we could consider a stochastic trend instead of a deterministic trend in our joint Bayesian dynamical model, such as the ARIMA model that is commonly used for non-stationary economic time series.

Variational Bayesian Method Beyond just implementing the long-memory version of our model, there are possible complications that might arise computationally from having a higher-dimensional latent state. One future research direction is to use variational Bayesian method to ease the computation [6, 29]. Instead of MCMC sampling, the variational Bayesian approach converts inference problems into optimization problems and yields a deterministic algorithm.

6.2.2 Future Work for Neighborhood Clustering

Our proposed cost minimization algorithm empirically performs well on the Seattle housing graph. However, the algorithm has some limitations that warrant significant further studies.

Convergence Our proposed algorithm falls into local optimum, since it is looking at each edge and is lowering the cost locally. The algorithm also has some convergence issues. As pointed out by Figure 5.3 and 5.7, the cost function is not monotonically decreasing. An important future research direction is to study the convergence and global optimality for this graph-based algorithm.

Learn Penalty Functions Instead of specifying a penalty function in the global cost objective as in Section 5.4, a possible future work is to learn the parameters of the penalty functions. One possible way to learn the penalization parameter is by cross validation.

Image Segmentation and Affinity Propagation Image segmentation in computer vision is a relevant area of research to the clustering problem on graph. Algorithms like normalized cut [54],

probabilistic aggregation [2] and affinity propagation [21] can be naturally extended to address the neighborhood clustering problem on network data. In particular, the affinity propagation method takes a similarity measure between two nodes as an input and evaluate all pairs of nodes. For our road network graph, the similarity of any two street segments can be a combination of housing feature similarity and spacial proximity. The latter can depend on edge distance between two nodes in the road graph. By working with the full matrix of similarity measures between any pair of nodes, the affinity propagation algorithm expands the node search beyond just adjacent nodes.

Appendix A

CONDITIONAL LIKELIHOOD OF DATA IN CLUSTER K

In this section, we provide the derivation for EM algorithm to estimate the parameters in a state space model. The unknown parameters in a state space model include $\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}$ for the doubly process. The joint log-likelihood of both $\mathbf{x}$ and $\mathbf{y}$ is

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \log p(\mathbf{x}, \mathbf{y}|\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}) && \text{(A.1)} \\ &= \log \left[\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{y}_t|\mathbf{x}_t) \right] && \text{(A.2)} \\ &= \sum_{t=1}^{T} \log p(\mathbf{x}_t|\mathbf{x}_{t-1}) + \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{x}_t). && \text{(A.3)} \end{aligned}$$

By substituting $\mathbf{x}_t|\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q})$ and $\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R})$, we get the explicit form of the joint log-likelihood as

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \sum_{t=0}^{T-1} \left(\frac{1}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)^T \mathbf{Q}^{-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t) \right) + && \text{(A.4)} \\ & \quad \sum_{t=0}^{T} \left(\frac{1}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t)^T \mathbf{R}^{-1} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t) \right) + \text{constant}. \end{aligned}$$

By utilizing the equation $a = \text{Tr}(a)$ for any scalar a, we get

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \frac{T}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} \left(\sum_{t=0}^{T-1} \text{Tr} \left[(\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)^T \mathbf{Q}^{-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t) \right] \right) + && \text{(A.5)} \\ & \quad \frac{T}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} \left(\sum_{t=0}^{T} \text{Tr} \left[(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t)^T \mathbf{R}^{-1} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t) \right] \right) + \text{constant}. \end{aligned}$$

Following the matrix equations $\text{Tr}(AB) = \text{Tr}(BA)$ and $\text{Tr}(A) + \text{Tr}(B) = \text{Tr}(A + B)$, Eqn. (A.5) becomes

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \frac{T}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{Q}^{-1} \sum_{t=0}^{T-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)^T \right) + \\ & \frac{T}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{R}^{-1} \sum_{t=0}^{T} (\mathbf{y}_t - \mathbf{C}\mathbf{x}_t)(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t)^T \right) + \text{constant.} \end{aligned} \tag{A.6}$$

Finally, we expand the summand and obtain the following,

$$\begin{aligned} l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) &= \frac{T}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{Q}^{-1} \sum_{t=0}^{T-1} (\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T - \mathbf{x}_{t+1}\mathbf{x}_t^T\mathbf{A}^T - \mathbf{A}\mathbf{x}_t\mathbf{x}_{t+1}^T + \mathbf{A}\mathbf{x}_t\mathbf{x}_t^T\mathbf{A}^T) \right) + \\ & \frac{T}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{R}^{-1} \sum_{t=0}^{T} (\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t\mathbf{x}_t^T\mathbf{C}^T - \mathbf{C}\mathbf{x}_t\mathbf{y}_t^T + \mathbf{C}\mathbf{x}_t\mathbf{x}_t^T\mathbf{C}^T) \right) + \text{constant.} \end{aligned} \tag{A.7}$$

Since the hidden states are unknown, we marginalize out the latent states by taking the expectation of the joint log-likelihood as follows, known as the E-step in EM algorithm:

$$\begin{aligned} & E_{\mathbf{x}|\mathbf{y}} \left[l(\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}|\mathbf{x}, \mathbf{y}) \right] \\ = \; & \frac{T}{2} \log |\mathbf{Q}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{Q}^{-1} \sum_{t=0}^{T-1} \left[E(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T) - E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{x}_t\mathbf{x}_{t+1}^T) + \mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{A}^T \right] \right) + \\ & \frac{T}{2} \log |\mathbf{R}^{-1}| - \frac{1}{2} \text{Tr} \left(\mathbf{R}^{-1} \sum_{t=0}^{T} \left[\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t E(\mathbf{x}_t^T)\mathbf{C}^T - \mathbf{C}E(\mathbf{x}_t)\mathbf{y}_t^T + \mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{C}^T \right] \right) + \text{constant.} \end{aligned} \tag{A.8}$$

The maximum likelihood estimator for the parameters can be computed by solving the follow-

ing first derivative equations with respect to each parameter, known as the M-step in EM algorithm:

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial A} = \frac{1}{2}\mathbf{Q}^{-1}\left(\sum_{t=0}^{T-1}\left[2E(\mathbf{x}_{t+1}\mathbf{x}_t^T) - 2\mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\right]\right) = 0 \tag{A.9}$$

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial C} = \frac{1}{2}\mathbf{R}^{-1}\left(\sum_{t=0}^{T}\left[2\mathbf{y}_t E(\mathbf{x}_t^T) - 2\mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\right]\right) = 0 \tag{A.10}$$

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial Q^{-1}} \tag{A.11}$$

$$= \frac{T}{2}\mathbf{Q} - \frac{1}{2}\left(\sum_{t=0}^{T-1}\left[E(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T) - E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{x}_t\mathbf{x}_{t+1}^T) + \mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{A}^T\right]\right)^T = 0$$

$$\frac{\partial E_{\mathbf{x}|\mathbf{y}}\left[l(\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}|\mathbf{x},\mathbf{y})\right]}{\partial R^{-1}} \tag{A.12}$$

$$= \frac{T+1}{2}\mathbf{R} - \frac{1}{2}\left(\sum_{t=0}^{T}\left[\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t E(\mathbf{x}_t^T)\mathbf{C}^T - \mathbf{C}E(\mathbf{x}_t)\mathbf{y}_t^T + \mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{C}^T\right]\right)^T = 0. \tag{A.13}$$

Therefore, the estimators for maximizing the expected log-likelihood are

$$\hat{\mathbf{A}} = \left(\sum_{t=0}^{T-1}E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\right)\left(\sum_{t=0}^{T-1}E(\mathbf{x}_t\mathbf{x}_t^T)\right)^{-1} \tag{A.14}$$

$$\hat{\mathbf{C}} = \left(\sum_{t=0}^{T}\mathbf{y}_t E(\mathbf{x}_t^T)\right)\left(\sum_{t=0}^{T}E(\mathbf{x}_t\mathbf{x}_t^T)\right)^{-1} \tag{A.15}$$

$$\hat{\mathbf{Q}} = \frac{1}{T}\left(\sum_{t=0}^{T-1}\left[E(\mathbf{x}_{t+1}\mathbf{x}_{t+1}^T) - E(\mathbf{x}_{t+1}\mathbf{x}_t^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{x}_t\mathbf{x}_{t+1}^T) + \mathbf{A}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{A}^T\right]\right) \tag{A.16}$$

$$\hat{\mathbf{R}} = \frac{1}{T+1}\left(\sum_{t=0}^{T}\left[\mathbf{y}_t\mathbf{y}_t^T - \mathbf{y}_t E(\mathbf{x}_t^T)\mathbf{C}^T - \mathbf{C}E(\mathbf{x}_t)\mathbf{y}_t^T + \mathbf{C}E(\mathbf{x}_t\mathbf{x}_t^T)\mathbf{C}^T\right]\right). \tag{A.17}$$

Appendix B

CONDITIONAL LIKELIHOOD OF DATA IN CLUSTER K

In this section, we describe how to compute the likelihood of the data from all time series assigned to a given cluster k, conditioned on the model parameters. We consider two mathematically equivalent methods: one based on the collection of observations directly, and the other using sufficient statistics of the observed house sales. In what follows, we drop the cluster index k for simplicity of notation.

B.1 Naive Kalman filtering

We consider a straightforward extension of the standard Kalman filter recursions to compute the marginal likelihood of all observations in cluster k when there can be multiple observations per time step. The derivation is as follows. The cluster marginal likelihood can be calculated as

$$\log P(\mathbf{y}_{1:T}) = \sum_{t=1}^{T} \log P(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \tag{B.1}$$

where the distribution of new observations at time t conditional on past time series is

$$\mathbf{y}_t|\mathbf{y}_{1:t-1} \sim \mathcal{N}\left(\mathbf{y}_t \left| C_t\mu_{t|t-1} + D_tU_t, S_t\right.\right). \tag{B.2}$$

The quantities $\mu_{t|t-1}$ and S_t are obtained by the Kalman filter:

$$\begin{aligned}
\text{Predict} \quad \mu_{t|t-1} &= A\mu_{t-1|t-1} \\
V_{t|t-1} &= AV_{t-1|t-1}A^T + Q \\
\text{Calculate} \quad S_t &= C_tV_{t|t-1}C_t^T + R_t \\
\text{Kalman gain matrix} \quad K_t &= V_{t|t-1}C_t^TS_t^{-1} \\
\text{Filter} \quad \mu_{t|t} &= \mu_{t|t-1} + K_t\left(\mathbf{y}_t - C_t\mu_{t|t-1} - D_tU_t\right) \\
V_{t|t} &= \left(\mathbf{I} - K_tC_t\right)V_{t|t-1}
\end{aligned} \tag{B.3}$$

The coefficient matrix C_t is an indicator matrix mapping each observation to its specific census tract. The matrix D_t is a coefficient matrix for hedonic effects. The filter should be applied to data for all tracts in cluster k together.

The purpose of doing filtering here is to evaluate the conditional likelihood of tract i belonging to cluster k, given observations of all the other tracts in cluster k. The conditional likelihood is

$$\begin{aligned} P(\mathbf{y}_{1:T,i}|\mathbf{y}_{1:T,-i}) &= P(\mathbf{y}_{1,i}|\mathbf{y}_{1,-i})P(\mathbf{y}_{2,i}|\mathbf{y}_1, \mathbf{y}_{2,-i}) \\ & \quad P(\mathbf{y}_{3,i}|\mathbf{y}_{1:2}, \mathbf{y}_{3,-i}) \cdots P(\mathbf{y}_{T,i}|\mathbf{y}_{1:T-1}, \mathbf{y}_{T,-i}). \end{aligned} \tag{B.4}$$

Therefore the log-likelihood of observations for tract i conditional on the other observations in cluster k is

$$\log P(\mathbf{y}_{1:T,i}|\mathbf{y}_{1:T,-i}) = \sum_{t=1}^{T} \log P(\mathbf{y}_{t,i}|\mathbf{y}_{1:t-1}, \mathbf{y}_{t,-i}). \tag{B.5}$$

At time t, we have the joint distribution $\mathbf{y}_{t,i}, \mathbf{y}_{t,-i}|\mathbf{y}_{1:t-1}$, which is $\mathbf{y}_t|\mathbf{y}_{1:t-1}$ in Eq. (B.2). We can then derive the conditional distribution $\mathbf{y}_{t,i}|\mathbf{y}_{1:t-1}, \mathbf{y}_{t,-i}$ by the conventional conditional multivariate normal distribution as follows:

$$A|B \sim \mathcal{N}\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right) \tag{B.6}$$

for the general form of a joint multivariate normal distribution

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA}, \Sigma_{AB} \\ \Sigma_{BA}, \Sigma_{BB} \end{pmatrix}\right]. \tag{B.7}$$

B.2 Sufficient statistic Kalman filter

If all $p^{(k)}$ tracts in a particular cluster k have observations at time t, the sufficient statistic multivariate Kalman filter algorithm is as follows:

$$\begin{aligned}
\text{Predict} \quad \mu_{t|t-1} &= A\mu_{t-1|t-1} \\
V_{t|t-1} &= AV_{t-1|t-1}A^T + Q \\
\text{Calculate} \quad S_t &= V_{t|t-1} + \bar{R}_t \\
\text{(Kalman gain matrix)} \quad K_t &= V_{t|t-1}S_t^{-1} \\
\text{Filter} \quad \mu_{t|t} &= \mu_{t|t-1} + K_t(\bar{\dot{\mathbf{y}}}_\mathbf{t} - \mu_{\mathbf{t}|\mathbf{t-1}}) \\
V_{t|t} &= (\mathbf{I} - K_t)\, V_{t|t-1}
\end{aligned} \tag{B.8}$$

where $\dot{\mathbf{y}}_t$ denotes the vector of observations with hedonic effects removed and $\bar{\dot{\mathbf{y}}}_\mathbf{t}$ the tract-specific mean of all observations at time t after removing hedonic effects. The matrix $\bar{R}_t$ is the diagonal matrix of size $p^{(k)}$-by-$p^{(k)}$ with (i,i)-th entry being $\sigma_i^2/L_{t,i}$. The variable σ_i^2 is the observational variance for tract i and the variable $L_{t,i}$ is the number of observations in tract i at time t. Note that all matrix operations above are of the size of the cluster, $p^{(k)}$. If some tracts at time t have no transactions, i.e. $L_{t,i} = 0$, we use the following recursion instead:

$$\begin{aligned}
\text{Predict} \quad \mu_{t|t-1} &= A\mu_{t-1|t-1} \\
V_{t|t-1} &= AV_{t-1|t-1}A^T + Q \\
\text{Calculate} \quad S_t &= \bar{C}_t V_{t|t-1}\bar{C}_t^T + \bar{R}_t \\
\text{(Kalman gain matrix)} \quad K_t &= V_{t|t-1}\bar{C}_t^T S_t^{-1} \\
\text{Filter} \quad \mu_{t|t} &\qquad\qquad \mu_{t|t-1} + K_t(\bar{\dot{\mathbf{y}}}_\mathbf{t} - \mu_{\mathbf{t}|\mathbf{t-1}}) \\
V_{t|t} &= \left(\mathbf{I} - K_t\bar{C}_t\right) V_{t|t-1}
\end{aligned} \tag{B.9}$$

In the formula above, $\bar{C}_t$ is an indicator matrix of non-zero sales at time t. The matrix has size $p^{(k)\prime} \times p^{(k)}$, where $p^{(k)\prime}$ is the number of tracts that have observations at time t (therefore $p^{(k)\prime} \leq p^{(k)}$). The response variance matrix $\bar{R}_t$ is of size $p^{(k)\prime} \times p^{(k)\prime}$ and includes the variance for tracts that have observations.

Appendix C

DERIVATION OF SAMPLING STEPS

In this section, we provide further details and derivations of the sampling steps outlined in Section 3.5 of the main paper.

C.1 Forward filter backward sampler for the intrinsic price dynamics

To sample the latent state sequence, we run a forward filter backward sampler.

Forward Kalman Filter

1. Initialize filter with $\mu_{0|0}, V_{0|0}$, where $X_0 \sim N(\mu_{0|0}, V_{0|0})$

2. Working forward in time, for $t = 1, \cdots, T$, implement the sufficient statistic filter of Appendix B.2 to obtain $\mu_{t|t}, V_{t|t}$ for $t = 1, \cdots, T$, where $X_t|\mathbf{y}_{1:t} \sim N(\mu_{t|t}, V_{t|t})$.

Backward Sampler

1. Draw X_T from $P(X_T|\mathbf{y}_{1:T}) = N(\mu_{T|T}, V_{T|T})$.

2. Sequentially sample backward, for $t = T-1, \cdots, 0$, x_t from $P(X_t|x_{t+1}, \mathbf{y}_{1:t})$:

$$x_t \sim N\left[\mu_{t|t} + J_t(x_{t+1} - \mu_{t+1|t}), \quad V_{t|t} - J_t V_{t+1|t} J_t^T\right] \tag{C.1}$$

where $J_t = V_{t|t} A^T V_{t+1|t}^{-1}$.

C.2 Sampling the latent factor η^*

For any t, the vector of latent states for all p tracts jointly follows a vector autoregressive (VAR) process as follows:

$$\begin{bmatrix} x_{t,1} \\ \vdots \\ x_{t,n} \end{bmatrix} = \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{pmatrix} \begin{bmatrix} x_{t-1,1} \\ \vdots \\ x_{t-1,n} \end{bmatrix} + (\Lambda \cdot Z) \begin{bmatrix} \eta^*_{t,1} \\ \vdots \\ \eta^*_{t,K} \end{bmatrix} + \tilde{\boldsymbol{\epsilon}}_t. \tag{C.2}$$

The VAR process can be written in the form of vectors and matrices:

$$\mathbf{x_t} = A\mathbf{x_{t-1}} + (\Lambda \cdot Z)\boldsymbol{\eta}^*{}_t + \tilde{\boldsymbol{\epsilon}}_t. \tag{C.3}$$

such that

$$\mathbf{x_t} - A\mathbf{x_{t-1}} \sim \mathcal{N}_n \left[(\Lambda \cdot Z)\boldsymbol{\eta}^*{}_t,\ \sigma_0^2 I_n \right]. \tag{C.4}$$

By first multiplying $(\Lambda \cdot Z)^T$, we get

$$(\Lambda \cdot Z)^T(\mathbf{x_t} - A\mathbf{x_{t-1}}) \ \sim \ \mathcal{N}_K \left[(\Lambda \cdot Z)^T(\Lambda \cdot Z)\boldsymbol{\eta}^*{}_t,\ \sigma_0^2 (\Lambda \cdot Z)^T(\Lambda \cdot Z) \right]. \tag{C.5}$$

We then multiply by $\left[(\Lambda \cdot Z)^T(\Lambda \cdot Z)\right]^{-1}$ and obtain

$$\left[(\Lambda \cdot Z)^T(\Lambda \cdot Z)\right]^{-1} (\Lambda \cdot Z)^T(\mathbf{x_t} - A\mathbf{x_{t-1}}) \sim \mathcal{N}_K \left\{ \boldsymbol{\eta}^*{}_t,\ \sigma_0^2 \left[(\Lambda \cdot Z)^T(\Lambda \cdot Z)\right]^{-1} \right\} \tag{C.6}$$

Given the prior of $\boldsymbol{\eta}^*{}_t \sim \mathcal{N}_K(\mathbf{0}, I_n)$ and the likelihood in Eq. (C.6) , by conjugacy, the full conditional distribution for $\boldsymbol{\eta}^*{}_t$ is

$$\boldsymbol{\eta}^*{}_t \left| \boldsymbol{\lambda}, \mathbf{z}, \mathbf{x}, \sigma_0^2 \right. \sim \mathcal{N}_K \left\{ \begin{array}{c} V \frac{1}{\sigma_0^2} (\Lambda \cdot Z)^T(\mathbf{x_t} - A\mathbf{x_{t-1}}), \\ V = \left[I_K + \frac{1}{\sigma_0^2} (\Lambda \cdot Z)^T(\Lambda \cdot Z) \right]^{-1} \end{array} \right\} \tag{C.7}$$

C.3 Sampling the factor loadings λ

For any t,

$$x_{t,i} = a_i x_{t-1,i} + \lambda_{ik} Z_{ik} \eta^*_{t,k} + \tilde{\epsilon}_{t,i} \tag{C.8}$$

If $Z_{ik} = 0$, then the full conditional distribution for λ_{ik} is just its prior,

$$\lambda_{ik}|\mathbf{x}, \mathbf{a}, \boldsymbol{\eta}^*, \mathbf{z}, \sigma_0^2 \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2) \tag{C.9}$$

If $Z_{ik} = 1$ then

$$\begin{aligned} & p(\lambda_{ik}|\mathbf{x}, \mathbf{a}, \boldsymbol{\eta}^*, \mathbf{z}, \sigma_0^2) \\ \propto \;& \mathcal{N}(\mu_\lambda, \sigma_\lambda^2) \prod_{t=1}^{T} \mathcal{N}(x_{t,i}|a_i x_{t-1,i} + \lambda_{ik}\eta^*_{t,k}, \sigma_0^2) \\ \propto \;& \mathcal{N}(\mu_\lambda, \sigma_\lambda^2) \prod_{t=1}^{T} \mathcal{N}\left(\frac{x_{t,i} - a_i x_{t-1,i}}{\eta^*_{t,k}} \middle| \lambda_{ik}, \frac{\sigma_0^2}{\eta^{*2}_{t,k}} \right) \\ \propto \;& \mathcal{N}\left[v\left(\frac{\mu_\lambda}{\sigma_\lambda^2} + \sum_{t=1}^{T} \frac{(x_{t,i} - a_i x_{t-1,i})/\eta^*_{t,k}}{\sigma_0^2/\eta^{*2}_{t,k}} \right), v = \left(\frac{1}{\sigma_\lambda^2} + \frac{1}{\sigma_0^2} \sum_{t=1}^{T} \eta^{*2}_{t,k} \right)^{-1} \right] \end{aligned} \tag{C.10}$$

In summary, the full conditional distribution for λ_{ik} is

$$\lambda_{ik}|\mathbf{x}, \mathbf{a}, \boldsymbol{\eta}^*, \mathbf{z}, \sigma_0^2 \sim \mathcal{N}(\mu^*_{ik}, v^*_{ik}) \tag{C.11}$$

where

$$(\mu^*_{ik}, v^*_{ik}) = \begin{cases} \mu_\lambda, \sigma_\lambda^2 & \text{if } Z_{ik} = 0 \\ v\left(\mu_\lambda \frac{1}{\sigma_\lambda^2} + \frac{1}{\sigma_0^2} \sum_{t=1}^{T} \epsilon_{t,i}\eta^*_{t,k} \right), v = \left(\frac{1}{\sigma_\lambda^2} + \frac{1}{\sigma_0^2} \sum_{t=1}^{T} \eta^{*2}_{t,k} \right)^{-1} & \text{if } Z_{ik} = 1 \end{cases}.$$

Here $\epsilon_{t,i} = x_{t,i} - a_i x_{t-1,i}$ and $\sum_{t=1}^{T} \epsilon_{t,i}\eta^*_{t,k}$ can be written as the inner product $\boldsymbol{\epsilon}_i^T \boldsymbol{\eta}^*_k$.

C.4 Sampling the autoregressive process parameters a_i

By the likelihood in Eq. (3.3) and Eq. (3.5) of the main paper, for $z_i = k$

$$x_{t,i} = a_i x_{t-1,i} + \lambda_{ik}\eta^*_{t,k} + \tilde{\epsilon}_{t,i}, \quad \tilde{\epsilon}_{t,i} \sim \mathcal{N}(0, \sigma_0^2). \tag{C.12}$$

Therefore,

$$x_{t,i} \sim \mathcal{N}(a_i x_{t-1,i} + \lambda_{ik}\eta^*_{t,k},\ \sigma_0^2). \tag{C.13}$$

By rearranging the terms, we get

$$\frac{x_{t,i} - \lambda_{ik}\eta^*_{t,k}}{x_{t-1,i}} \sim \mathcal{N}\left(a_i,\ \frac{\sigma_0^2}{x^2_{t-1,i}} \right), \quad \text{i.i.d. for } t = 1, \cdots, T. \tag{C.14}$$

By conjugacy, the posterior distribution of the AR process coefficient a_i is

C.5 Sampling the covariate parameters $\beta_{i,h}$

For tract i and hedonic covariate h, the posterior distribution for covariate effect $\beta_{i,h}$ is

$$\begin{aligned}
& p(\beta_{i,h}|\mu_h, \sigma_h^2, R_i, \mathbf{x}_{1:T,i}, \mathbf{y}_{1:T,i}) \\
\propto\ & N(\beta_{i,h}|\mu_h, \sigma_h^2) \prod_{t=1}^{T} \prod_{l=1}^{L_t} N\left(y_{t,i,l} \left| x_{t,i} + \sum_{s \neq h} \beta_s U_{l,s} + \beta_h U_{l,h} \right., R_i \right) \\
\propto\ & N(\beta_{i,h}|\mu_h, \sigma_h^2) \prod_{t=1}^{T} \prod_{l=1}^{L_t} N\left(y_{t,i,l} - x_{t,i} + \sum_{s \neq h} \beta_s U_{l,s} \middle| \beta_h U_{l,h}, R_i \right) \\
\propto\ & N(\beta_{i,h}|\mu_h, \sigma_h^2) \prod_{t=1}^{T} \prod_{l=1}^{L_t} N\left[\frac{1}{U_{l,h}} \left(y_{t,i,l} - x_{t,i} + \sum_{s \neq h} \beta_s U_{l,s} \right) \middle| \beta_h, \frac{R_i}{U_{l,h}^2} \right] \\
\propto\ & N \left\{ \begin{array}{c} v \left[\frac{1}{\sigma_h^2} \mu_h + \frac{1}{R_i} \sum_{t=1}^{T} \sum_{l=1}^{L_t} U_{l,h} \left(y_{t,i,l} - x_{t,i} - \sum_{s \neq h} \beta_s U_{l,s} \right) \right], \\ v = \left(\frac{1}{\sigma_h^2} + \frac{1}{R_i} \sum_{t=1}^{T} \sum_{l=1}^{L_t} U_{l,h}^2 \right)^{-1} \end{array} \right\}
\end{aligned}$$

C.6 Sampling the DP hyperparameter α

Following [17] and [1], we assume a gamma distribution prior for the concentration parameter $\alpha \sim \text{Gamma}(\alpha_\alpha, \beta_\alpha)$. We sample an auxiliary variable κ to help us sample α:

1. Sample $\kappa \sim \text{Beta}(\alpha + 1, n)$, where n is the total number of tracts.

2. Sample α from the a mixture of two gamma distributions as follows:

$$\begin{aligned}
\alpha|\kappa, K \ \sim\ & \pi \text{Gamma}\left(\alpha_\alpha + K, \beta_\alpha - \log(\kappa)\right) \\
& + (1 - \pi) \text{Gamma}\left(\alpha_\alpha + K - 1, \beta_\alpha - \log(\kappa)\right),
\end{aligned}$$

   where K is the number of unique clusters, and the mixture weight π is defined by $\pi/(1-\pi) = (\alpha_\alpha + K - 1)/\left(p\left[\beta_\alpha - \log(\kappa)\right]\right)$.

Appendix D

PARALLEL DPMM SAMPLER

Sampling the cluster membership z_i in parallel includes the following two steps:

1. **Local step** on each machine in parallel:

   Conditioned on the processor assignments $\boldsymbol{\gamma}$, we sample the cluster assignments $\{z_i : \gamma_i = j\}$ as in a conventional Dirichlet process mixture model (Section 1 of the main paper) with concentration parameter α/P, for data points assigned to a machine j. Since the DPMMs are independent given the processor allocations, we can sample $\{z_i : \gamma_i = j\}$ in parallel across machines.

2. **Global step** over machines:

   Each cluster is associated with a single processor. One processor can have multiple clusters. We jointly resample the processor allocations of all data points within a given cluster. We use a Metropolis-Hastings step with a proposal distribution that independently assigns cluster k to processor j with probability $1/P$. This means our accept/reject ratio depends only on the ratio of the likelihoods of the current processor assignments $\{\gamma_i\}$ and the proposal $\{\gamma_i^*\}$.

   The likelihood ratio is given by:

$$\frac{p\left(\{\gamma_i^*\}\right)}{p\left(\{\gamma_i\}\right)} = \frac{p\left(\{z_i\}|\gamma_i^*\right) p\left(\{\gamma_i^*\}|\alpha, P\right)}{p\left(\{z_i\}|\gamma_i\right) p\left(\{\gamma_i\}|\alpha, P\right)} \tag{D.1}$$

$$= \prod_{j=1}^{P} \prod_{i=1}^{\max(N_j, N_j^*)} \frac{a_{ij}!}{a_{ij}^*!} \tag{D.2}$$

   where N_j is the number of data points on machine j, and a_{ij} is the number of clusters of size i on machine j. The derivation is shown in the supplementary material of [62].

Appendix E

HYPERPRIOR SETTINGS

E.1 Hyperprior for σ_0^2

We set the hyper priors for $\sigma_0^2 \sim \text{IG}(\alpha_{\epsilon 0}, \beta_{\epsilon 0})$ with hyperparameters $\alpha_{\epsilon 0} = 0.5, \beta_{\epsilon 0} = 1$. When examining the housing data, for numerical stability we multiply the observations $[\log(\text{Price}_{t,i,l}) - \log(g_t)]$ by a factor of 200. As a result, 99% of outcome values are covered by the interval $[-1.10, 1.61]$. The chosen hyper prior has a long and flat tail distribution over the range of variance.

E.2 Hyperprior for R_i

We set the hyper priors for $R_i \sim \text{IG}(\alpha_{R0}, \beta_{R0})$ with hyperparameters $\alpha_{R0} = 3, \beta_{R0} = 1$. The chosen hyper prior has a long and flat tail distribution over the range of variance.

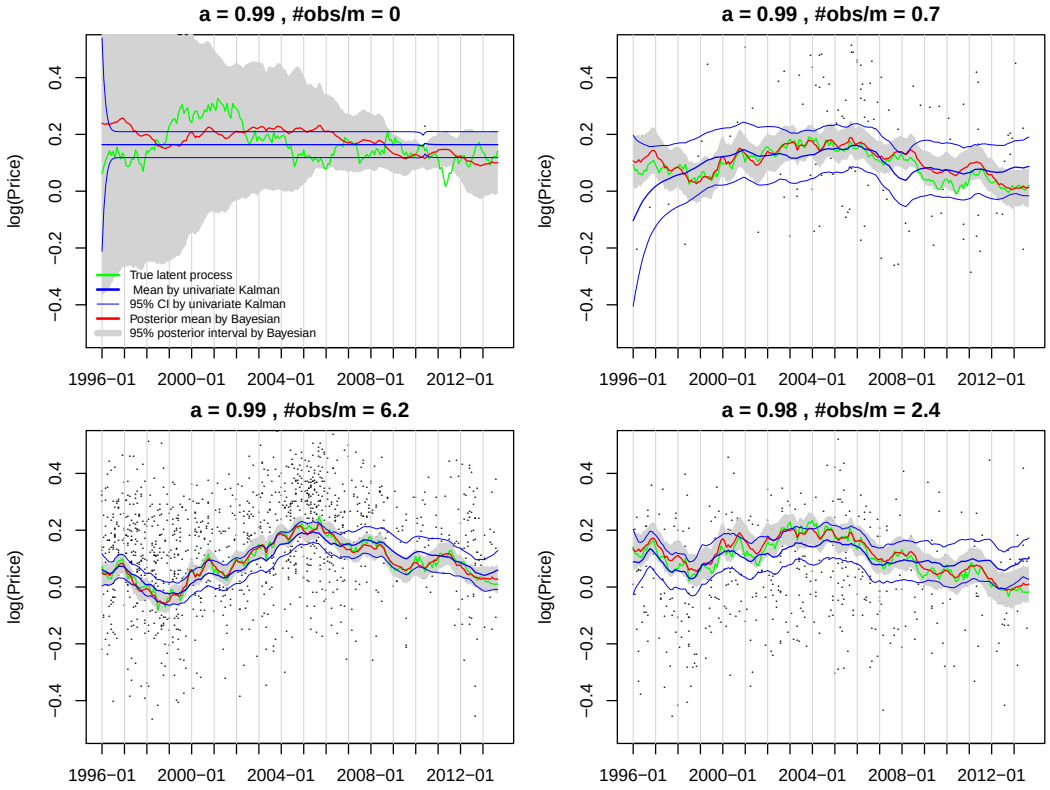


Figure F.1: Performance of estimating the latent process for Cluster 2.

Appendix F

EXTENDED SIMULATION RESULTS

In this section, we provide a performance analysis of the remaining clusters not examined in Section 3.7.2 of the main paper. Figures F.1, F.2 and F.3 directly parallel Figure 3.7 of the main paper and show our performance in estimating the simulated intrinsic price dynamics compared to an independent Kalman-filter-based analysis of the tracts.

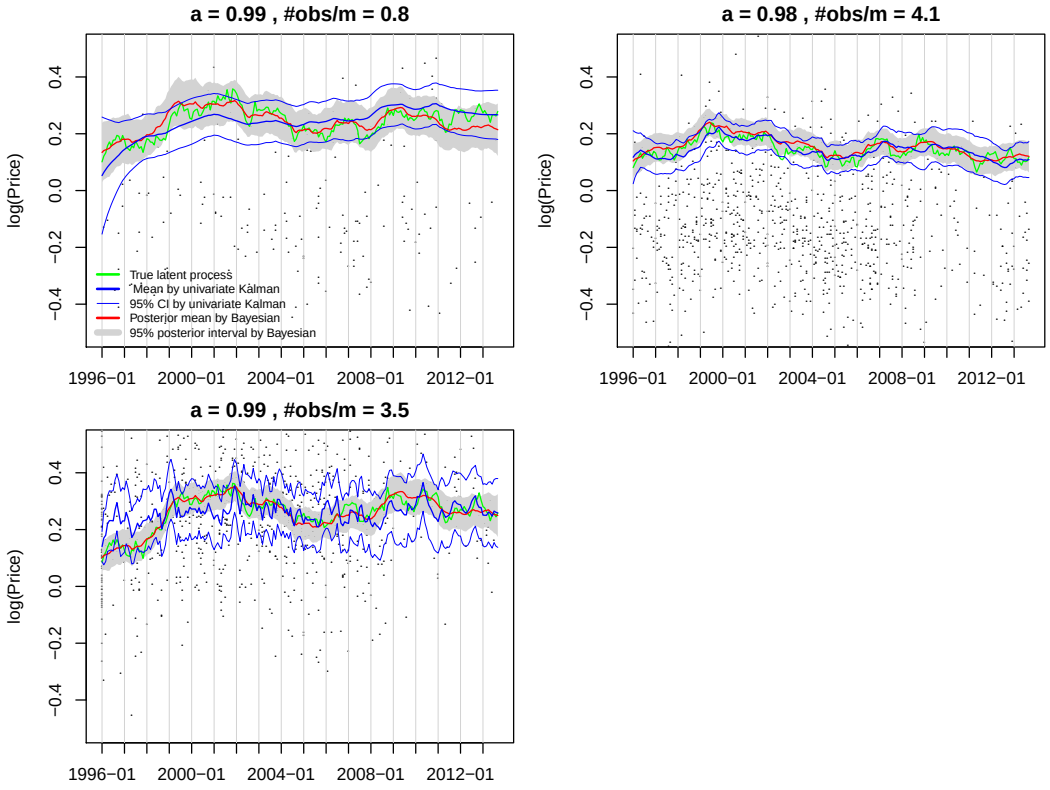


Figure F.2: Performance of estimating the latent process for Cluster 3.

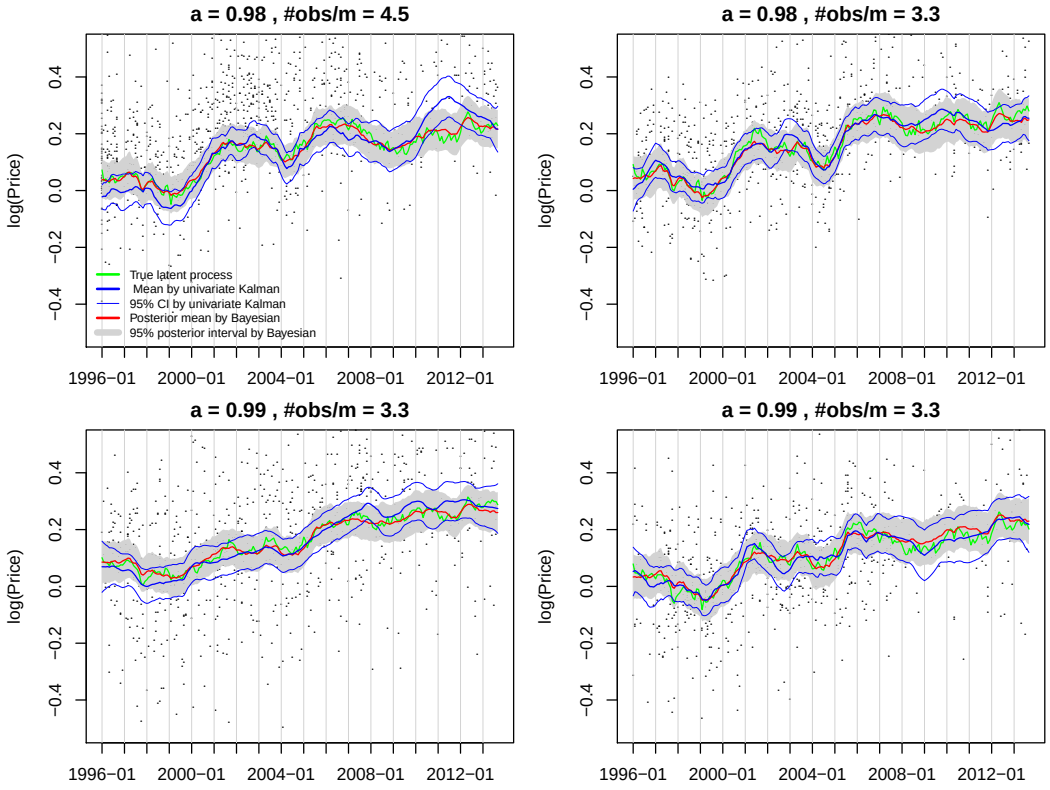


Figure F.3: Performance of estimating the latent process for Cluster 4.

Appendix G

EXTENDED SEATTLE CITY RESULTS

In this section, we present a set of figures from our Seattle City data analysis to augment those presented in the main paper. For the MAP MCMC sample, Figure G.1 displays the average of the intrinsic price processes within a cluster, for each of the 16 inferred clusters. This plot parallels that of Figure 3.10 of the main paper, but here in the raw price space instead of $\log$ space and with the global trend added without the seasonality component (for clarity). We additionally hold on the estimated global trend without seasonality for comparison. In Figure G.2, we compare the resulting housing index produced by S&P Case–Shiller, Zillow Home Value Index (ZHVI), and our Bayesian method at the Seattle City level.

G.1 Sales volume and variance over time

Figure G.3 shows the sales volume and its variance over time, as discussed in Section 3.8.1, together with Figure 3.12, of the main paper. The market boom, roughly 2006-2007, and subsequent stabilization, roughly 2010-2011, were manifested in the different housing sectors in disparate ways. The index formed from the model based on DP clustering is able to capture the dynamics of the change in value for different housing sectors during these two periods.

G.2 Trace plots for convergence diagnostic

Figure G.4 shows some selected trace plots of MCMC chains. Three chains (in *black, red, green*) are initialized with values randomly sampled from the prior. It shows good mixing and convergence.

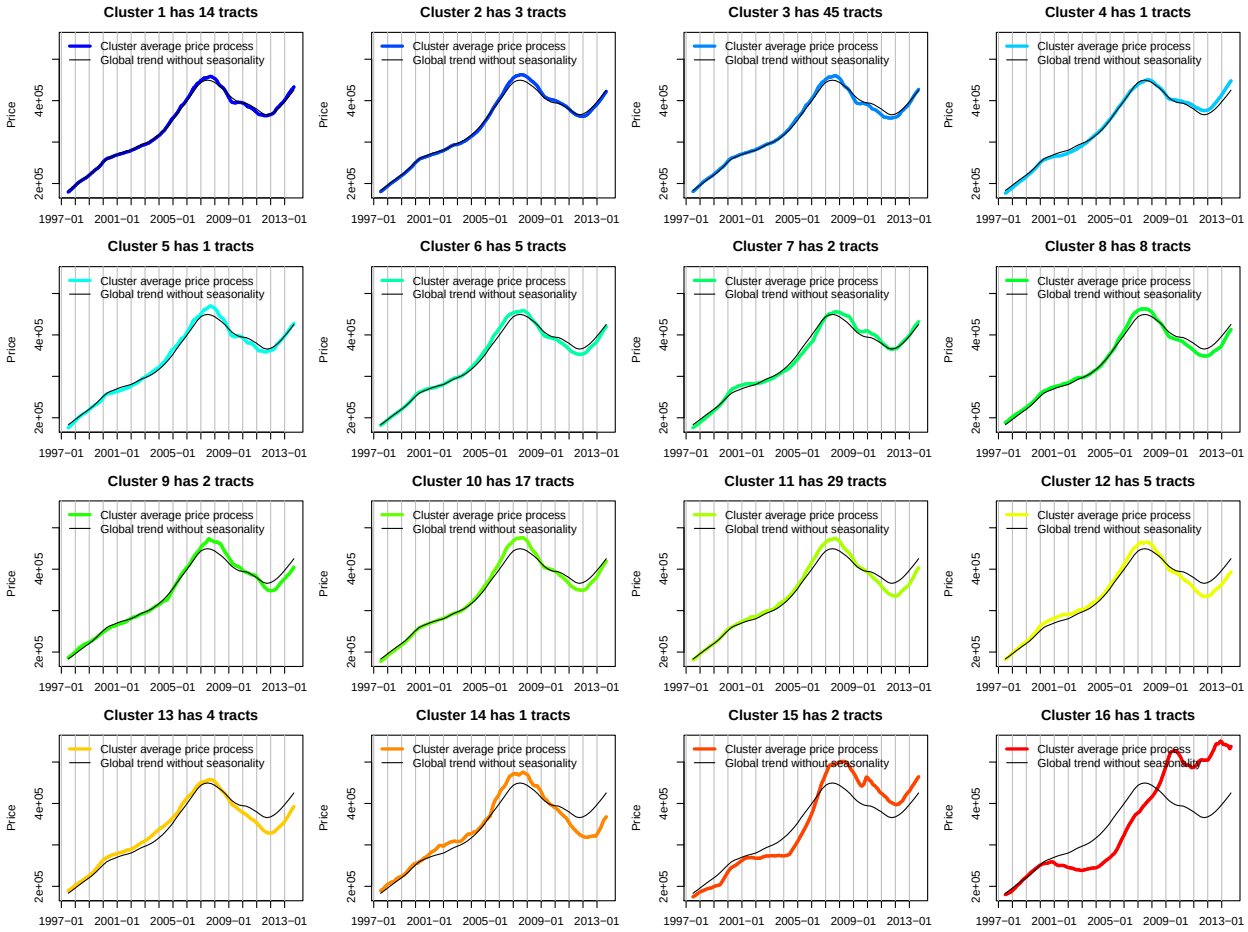


Figure G.1: Under the MAP sample, we first compute the cluster-average intrinsic price dynamics by averaging $x_{1:T,i}$ over all i with $z_i = k$ for $k = 1, ..., 16$ (all of the estimated clusters). We then add this cluster-average price to the global trend without seasonality (*various colors*) and hold on the seasonally adjusted global trend (*black*) for comparison.

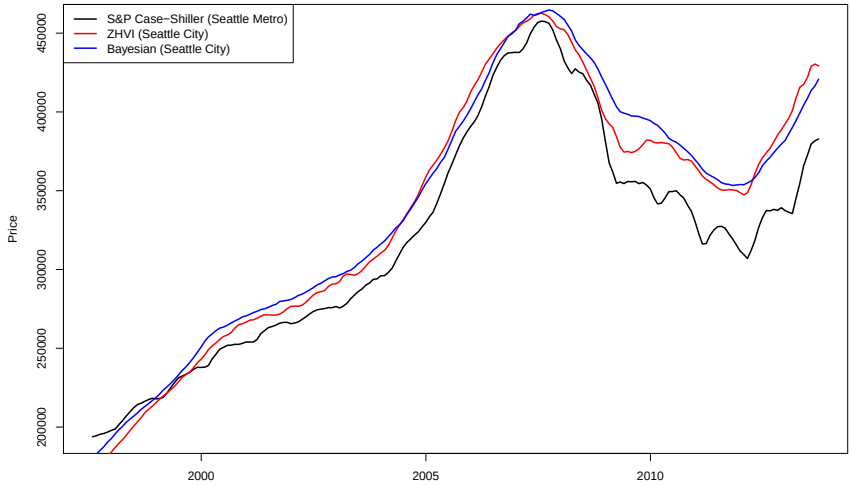


Figure G.2: Seattle City Price Index by S&P Case–Shiller, Zillow Home Value Index (ZHVI) and our proposed Bayesian method.

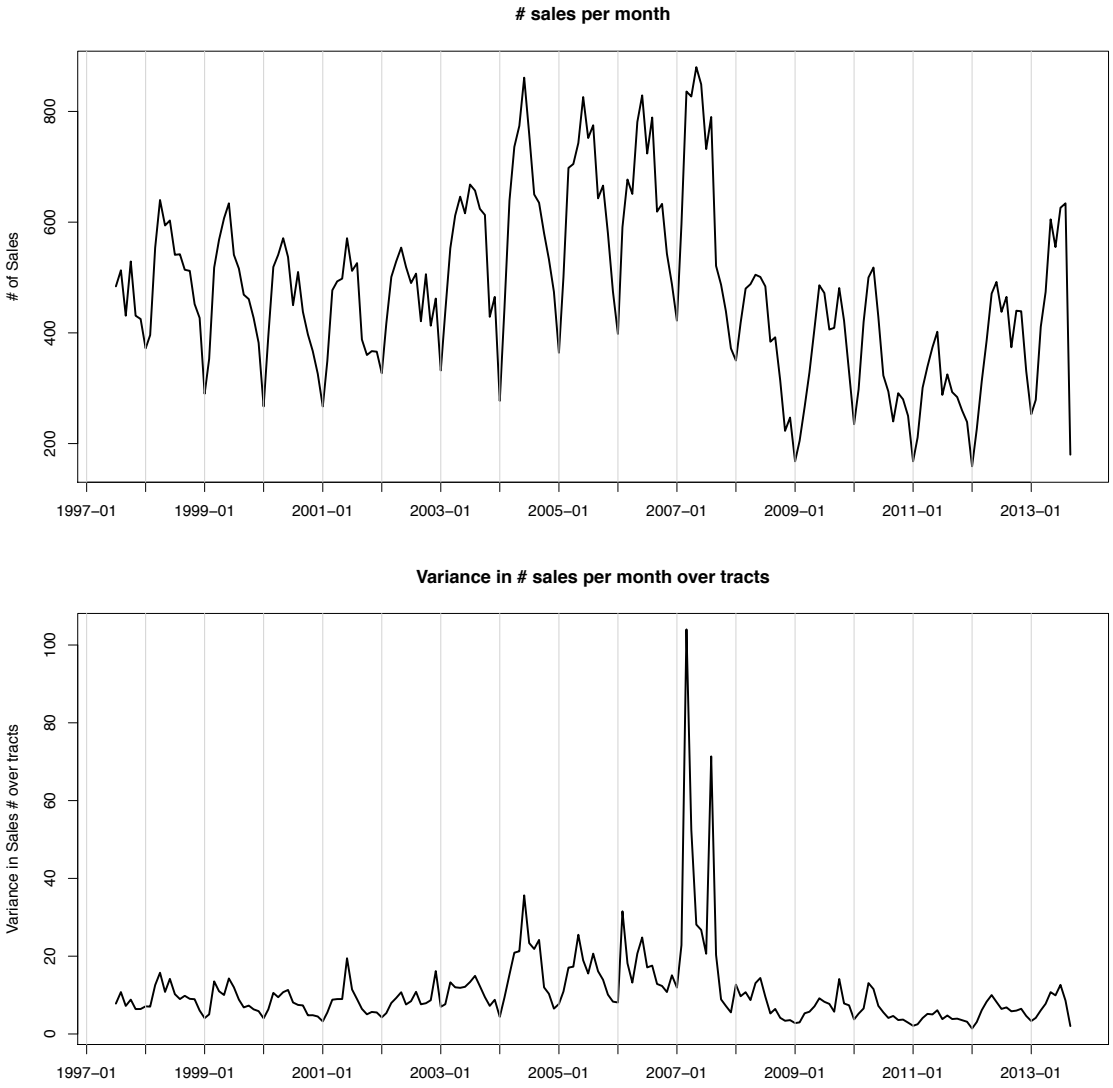


Figure G.3: Sales volume (*top*) and variance (*bottom*) versus time.

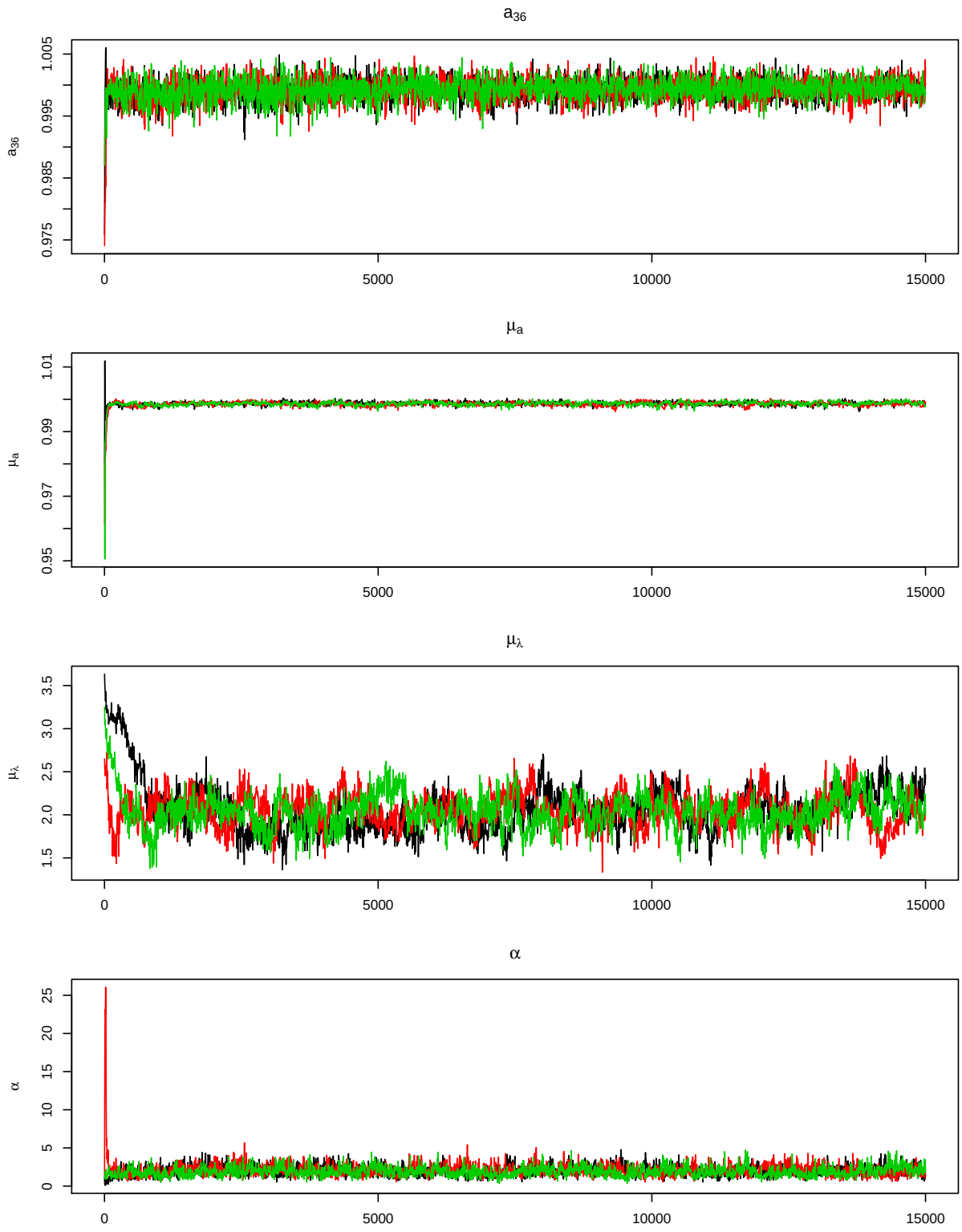


Figure G.4: Selected trace plots for convergence diagnostic. Three chains (in *black, red, green*) are initialized with values randomly sampled from the prior. It shows good mixing and convergence.

BIBLIOGRAPHY

[1] S. Aldor-Noiman, L. D. Brown, E. B. Fox, and R. A. Stine. Spatio-Temporal Low Count Processes with Application to Violent Crime Events. *ArXiv e-prints*, April 2013.

[2] Sharon Alpert, Meirav Galun, Achi Brandt, and Ronen Basri. Image segmentation by probabilistic bottom-up aggregation and cue integration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):315–327, 2012.

[3] Martin J. Bailey, Richard F. Muth, and Hugh O. Nourse. A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):pp. 933–942, 1963.

[4] Ramaprasad Bhar and Shigeyuki Hamori. *Hidden Markov models: applications to financial economics*, volume 40. Springer Science & Business Media, 2006.

[5] D. Blackwell and J. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

[6] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[7] Pascal Bondon and Wilfredo Palma. Asymptotics for linear predictors of strongly dependent time series. In *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pages 847–852. IEEE, 2005.

[8] George EP Box and Gwilym M Jenkins. *Time series analysis: forecasting and control, revised ed*. Holden-Day, 1976.

[9] B. Case and J. M. Quigley. The dynamics of real estate prices. *Rev. Econ. Statist.*, pages pp. 50–58, 1991.

[10] Karl E. Case and Robert J. Shiller. Prices of single family homes since 1970: New indexes for four cities. *N. Engl. Econ. Rev.*, pages 45–56, Sept./Oct. 1987.

[11] Karl E. Case and Robert J. Shiller. The efficiency of the market for single-family homes. *Amer. Econ. Rev.*, pages 125–137, 1989.

[12] Ngai Hang Chan and Wilfredo Palma. State space modeling of long-memory processes. *Annals of Statistics*, pages 719–740, 1998.

[13] Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73, 1990.

[14] Dato. Graphlab create. `http://dato.com/products/create/`.

[15] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[16] P. Englund, J. M. Quigley, and C. L. Redfearn. The choice of methodology for computing housing price indexes: Comparisons of temporal aggregation and sample definition. *Real Estate Fin. Econ.*, pages pp. 91–112, 1999.

[17] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1994.

[18] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[19] Gary William Flake, Steve Lawrence, and C Lee Giles. Efficient identification of web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160. ACM, 2000.

[20] E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Ph.D. thesis, MIT, Cambridge, MA, 2009.

[21] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:2007, 2007.

[22] D. H. Gatzlaff and D. R. Haurin. Sample selection bias and repeat-sales index estimates. *J. real Estate Fin. Econ.*, pages pp. 33–50, 1997.

[23] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[24] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[25] W. N. Goetzmann and L. Peng. The bias of the rsr estimator and the accuracy of some alternatives. *Real Estate Econ.*, pages pp. 13–39, 2002.

[26] Stefano Grassi and Paolo Santucci de Magistris. When long memory meets the kalman filter: A comparative study. *Computational Statistics & Data Analysis*, 76:301–319, 2014.

[27] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

[28] Jonathan RM Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 1981.

[29] Michael C Hughes and Erik Sudderth. Memoized online variational inference for dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 1133–1141, 2013.

[30] Matteo Iacoviello. Housing wealth and consumption. *Board of Governors of the Federal Reserve System, International Finance Discussion Papers*, 2011.

[31] Yannis M Ioannides and Jeffrey E Zabel. Neighbourhood effects and housing demand. *Journal of applied Econometrics*, 18(5):563–584, 2003.

[32] Yannis M Ioannides and Jeffrey E Zabel. Interactions, neighborhood selection and housing demand. *Journal of urban economics*, 63(1):229–252, 2008.

[33] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[34] Daniel Jurafsky and James Martin. Speech and language processing: An introduction to speech recognition. *Computational Linguistics and Natural Language Processing. Prentice Hall*, 2008.

[35] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[36] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2):291–307, 1970.

[37] Mingche M Li and H James Brown. Micro-neighborhood externalities and hedonic housing prices. *Land economics*, pages 125–141, 1980.

[38] Dougal Maclaurin and Ryan P. Adams. Firefly monte carlo: Exact mcmc with subsets of data. In *Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, 07/2014 2014.

[39] R. A. Meese and N. E. Wallace. The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *J. Real Estate Fin. Econ.*, pages pp. 51–73, 1997.

[40] Peter Müller and Fernando A Quintana. Nonparametric bayesian data analysis. *Statistical science*, pages 95–110, 2004.

[41] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1):32–38, 1957.

[42] Chaitra H. Nagaraja, Lawrence D. Brown, and Linda H. Zhao. An autoregressive approach to house price modeling. *The Annals of Applied Statistics*, 5(1):124–149, 03 2011.

[43] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[44] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[45] Luis E. Nieto-Barajas and Alberto Contreras-Cristn. A bayesian nonparametric approach for time series clustering. *Bayesian Anal.*, 9(1):147–170, 03 2014.

[46] Konstantina Palla, Zoubin Ghahramani, and David A. Knowles. A nonparametric variable clustering model. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2987–2995. Curran Associates, Inc., 2012.

[47] Robert S Pindyck and Daniel L Rubinfeld. *Econometric models and economic forecasts*, volume 4. Irwin/McGraw-Hill Boston, 1998.

[48] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[49] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[50] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

[51] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[52] Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, et al. mgene: accurate svm-based gene finding with an application to nematode genomes. *Genome research*, 2009.

[53] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[54] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[55] R. Shiller. Arithmetic repeat sales price estimators. *J. Housing Econ.*, pages pp. 110–126, 1991.

[56] Robert H Shumway and David S Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

[57] P. L Smith. Splines as a useful and convenient statistical tool. *The American Statistician*, 33(2):5762, 1979.

[58] Erik B Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Citeseer, 2006.

[59] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[60] Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Science & Business Media, 2013.

[61] Stephen G Walker, Paul Damien, Purushottam W Laud, and Adrian FM Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 485–527, 1999.

[62] Sinead Williamson, Avinava Dubey, and Eric P. Xing. Parallel markov chain monte carlo for nonparametric mixture models. In *ICML*, pages 98–106, 2013.

[63] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

[64] Zillow. Zillow home value index: Methodology. http://www.zillow.com/research/zhvi-methodology-6032/, 2014.