Exploring Spatio-Temporal Pattern of Connecticut Housing

Yuen Tsz Abby Lau

5/12/2019

Abstract

This study examines the Bayesian spatio-temporal methods to analyze local patterns of housing rate over time at the county level by an application to the housing data in the state of Connecticut, United States.

The method used in this paper was proposed by Napier, Lee, and Robertson (2016) in the paper *A Model to Estimate the Impact of Changes in MMR Vaccination Uptake on Inequalities in Measles Susceptibility in Scotland* where they explored the Bayesian spatio-temporal method to study epidemiological questions about vaccination rates in Scotland. In this paper, the Bayesian model was extended to monitor the spatio-temporal change of CT housing rate and it estimated (1) overall temporal trend, and (2) area-specific differential trend.

Introduction

Bayesian Spatio-Temporal Approaches

As spatio-temporal data has become increasingly popular, a comprehensive statstical understanding of this data is meaningful. And considering both the spatial effects and temporal effects simultaneously renders richer information from the data and thus forms a bigger, clearer, and more interesting picture for data.

Currently there are many statistical models available for capturing the spatio-temporal pattern, and in this study, Bayesian spatio-temporal methods are applied. The methods are hierarchical and utilize parameters which follow certain probability distributions, which are specified by some prior believes or knowledge. For instance, neighborhood information or other known information can be included in the models to provide a better representation of the autocorrelation among all spatial units. Moreover, the posterior probability can render an idea of whether a parameter value exceeds certain threshold. As a result, the Bayesian methods take a full advantage of all the available information, provide a more meaningful representation of parameters, and return more useful results. Moreover, Bayesian methods are favored because they stabilize the area-specific risks as they can borrow strength of neighborhood areas.

Datasets

The primary dataset is given and maintained by the Office of Policy and Management, State of Connecticut. It lists all real estate sales with a sales price of $2,000 or greater that occur between October 1 and September 30 of each year. For each sale record, the file includes: town, property address, date of sale, property type (residential, apartment, commercial, industrial or vacant land), sales price, and property assessment. Number of sale records by year for each county can be aggregated by this dataset.

The additional dataset is the annual housing inventory data by town provided by the Department of Economic and Community Development. This dataset has the housing inventory from 2000 to 2017 based on census 2000 and 2010 data. Number of properties by year for each county can be obtained from this dataset.

The average sales rate for county in each year is then calculated as

$$\frac{\text{number of sale records}}{\text{number of properties}}$$

Analysis Tool

`CARBayesST` is an `R` package developed by Duncan Lee, Alastair Rushworth, and Gary Napier. It provides multiple models for spatial data relating to non-overlapping areal units, and this package stands out as it includes models that explore the overall spatial-temporal trend with different focuses. Many of the models utilize the conditional autoregressive (CAR) priors to capture the spatial auto-correlation in the data, and depending on the nature of the data, the users can select certain models to fulfill their needs.
Furthermore, in all the cases the inference in this software is based on Markov chain Monte Carlo (MCMC) stimulation which is used in the following proposed method. Hence, this package is selected to conduct Bayesian analysis in this study.

Proposed Method

The study region is first divided as $k = 1, \ldots, K$ non-overlapping areal units, and data are recorded for each unit for $t = 1, \ldots, N$ time periods.
For this study, $N = 16$ and $K = 8$ because the primary dataset has sales records from 2011 to 2016 for all 8 counties in Connecticut. Thus, Y_{kt}, the response variable, is the number of observed housing sales records.
It is assumed that Y_{kt} follows a possion distribution with a parameter $\lambda_{kt} = n_{kt}\theta_{kt}$, where θ_{kt} is the rate of property sales as a proportion of the total number of properties n_{kt}. The possion distribution is considered here as a property could be traded multiple times throughout years. As a result, θ_{kt} is not strictly the proportion of properties that sell in a year, but is on approximately the same scale for interpretation purposes (Lee, Rushworth, Napier, 2018). Hence,

$$Y_{kt} \sim Poisson(\lambda_{kt})$$

.

Then, by looking at the log of λ_{kt},

$$ln(\lambda_{kt}) = \boldsymbol{x}_{kt}^T \beta + O_{kt} + \psi_{kt}$$

where $\boldsymbol{x}_{kt} = (x_{kt1}, \ldots, x_{ktp})$ is a $1 \times P$ vector which represents p known covariates for unit k and time period t, $\boldsymbol{O}_t = (O_{1t}, \ldots, O_{Kt})$ denotes the $1 \times K$ column vector of offsets for time period t, and $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N)$, where $\boldsymbol{\psi}_t = (\psi_{1t}, \ldots, \psi_{Kt})$. The ψ_{kt} term is a latent component for area k and time period t encompassing one or more sets of spatio-temporally autocorrelated random effects.

Under the spatio-temporal random effects models, the spatial autocorrelation is controlled by the random effects in the neighborhood areas. A non-negative and symmetric $K \times K$ neighborhood matrix is defined to describe whether each pair of areas share one or more common vertex between boundaries nor not. Let's denote the neighborhood matrix as $\boldsymbol{W} = (w_{kj})$, where w_{kj} denotes the spatial adjacency between area i and j. If area i and j share a common border, $w_{kj} = 1$; and if area i and j are not spatially adjacent, then $w_{kj} = 0$.

Napier *et al.* (2016) proposed the following specification of the spatio-temporal structure for ψ_{kt},

$$\psi_{kt} = \phi_{kt} + \delta_t$$

$$\phi_k t | \boldsymbol{\phi}_{-kt}, \boldsymbol{W} \sim N(\frac{\rho_S \sum_{j=1}^K w_{kj}\phi_{jt}}{\rho_S \sum_{j=1}^K w_{kj} + 1 - \rho_S}, \frac{\tau_S^2}{\rho_S \sum_{j=1}^K w_{kj} + 1 - \rho_S})$$

$$\delta_t | \boldsymbol{\delta}_{-t}, \boldsymbol{D} \sim N(\frac{\rho_T \sum_{j=1}^N d_{tj}\delta_j}{\rho_T \sum_{j=1}^N d_{tj} + 1 - \rho_T}, \frac{\tau_T^2}{\rho_T \sum_{j=1}^K w_{kj} + 1 - \rho_T})$$

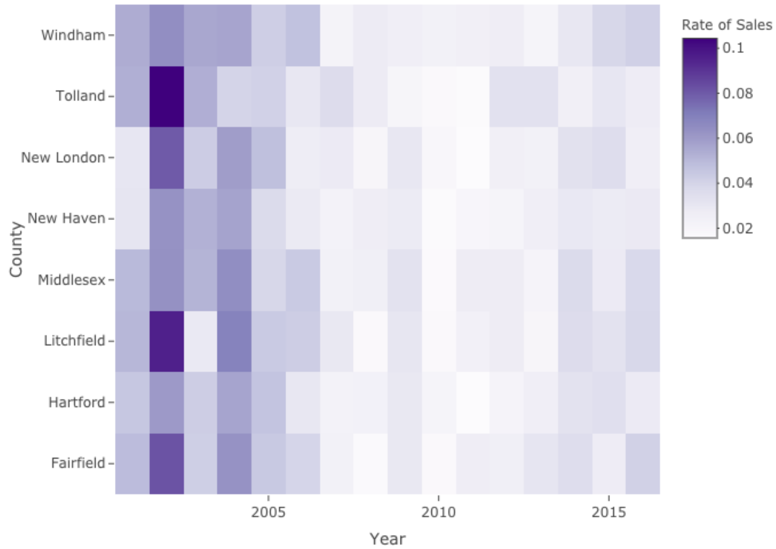

Figure 1: Heatmap showing that the spatial-temporal trend in the rate of property sales as a porportion of the total number of the propoerties between 2001 and 2016

$$\tau_1^2, \ldots, \tau_N^2, \tau_T^2 \sim Inverse - Gamma(a, b)$$

$$\rho_S, \rho_T \sim Uniform(0, 1)$$

where $\boldsymbol{\psi}_{-kt} = (\psi_{1t}, \ldots, \psi_{k-1,t}, \psi_{k+1,t}, \ldots, \psi_{Kt})$ and $\boldsymbol{\delta}_{-t} = (\delta_1, \ldots, \delta_{t-1}, \delta_{t+1}, \ldots, \delta_K)$. The model explain the overall spatio-temporal trend by a common temporal trend and a spatial variation term associated with each spatial unit. CAR priors suggested by Leroux *et al.* (2000) are used here for the model.

Explanatory Analysis

To start with, the duplicate observations from the primary dataset has been excluded. The number of property sales is then calculated from the dataset.

Before fitting model, the spatio-temporal pattern of average sales rate can be examined by a heatmap, where the y-axis is the spatial stamp and x-axis is the temporal index. Figure 1 is a heatmap shows the average sales rate for each county from year 2001 to year 2016. It is obvious that the rates were higher for all counties for early 2000s, but they reached the minimum around 2007 and 2008, when the financial crisis occurred. It is also important to notice that the rates are different for each county. Hence, there are both a spatial and temporal pattern for the sales rate. Moreover, the spatial-temporal pattern for each county is not necessarily the same, and it seems that similarities exist for close counties in consecutive time periods.

To include the spatio-temporal structure in the model, a neighborhood matrix is created so that the areas which share a common border have entries of 1 and the areas that are not spatially adjacent get entries of 0.

	fairfield	hartford	litchfield	middlesex	new haven	new london	tolland	windham
fairfield	0	0	1	0	1	0	0	0
hartford	0	0	1	1	1	1	1	0
litchfield	1	1	0	0	1	0	0	0
middlesex	0	1	0	0	1	1	0	0

	fairfield	hartford	litchfield	middlesex	new haven	new london	tolland	windham
new haven	1	1	1	1	0	0	0	0
new london	0	1	0	1	0	0	1	1
tolland	0	1	0	0	0	1	0	1
windham	0	0	0	0	0	1	1	0

Model and Findings

With the neighborhood matrix being constructed, the data can be fitted using `CARsepspatial()`, and the following code is used for fitting the model. The model uses 220,000 MCMC samples, and the first 20,000 of them are removed by the burn-in period. The samples are then thinned by 10 to reduce the autocorrelation of the Markov chain, resulting in 20,000 samples for inference (Lee, *et al.*, 2018).

```
setorder(dat_combine, ListYear)

formula <- Freq ~ offset(log(stock))

model1 <- ST.CARsepspatial(formula = formula,
                           family = "poisson",
                           data = dat_combine,
                           W = W,
                           burnin = 20000,
                           n.sample = 220000,
                           thin = 10)
```

The offset term renders the desired response variable, θ_{kt}, the rate of property sales. The estimated result can be interpreted as follows,

$$\theta_{kt} = exp(\beta_1 + \phi_{kt} + \delta_t)$$

which is the sum of an overall intercept term β_1, a space-time covariate ϕ_{kt} with a time period specific variance, and a region-wide temporal trend δ_t (Lee, *et al.*, 2018). The mean and standard deviation of θ_{kt} over space for each year is then plotted as follows, which displays the posterior median and spatial variation with 95% credible intervals for each quantity for each year. A heatmap (Fig 3) is also created to plot all the posterior median sales rate θ_{kt}.

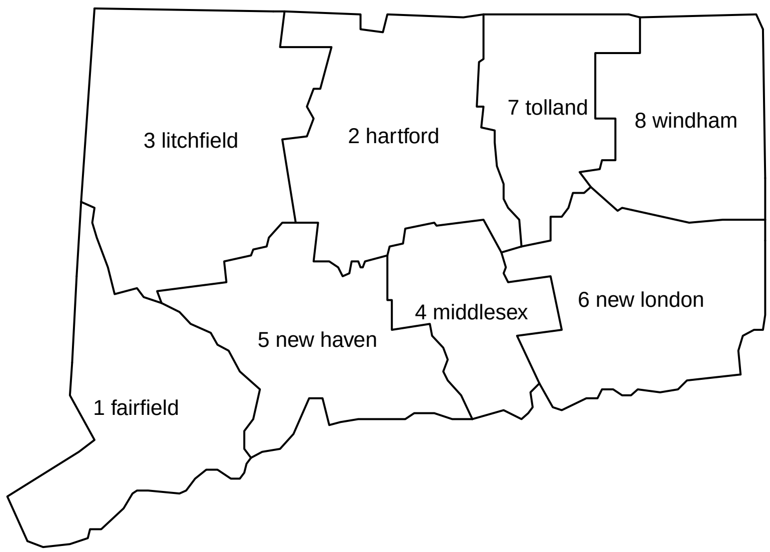

Figure 2: Map of Connecticut Counties

Conclusion and Discussion

Based on the previous plots, the average sales rate started to drop since 2005 and has reached its minimum during the financial crisis. It also started to climb up recently, starting from 2011. The spatial variations have similar trends, which suggests that the finical crisis negatively impacted the sales rate. It is also important to notice that the heatmap show the a clear changing spatial pattern in sales rates over time. The spatial rates for 2001 to 2005 are largely consistent, but a clear change is evident between 2005 and 2008, which is coincident with the beginning of the global financial crisis. In 2013, the spatial pattern is different again. Generally speaking, there is a spatio-temporal pattern existing in the calculated sales rates, which provides some insights about the CT housing market, and indirectly reflects the occurrence of some major economic events in Connecticut, such as global financial crisis in late 2000s. It might helpful to incorporate some covariates to understand and explain the underlying patterns better.

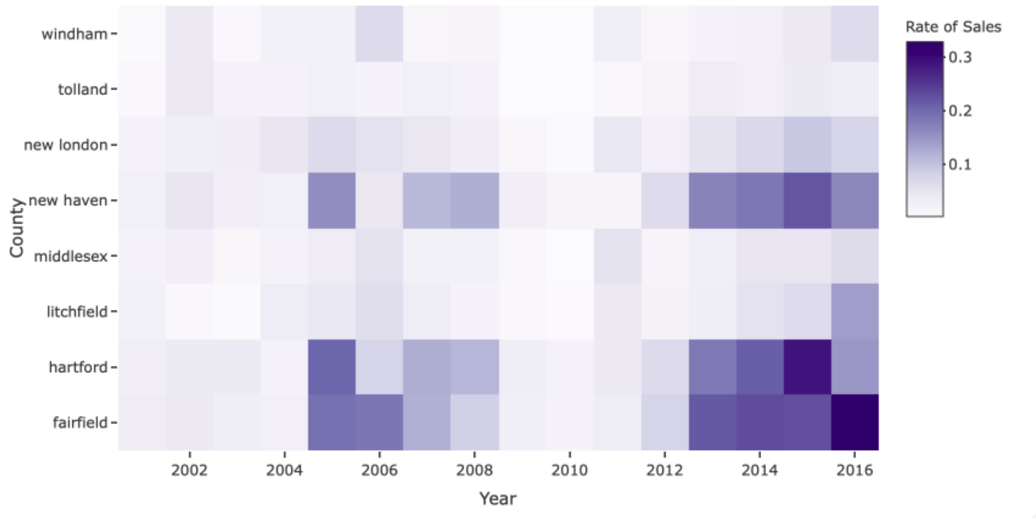

Figure 3: Posterior Median Sales Rate

References

Napier G, Lee D, Robertson C, Lawson A, Pollock K (2016). "A Model to Estimate the Impact of Changes in MMR Vaccination Uptake on Inequalities in Measles Susceptibility in Scotland." *Statistical Methods in Medical Research*, 25, 1185–1200.

Lee D, Rushworth A, Napier G (2018). "Spatio-Temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST Package." *Journal of Statistical Software*, *84*(9), 1-39. doi: 10.18637/jss.v084.i09 (URL:http://doi.org/10.18637/jss.v084.i09).

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019), Spatio-Temporal Statistics with R, Boca Raton, FL: Chapman & Hall/CRC. 2019 Wikle, Zammit-Mangion, Cressie. https://spacetimewithr.org