Spatial Statistical Modeling of housing prices in Yangtze River Delta

Comparing with Pearl River Delta and BTH Coordinated Development

Yunsong Liu

Master's thesis
2024:E24

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Abstract

This paper aims to explore the spatial statistical model of housing prices in the Yangtze River Delta (YRD) and compare it with prices in the Pearl River Delta (PRD) and the Beijing-Tianjin-Hebei Cooperative Development Area (BTH). By collecting the relevant data such as population structure, economic development level, and education level, and using a spatial statistical method, the paper constructs a spatial statistical model of the housing prices in these three regions. It is found that the housing prices in the YRD region show unique spatial distribution characteristics that are mainly affected by geography. In contrast, housing prices in the PRD and BTH regions also have their own characteristics, showing spatial agglomeration centered on a certain city. The models are based on a Bayesian approach, and inference is made using Integrated Nested Laplace Approximation (INLA).

Keywords: Housing price; Spatial statistical model; Temporal component; Bayesian inference

Acknowledgements

Time flies. Two years of college life will come to an end. In this precious time, I have gained a lot, but also grown up a lot. The completion of my graduation thesis is not only an important milestone in my academic career, but also a valuable asset on my life path. Here, I would like to express my heartfelt thanks to all those who have given me help and support.

Firstly, I'd like to thank my mentor. From the topic selection to the writing of the paper, the tutor has always given me careful guidance and patient help. My mentor's rigorous academic attitude, profound academic attainments, and selfless dedication deeply inspired and moved me. In the process of writing my thesis, my tutor not only helped me sort out my thoughts, but also reviewed my thesis word by word and pointed out the existing problems and shortcomings, so that my thesis was constantly improved. Here, I would like to express my sincerest respect and gratitude to my teacher.

And I would like to thank the school for providing me with a good academic atmosphere and learning environment. The university's library, laboratory, and other resources are abundant, which provides great convenience for me to write my thesis. At the same time, the school teachers also gave me a lot of guidance and help, so that I have made academic progress.

Looking back on the past, I feel that my growth and progress cannot be achieved without the help and support of so many people. In the coming days, I will continue to work hard, constantly pursue progress, and repay all those who care and support me with practical actions. Thanks again to everyone who has given me help and support!

Contents

1 Introduction

Since the commercialization reform of China's urban housing system, despite China's housing market being hit by multiple financial crises, urban housing prices have generally maintained a steady and rapid growth. In order to promote the stable and healthy development of the real estate market, the central government proposed a new policy in 2016 that "*houses are for living in, not for speculation.*" However, judging from the actual results in recent years, the rapid rise in housing prices in large cities has not been fundamentally curbed. Against this background, economics and geographers in China and abroad are increasingly aware of the necessity and urgency of studying the housing price issue in Chinese cities.

The Yangtze River Delta (YRD), located on the eastern coast of China, is one of the most economically vibrant and densely populated regions in the country. With rapid urbanization and economic growth, the demand for housing in the YRD has skyrocketed, leading to significant variations in housing prices across different cities and localities. Understanding the spatial patterns and determinants of housing prices in the YRD is crucial for policymakers, real estate developers, and investors to make informed decisions. Comparative analyses with other economically advanced regions, such as the Pearl River Delta (PRD) and Beijing-Tianjin-Hebei (BTH) Coordinated Development Area, provide valuable insights into the unique characteristics and challenges of the YRD's housing market.

A previous study [1] concludes that the spatial differentiation of housing prices among Chinese cities is significant, showing patterns of both spatial agglomeration differentiation (between the three southeastern coastal urban agglomerations and inland cities) and administrative hierarchical differentiation (between provincial capital cities and prefecture-level cities).

This thesis establishes a spatial model for the housing prices of three megalopolis on the southeast coast, so as to gain insight into the differences in housing prices and the growth patterns of housing prices in cities within that region.

1.1 History of spatial analysis

Statistical analysis of spatial data was born in Germany at the end of the 18th century and the beginning of the 19th century. In 1909, German economist Alfred Weber [2] founded the industrial location theory. The central place theory proposed by geographer Walter Christaller [3] in 1933 included non-productive service industries for the first time. In economic activities, the urban level, scale, and spatial distribution patterns were systematically discussed. Homer Hoyt's Sector Model [4] in 1939 described urban land use and the spatial patterns of residential areas.

On the theoretical side, one of the most influential contributors during this period was Georges Matheron, a French mathematician who pioneered the field of geostatistics. In the 1960s, Matheron [5] developed the theory of regionalized variables and introduced kriging, a method for spatial interpolation. Kriging, named after South African mining engineer Danie Krige [6], provided a means to make the best linear unbiased predictions for spatially correlated data. This technique became essential not only in mining and environmental science but also in the analysis of economic and housing data, allowing for more accurate spatial predictions and modeling.

In the realm of spatial statistics, Peter Whittle made significant strides. A British mathematician, Whittle focused on the theory of stochastic processes. In 1954, he worked on the spatial prediction of random fields and provided a robust mathematical foundation for analyzing spatially distributed data [7]. Whittle's contributions were crucial in the development of spatial econometric models, which account for the interdependence of spatial data points, leading to more accurate and reliable analyses.

The 1960s also witnessed important advancements in the quantitative analysis of urban land use. William Alonso introduced the Bid Rent Theory [8] in 1964, a seminal work that explained how land values and housing prices decrease with distance from the central business district. Alonso's theory provided an economic framework to understand urban spatial structures, influencing subsequent studies on urban development and housing markets.

In 1970, Tobler [9], a geographer at the University of Michigan, proposed the first law of geography: "*Everything is related, but things that are close are more closely related than things that are far away.*", it laid the foundation for spatial statistics.

In the 1990s, The combination of spatial data statistical analysis and GIS symbolizes the maturity of spatial statistics. Today, data accumulated over time (called spatio-temporal big data) and spatiotemporal statistics that combine spatial statistics and time statistics (time series analysis) are widely used in many aspects of society.

Internationally, scholars now have mainly discussed the status quo and connotation of urbanization quality from the ecological perspective, such as sustainable development and urban ecological civilization construction. For example, Yehua Dennis Wei [10] and Simon Elias Bibri [11] both discussed urbanization and sustainable urban development. Sajal Ghosh [12] analyzed the balance between urbanization, energy consumption, and economic activities by summarizing the practical experience of India.

1.2 A basic model for housing prices

In this thesis, housing prices refer to unit housing prices per square meter. Previous research [13] has found that due to the difference in housing prices potential energy between cities and the existence of the "ripple effect", housing prices are usually transmitted from economic center cities to adjacent cities and peripheral areas. Especially in integrated areas with close economic connections, urban housing prices show significant "spatial dependence" and spillover effects from high to low.

Typically, Hedonic housing price models [14] are powerful in econometrics. Although its theoretical basis is sound and appealing, these applications often encounter difficulties relating to its specification. The Hedonic house price function relates the price of a house to the implicit prices of its housing attributes. Thus

$$P_i = \alpha + \sum \beta_k S_{ki} + \sum \gamma_q L_{qi} + \epsilon_i \tag{1.1}$$

where P_i is the housing prices, S_{ki} are structural attributes, L_{qi} are locational attributes and α, β, γ are parameters. Hedonic house price models imply that the spatial model can be built as a linear predictor with covariates replacing S_{ki} and spatial effects replacing L_{qi}. In this thesis, we will also incorporate and study the temporal effect, considering the changes in house prices over time (years), which is called time series analysis. Usually, an AR model is a preliminary method for analyzing time series [15]. However, in practical applications, time is typically used as a covariate, and its relationship with the dependent variable needs to be addressed before fitting the model.

1.3 Data

This thesis collected data from 2019 to 2022 from **JuHuiData** (`gotohui.com`) for the Yangtze River Delta, Pearl River Delta, and Beijing-Tianjin-Hebei region. This geographical division is mainly derived from history and discussed in more details in Chapter 2. However, the development and sale of real estate also have geographical, political, and humanistic factors. For example, houses near rivers or subways are more popular, and Chinese policies have increased housing prices near schools which implies that the studied areas will likely not be homogeneous with respect to demography and socioeconomic status.

1.4 Purpose of thesis

The rational development of the real estate market and stable housing prices have an important impact on China's economic development. By studying the influencing factors and mechanisms of housing prices, while ensuring economic development and controlling housing prices within a reasonable range, sustainable social and economic development can be achieved, which will help the country and the government to formulate corresponding real estate market control measures and to ensure the stable, orderly and harmonious development of the real estate market.

This study aims to contribute to the existing literature on housing price modeling by providing a comprehensive spatial statistical analysis of housing prices in the YRD while drawing comparisons with the PRD and BTH, and to reveal key factors affecting prices and similarities and differences in the spatial distribution of housing prices and their determinants across different regions and also discuss the possibilities of using the generalized model in practice.

2 Regions and data

An overview of the regions and available data is presented in this Chapter.

2.1 The regions

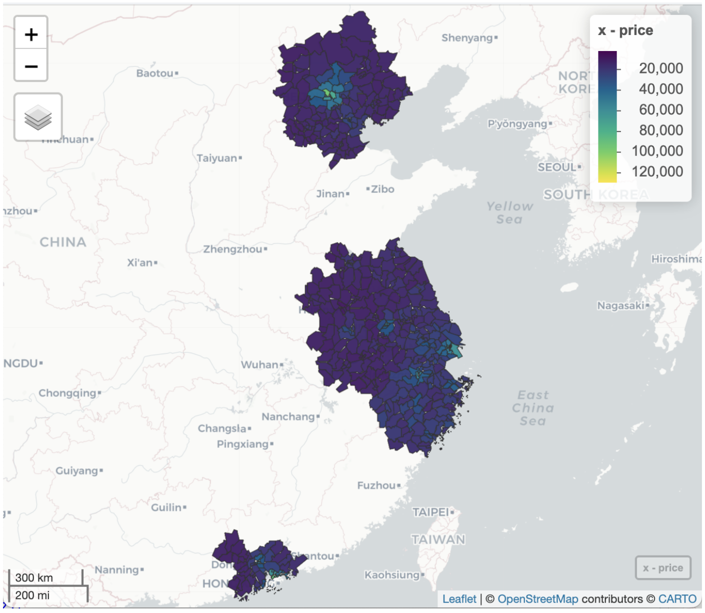


Figure 2.1: China's three major metropolitan areas. The topmost one is the BTH Cooperative Development Area, the middle one is the YRD, and the bottom one is the PRD.

2.1.1 The Yangtze River Delta

The Yangtze River Delta urban agglomeration (YRD) is a highly economically developed region along the eastern coast of China. Its prototype was the Shanghai Economic Zone established on December 22, 1982. It is positioned as an advanced manufacturing base and modern service industry base, a national scientific and technological innovation and technology research and development center, and a leader in radiating the development of the Yangtze River Basin. Cities in the region are geographically adjacent and diverse.

From the perspective of urban housing prices, housing prices in the Yangtze River Delta region are relatively high, growing rapidly, with diverse levels and significant gaps, which better reflect the overall characteristics of housing prices in China. They are highly representative among China's major urban agglomerations and can provide experience and development paths for other high-quality integrated urban agglomerations.

2.1.2 The Pearl River Delta

In December 2008, the regional economic development of the Pearl River Delta (PRD) was elevated to a national strategy. It now concentrates more than 80% of the total economic output of Guangdong Province and has developed into an important economic region and manufacturing center in China. The population density of the Pearl River Delta region ranks among the top three in the world. The main source of population growth is migration from other parts of China. Natural population growth does not have a major impact on population growth.

Compared with the Yangtze River Delta, the Pearl River Delta started later and has lower price levels. Because a large number of working people have flowed into the area, population size has become an important factor affecting urban development. At the same time, the average education level has also been lowered.

2.1.3 The Beijing-Tianjin-Hebei Cooperative Development Area

1981 was the beginning of government departments' research on the integration of Beijing, Tianjin, and Hebei. The Beijing-Tianjin-Hebei Coordinated Development Zone (BTH), officially established in 2015, is an expansion of China's capital economic circle with Beijing as its core. In particular, Beijing is the political and cultural center of China. Compared with the other two economic development zones, this urban agglomeration does not primarily focus on economic development and has a higher level of education. Educational factors and political factors have a greater impact on housing prices in this area.

2.1.4 Dynamics in the regions

To further elaborate on the spatial statistical modeling of housing prices in the Yangtze River Delta (YRD), Pearl River Delta (PRD), and Beijing-Tianjin-Hebei (BTH) coordinated development regions, a few key points are worth mentioning:

- **Economic Policies and Market Dynamics**: The YRD, PRD, and BTH regions all have their unique economic policies and market dynamics that influence housing prices. The YRD, for instance, benefits from robust economic growth and favorable policies that attract both domestic and foreign investments. This drives up demand for housing, particularly in major cities like Shanghai and Hangzhou; the PRD, on the other hand, has a more diversified economy with a strong focus on manufacturing and services. Its real estate market is maturing, offering a wide range of housing options to suit different budgets and preferences; the BTH region, under the influence of the coordinated development strategy, is seeing increasing investments in infrastructure and public services. This is likely to impact housing prices positively in the long run.

- **Population Dynamics and Demographics**: Population growth and migration patterns play a crucial role in shaping housing prices. The YRD, PRD, and BTH regions all have large populations and high urbanization rates. However, the demographic composition and migration trends vary; the YRD, for instance, attracts a significant number of migrants from other parts of China due to its economic opportunities. This drives up demand for housing, particularly in urban areas. The PRD also sees a high level of migration, but its population is more evenly distributed across the region; in the BTH region, under the influence of the coordinated development strategy, the level of population mobility is low and the population numbers in each district have tended to be stable. This could have a negative impact on house prices, but the impact could vary by specific city or region.

2.2 Available data

There are 303 districts and counties in the YRD, 50 districts and counties in the PRD, and 127 districts and counties in the BTH. The housing price refers to the average price per square meter in each district annually, measured in Chinese Yuan (RMB) and 1 yuan is about 1.45 SEK (2024-06). The housing price information provided by **JuHuiData** has the advantages of continuous time, complete samples, and accurate data. Population data comes from "The Seventh National Population Census of the People's Republic of China" [16]. And other impact factor data, such as GDP (gross domestic product) for each area in 100 million RMB and the average years of education for people over 15 years old, comes from "*China Statistical Yearbook*" [17].

An overview of part of the data is presented in Table 2.1. Administrative divisions from the government are uneven, but most regions have similar sizes (about 11000 km^2). During data cleaning, we removed some special areas, such as island groups and very small areas. A rough comparison shows that the housing price distribution in the PRD is more even (the range is smaller) while the data in the YRD has a large range. Hence, it is reasonable that the YRD will show more significant spatial agglomeration. The PRD region has a remarkable population density (2-3 times higher than other regions on average) and the BTH has a higher education level, which illustrates the characteristics of these two regions. Taking YRD as an example, the variates shown imply a spatial pattern from the boundary to the center (Shanghai) in Figure 2.2. Spatial patterns in housing prices and possibly covariates suggest that spatial effects should be considered in modeling.

Region		YRD	PRD	BTH
		n=303	n=50	n=127
price(RMB/m^2)	min	2670	5133	4133
	mean	10792	14251	15655
	max	112819	107078	128952
area(km^2)	min	20	22.8	9.98
	mean	1177	1074.4	1248.06
	max	4452	3561.0	9037.00
GDP(10^8 RMB)	min	13.80	125.10	22.30
	mean	554.60	993.80	568.80
	max	16013.40	60122.20	10200.00
pd(ppl/km^2)	min	56.33	137.6	37.3
	mean	1818.75	5562.8	2359.0
	max	33101.50	45554.5	35571.1
eduy	min	7.59	8.24	8.04
	mean	9.12	10.25	10.74
	max	12.93	12.65	13.26

Table 2.1: An overview of part of the data in three regions. ”pd” is the population density in ppl/km^2 (people per square kilometer). ”eduy” is the average years of education for people over 15 years old.

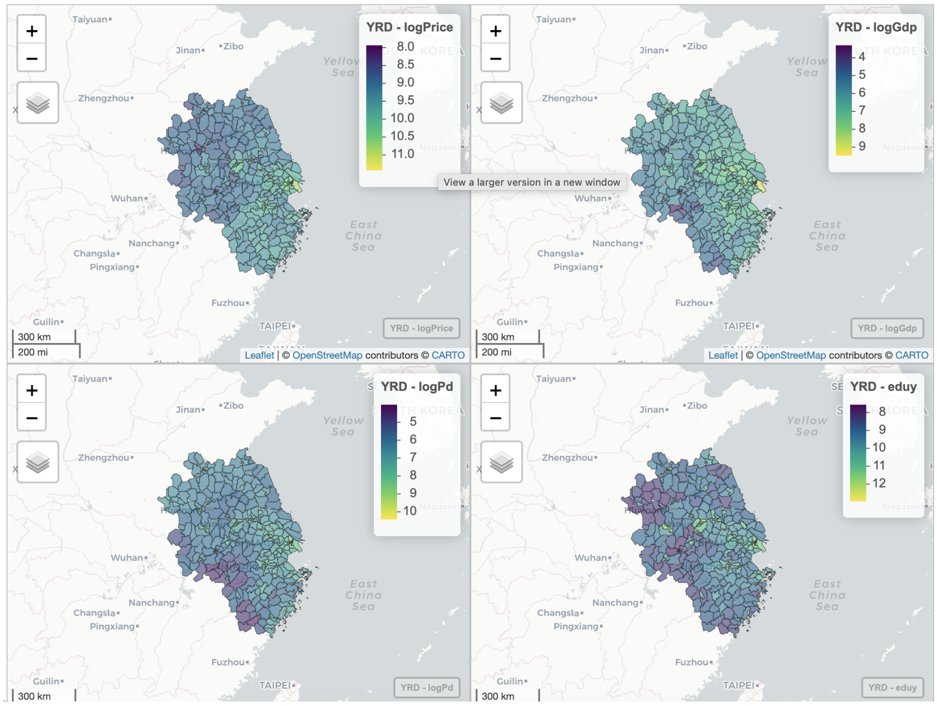


Figure 2.2: The spatial pattern of the log-transformed housing prices (YRD-logPrice), GDP (YRD-logGdp) and population density (YRD-logPd) and the average years of education for people over 15 years old (YRD-eduy) in the YRD in 2020.

3 Model

Section 3.1 explains why the lognormal distribution should be used for the housing prices and the log-transformed housing prices follow a normal distribution. Then a spatial effect is added to the Hedonic housing price model in Section 3.2. Further theory on spatial model is explained in Section 3.3. The general model and the joint model for housing prices are specified in Section 3.4.

3.1 Log-normal distribution for housing prices

Housing prices often do not follow a simple, single distribution due to the complexity and variability of real estate markets. Housing prices are commonly right-skewed (positively skewed), which means there are a larger number of houses priced below the median, with a long tail of higher-priced properties. Therefore, log-normal distribution, Gamma distribution, and generalized beta distribution are often used to analyze the raw data. Sometimes, Weibull distributions are also used for housing prices if the data is more severely skewed. Specially, the log-normal distribution is the most popular. The log-normal distribution is defined for positive real numbers. This property makes it suitable for modeling variables that cannot take negative values, such as prices, incomes, and sizes. Also, the log-normal distribution is positively skewed and its log-transformation follows a normal (bell-shaped) distribution.

The housing prices collected in the thesis fulfill a log-normal distribution, and have the following probability density function

$$f(Y_i|\mu_i, \sigma) = \frac{1}{Y_i \sigma \sqrt{2\pi}} e^{-\frac{(\log Y_i - \mu_i)^2}{2\sigma^2}} \tag{3.1}$$

Where $\log Y_i \sim N(\mu_i, \sigma^2)$ and Y_i has mean $E[Y_i] = e^{\mu_i + \sigma^2/2}$ and variance $Var[Y_i] = (e^{\sigma^2 - 1})e^{2\mu_i + \sigma^2}$. Additionally, the geometric mean and median are both equal to e^{μ}.

However, many predictors do not work ideally under these distributions. To avoid this problem, mathematicians noted that log-transformed housing prices data follows a Gaussian distribution [18]. In this case, $y_i = \log Y_i$, is normal with mean μ_i and variance σ^2. In the following thesis, we will use log-transformed housing prices, y_i, to build a model such that the model will be conditioned on Gaussian distributions.

3.2 Modeling for housing prices

Based on the Hedonic housing price model in Section 1.2, the linear additive predictor will be on the following form

$$\begin{aligned}\eta_{it} &= \underbrace{\beta_0 + \boldsymbol{z}_{it}\boldsymbol{\beta}}_{\text{fixed effect}} + \underbrace{u_i}_{\text{spatial effect}} + \underbrace{v_i}_{\text{i.i.d. effect}} \\ y_{it}|\eta_{it} &\sim N(\eta_{it}, \tau^{-1})\end{aligned} \tag{3.2}$$

Here, $\boldsymbol{z}$ is a (suitable) collection of underlying covariates in each area and for each year, $\boldsymbol{\beta}$ are parameters, u_i is a correlated spatial effect, v_i is an unstructured spatial effect, and y_{it} is the log-transformed housing prices. Note that time is included in the factors $\boldsymbol{z}$ and observations $\boldsymbol{y}$. And we have $E[y_{it}|\eta_{it}] = \eta_{it}$.

When working with spatial data it is important to account for a possible spatial trend in the model to avoid biases in the estimates. In this setting, the Bayesian approach for inference is particularly effective [19], given that we can solve the inference problem in a feasible amount of time. One of the main challenges in Bayesian inference for spatial models is computational, given the added complexity due to the spatial structure.

3.3 Spatial models

3.3.1 Gaussian Markov random fields

Let the neighbours $\mathcal{N}_i$ to a point n_i be the points $\{n_j, j \in \mathcal{N}_i\}$ that are close to n_i. A Gaussian random field $x \sim N(\mu, \Sigma)$ that satisfies

$$p(x_i|\{x_j : j \neq i\}) = p(x_i|\{x_j : j \in \mathcal{N}_i\}) \tag{3.3}$$

is called a Gaussian Markov random fields (GMRF). The simplest example of a GMRF is the AR(1)-process $x_t = ax_{t-1} + \epsilon_t$, where a is a parameter and $\epsilon_t \sim N(0, \sigma^2)$ and independent. However, the AR(1) model is primarily used in time series analysis to capture temporal dependencies. To capture the conditional dependence on neighboring locations, the CAR(1) model is used in spatial analysis [20]. Mathematically, the CAR(1) model is defined as:

$$X_i|X_{-i} \sim N\left(\sum_{j \in \mathcal{N}_i} \rho_{ij} X_j, \tau^2\right)$$

where X_i is the value at location i; X_{-i} represents the values at all other locations except i; $\boldsymbol{N}(i)$ denotes the set of neighbors of location i; ρ_{ij} are the spatial autoregressive parameters; τ^2 is the conditional variance.

The Markov property, combined with the Gaussian assumption, results in the linear predictor $\boldsymbol{\eta}$ forming a Gaussian Markov Random Field (GMRF). A notable characteristic of the GMRF is that its precision matrix (the inverse of the covariance matrix) is sparse. This sparsity provides significant computational advantages during the inference process. The general theory on GMRFs and the associated computational benefits is given by Rue and Held [21].

3.3.2 Spatial referenced data

Available house prices data provided by *JuhuiData* in combination with the administrative geographical division of China yields spatially referenced data. Such that data is characterized by each of the N areas in the analysis window being associated with one observation, i.e. one average price. This spatial data can be viewed as realizations of a (possibly multivariate) stochastic field

$$\boldsymbol{y} := \{y(x), x \in \mathcal{D}\} \tag{3.4}$$

where $\mathcal{D}$ is a fixed countable subset of $\mathcal{R}^2$, and x is the coordinates giving the location of each area. Here $\mathcal{D}$ will be the division into administrative districts provided by the Chinese government. When constructing spatial models using areal data the spatial dependency is modeled through the neighbourhood structure of $\mathcal{D}$. The neighbours can be defined through $\mathcal{N}_i$ as the set of all areas which share borders to area i, as illustrated in Figure 3.1. These are called the first-order neighbours. It would also be possible to consider $\mathcal{N}_i$ as the set of second-order neighbours, i.e. all the areas that share borders with it (first-order) plus the areas which share borders with the first-order neighbours (second-order). But in fact, second-order neighbours are not a common method. Hence, we can obtain the adjacency matrix W with $w_{ij} = 1$ if i and j are neighbours and $w_{ij} = 0$ if $i = j$ or i and j are not neighbours. The spatially correlated random effects can be simply defined by the adjacency matrix W, taking $Q = I - \rho W$, where I is the identity matrix and ρ is a spatial autocorrelation parameter. Hence, the precision matrix τQ of the linear predictor $\boldsymbol{\eta}$ can be computed where τ is a precision hyperparameter. Note that the elements in Q are

$$Q_{ij} = \begin{cases} 1, & i = j, \\ -\rho, & i \in \mathcal{N}_j \\ 0, & otherwise \end{cases} \tag{3.5}$$

3.3.3 Conditional autoregression

To begin, consider first the vector of all linear predictors $\boldsymbol{\eta} = [\eta_1 \cdots \eta_N]^T$ as a random vector, which indeed is the case in the Bayesian framework. If each component were located at a time point rather than a spatial point - under a Markov assumption - the joint density of $\boldsymbol{\eta}$ could be decomposed as

$$\pi(\boldsymbol{\eta}) = \pi(\eta_1) \times \pi(\eta_2|\eta_1) \times \cdots \times \pi(\eta_N|\eta_{N-1}) \tag{3.6}$$

where $\pi(\cdot)$ generically denotes the probability density function of its argument. An intuitive model specification would be to specify the conditional marginal distributions $\pi(\eta_i|\eta_{i-1})$ for $i = 2, ..., N$. Since the assumption is first-order Markov, this compares to an AR(1) model in the standard time series analysis [22]. However, this is not a useful specification in the spatial setting. Nevertheless, using the intuition of specifying the conditional distributions, it is natural to define the conditional distribution $\pi(\eta_i|\eta_{-i})$, where η_{-i} denotes all elements of $\boldsymbol{\eta}$ except η_i. The Markov assumption in the spatial setting refers to the property that η_i should only depend on a few components of η_{-i},

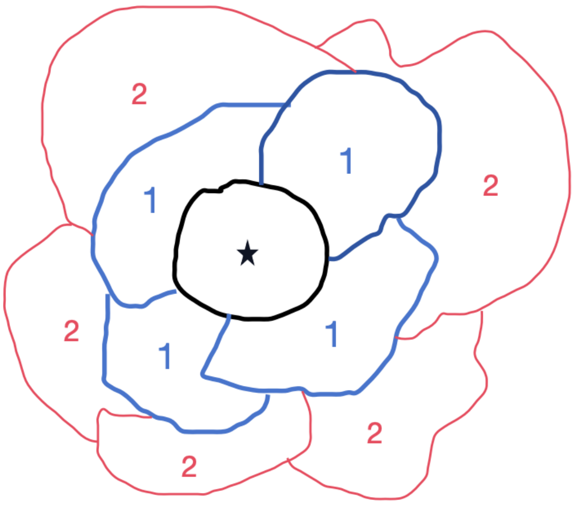


Figure 3.1: Neighbourhood structure. The black area with a star represents the current location; the blue areas with number 1 represents the first-order neighbour; the red areas with number 2 represents the second-order neighbour.

namely the set of near neighbours $\mathcal{N}_i$. If the model is specified as conditional on only the set of first-order neighbours it is referred to as a conditional auto-regressive model of order one (CAR(1)). One could also consider a CAR(2) model, which extends the Markovian property to the second-order neighbours. These model specifications are thoroughly discussed in Waller and Gelfand et al. [22]. The precision mentioned in the previous section represents a CAR(1) model with conditional probability

$$E(\eta_i|\eta_{-i}) = E(\eta_i|\eta_j, j \in \mathcal{N}_i) = \sum_{j\in\mathcal{N}_i} \rho\eta_j$$

.

3.4 Model specification

Let $\pi(y_i|\eta_i, \tau)$ denote the observation likelihood for the i:th observation conditioned on the linear predictor η_i. Recall that $\mathcal{N}_i$ denotes the set of neighbours to area i and let $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2. The assumed model will be on the following form.

$$\begin{aligned} y_{it}|\eta_{it} &\sim \pi(y_{it}|\eta_{it}, \tau) \\ \eta_{it} &= \beta_0 + \boldsymbol{z_{it}\beta} + u_i + v_i \\ \boldsymbol{u} &\sim N(0, \tau_u^{-1}Q_u^{-1}(\rho)) \\ v_i &\sim N(0, \tau_v^{-1}) \end{aligned} \tag{3.7}$$

$\pi(y_{it}|\eta_{it})$ is a Gaussian distribution with mean η_{it} and precision τ, see Equation 3.2. The Gaussian part in a latent Gaussian model stems from that Gaussian priors are assigned to the vector of parameters in the predictor η_{it}. The unstructured effects $\boldsymbol{v} =$

$[v_1, ..., v_N]^T$ models additional random variation, independent of geographic location. It would be possible to exclude $\boldsymbol{v}$ from the model If the data is highly correlated in space. The larger the effect of $\boldsymbol{v}$ compared to $\boldsymbol{u}$ is in the model, the less exchange of information between areas is allowed. As seen, both u_i and v_i are specified as Gaussian, with variance determined by the hyperparameters ρ, τ_u and τ_v. A common choice of prior distributions of the hyperparameters is a Gamma distribution [19].

3.4.1 Priors

By default, the intercept has a Gaussian prior with mean equal to zero. Coefficients of the fixed effects also have a Gaussian prior by default with zero mean and precision equal to 0.001. The spatial autocorrelation parameter ρ can be calculated using Moran's I. More details of Moran's I will be introduced in Section 4.6.4. The prior on the precision of the error term (τ, τ_u and τ_v) is, by default, a Gamma distribution with parameters 1 and 0.00005 (shape and rate, respectively). Theoretically, hyperparameters have no impact on the performance of the model, but will affect the speed and quality of learning. Given the absence of prior information in this scenario, opting for a relatively flat prior is a sensible choice. Such a choice allows the data to exert a stronger influence on the inference process, effectively letting the observed data "speak for themselves." It is also the default in the R-INLA package.

3.4.2 BESAG model

The Besag model, introduced by Julian Besag, is designed to capture spatial correlations, particularly in fields like geography and epidemiology. The Besag model is a type of Conditional Autoregressive (CAR) model, which defines the conditional distribution of each area's value based on its neighboring areas' values. Specifically, the Besag model assumes that the random effect of each area follows a normal distribution with a mean that is the average of the neighboring areas' random effects and a variance that depends on the number of neighboring areas. The besag model for random vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ is defined as

$$x_i | x_j, i \neq j, \tau \sim N(\frac{1}{n} \sum_{i\ j} x_j, \frac{1}{n_i \tau})$$

where n_i is the number of neighbours of node i, $i \sim j$ indicates that the two nodes i and j are neighbours and the hyperparameters τ is the precision of $\boldsymbol{x}$.

3.4.3 BYM model

The BYM model (Besag-York-Mollié model) for random vector $\boldsymbol{x}$ is simply a union of the besag model (spatial effect) $\boldsymbol{u}$ and a iid model (unstructured effect) $\boldsymbol{v}$, so that

$$x = \begin{pmatrix} u + v \\ u \end{pmatrix}$$

The hyperparameters are the precision τ_u of the besag mode u and τ_v of the iid model v. The benefite is that this allows to get the posterior marginals of the sum of the spatial and iid model.

3.4.4 Joint model

Sometimes it is necessary to share an effect that is estimated from two or more parts of the dataset, so that all of them provide information about the effect when fitting the model. This is known as a *copy effect*, as the new effect will be a copy of the original effect plus some tiny noise.

Formally, the copy feature is used when a latent field is needed more than once in the model formulation [23]. When using the feature we then create a (almost) identical copy of $\boldsymbol{Z}$, denoted by $\boldsymbol{Z}^*$, that can then be used in the model formulation. In this case, latent field can be extended from $\boldsymbol{Z}$ to $\boldsymbol{Z}^C = (\boldsymbol{Z}, \boldsymbol{Z}^*)$, where $\pi(\boldsymbol{Z}^C) = \pi(\boldsymbol{Z})\pi(\boldsymbol{Z}^*|\boldsymbol{Z})$ and

$$\pi(\boldsymbol{Z}^*|\boldsymbol{Z}, \tau_Z, \lambda) \propto \exp(-\frac{\tau_Z}{2}(\boldsymbol{Z}^* - \lambda\boldsymbol{Z})^T(\boldsymbol{Z}^* - \lambda\boldsymbol{Z})) \tag{3.8}$$

where λ is a scale parameter and a default value of τ_Z is large e.g. $\tau_Z = \exp(15)$ so that the degree of closeness between $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$ is controlled by the fixed high precision τ_Z.

Generally, the population and education level change a little each year for each district at the macro level, but in terms of the economy, the annual changes are more obvious. Hence, temporal effects can be considered for extraction from GDP. And the assumed layered model will be on the following form.

$$\begin{aligned}
y_{it}|\eta_{it}^P &\sim N(\eta_{it}^P, \tau_P^{-1}) \\
\log \mathrm{GDP}_{it} &\sim N(\eta_{it}^G, \tau_G^{-1}) \\
\eta_{it}^G &\sim AR(1) \\
\eta_{it}^P &= \beta_0 + \beta^*\eta_{it}^G + \boldsymbol{z}_{it}^{-G}\boldsymbol{\beta} + u_i + v_i \\
u_i &\sim \mathrm{CAR}(1) \\
v_i &\sim \text{i.i.d.}
\end{aligned} \tag{3.9}$$

Here, η_{it}^G is the copied effect from $\log \mathbf{GDP}$ and $\boldsymbol{z}_{it}^{-G}$ is all the covariates except GDP. Note the copied effect has a scaling factor, β^*, and this is fixed to 1 by default. Furthermore, the precision of η_{it}^G is set to a very large value, ensuring that the copied effect is very close to $\log \mathbf{GDP}$ [24].

4 Inference

A brief introduction to Bayesian inference and the conditions is given in Section 4.2. Then the inference of Integrated Nested Laplace Approximation (INLA) [25] is described in Section 4.5. In Section 4.6, methods for solving the model selection and validation problem are presented.

4.1 The R-INLA package

R-INLA is the R package that implements approximate Bayesian inference using integrated nested Laplace approximation [26]. It is a rich package, which enables a lot of models to be specified. It is available at `http://www.r-inla.org`.

An assumption required by **R-INLA** is that all observations y_i should be independent conditional on the latent field $\boldsymbol{\eta}$. This conditional independence is crucial for simplifying the computations and ensuring the efficiency of the inference procedures. The joint likelihood of the observations can be decomposed into a product of individual likelihoods. This decomposition simplifies the computation of the likelihood function, making the model easier to handle computationally.

4.2 Bayesian Inference

In Bayesian inference, all unknown parameters in the model are treated as random variables. The aim is to compute or estimate the joint posterior distribution, which represents the distribution of the parameters ψ conditional on the observed data y. From Equation 3.7, the latent Gaussian field is

$$\begin{aligned} \boldsymbol{x} &= [\quad \beta_0 \quad \boldsymbol{\beta}^T \quad \boldsymbol{u}^T \quad \boldsymbol{v}^T]^T \\ A &= [1 \quad \boldsymbol{z} \quad I \quad I] \\ \boldsymbol{\eta} &= \beta_0 + \boldsymbol{z}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{v} = A\boldsymbol{x} \\ \boldsymbol{y}|\boldsymbol{\eta} &\sim N(A\boldsymbol{x}, \tau^{-1}) \end{aligned} \tag{4.1}$$

with $\dim(\boldsymbol{x}) = n$ and I is the identity matrix. Since $\boldsymbol{x}$ is now a joint Gaussian density with block precision matrix, we can rewrite the model as

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{x} &\sim \prod \pi(y_i|\eta_i, \boldsymbol{\psi}) \\ \boldsymbol{x}|\boldsymbol{\psi} &\sim N(0, Q^{-1}(\boldsymbol{\psi})) \\ \boldsymbol{\psi} &\sim \pi(\boldsymbol{\psi}) \end{aligned} \tag{4.2}$$

where $\boldsymbol{\psi}$ represents all hyperparameters $[\tau, \tau_u, \tau_v]$.

Observations are assumed to be independent, conditioned on the vector of latent effects and the hyperparameters. By assumption of the model, we already have the distributions $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})$, $\pi(\boldsymbol{x}|\boldsymbol{\psi})$ and $\pi(\boldsymbol{\psi})$. The main aim now is to compute the posterior marginal distributions

$$\pi(\boldsymbol{x},\boldsymbol{\psi}|\boldsymbol{y}), \quad \pi(x_i|\boldsymbol{y}), \quad \pi(\psi_i|\boldsymbol{y}) \tag{4.3}$$

allowing us to estimate parameters and the latent fields.

4.3 Advantages of INLA

A typical strategy for approximating the posterior distributions in Equation (4.3) is the Markov chain Monte Carlo (MCMC) method. However, in spatial contexts, particularly when dealing with large spatial fields $\boldsymbol{u}$ and potentially numerous parameters in $\boldsymbol{\beta}$ resulting in a high dimension, MCMC methods often fail to deliver adequately precise approximations within a feasible computational timeframe. Rue et al.[25] develop the integrated nested Laplace approximation (INLA) for approximate Bayesian inference as an alternative to traditional Markov chain Monte Carlo methods, thereby solving the following two problems: the latent field $\boldsymbol{z}$ are strongly dependent on each other and $\boldsymbol{\beta}$ and $\boldsymbol{z}$ are also strongly dependent, especially when the size of data is large. And the Metropolis-Hastings algorithm and Gibbs sampling also play an important role in Bayesian inference. INLA prioritizes models that can be represented as latent Gaussian Markov random fields (GMRF) due to their favorable computational characteristics [21]. Specifically, the focus lies on estimating the posterior marginals of the model parameters. Consequently, instead of tackling the highly multivariate joint posterior distribution $\pi(\boldsymbol{\psi}|\boldsymbol{y})$, the emphasis shifts towards obtaining approximations for univariate posterior distributions $\pi(\psi_i|\boldsymbol{y})$. This approach enables the algorithm to generate approximations with acceptable error using significantly fewer computations, thereby facilitating sufficiently rapid approximations in the case of GMRFs.

4.4 The joint posterior distribution

The joint posterior distribution of the effects and hyperparameters $\boldsymbol{x}$, $\boldsymbol{\psi}|\boldsymbol{y}$ can be expressed as

$$\begin{aligned}
\pi(\boldsymbol{x},\boldsymbol{\psi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\psi}) \cdot \pi(\boldsymbol{x}|\boldsymbol{\psi}) \cdot \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi}) \\
&\propto \pi(\boldsymbol{\psi}) \cdot \pi(\boldsymbol{x}|\boldsymbol{\psi}) \cdot \prod_{i=1}^{N} \pi(y_i|\boldsymbol{x},\boldsymbol{\psi}) \\
&\propto \pi(\boldsymbol{\psi}) \cdot |Q(\boldsymbol{\psi})|^{1/2} \exp(-\frac{1}{2}\boldsymbol{x}^T Q(\boldsymbol{\psi})\boldsymbol{x}) \cdot \prod_{i=1}^{N} \exp(\log(\pi(y_i|\boldsymbol{x},\boldsymbol{\psi}))) \\
&\propto \pi(\boldsymbol{\psi}) \cdot |Q(\boldsymbol{\psi})|^{1/2} \exp(-\frac{1}{2}\boldsymbol{x}^T Q(\boldsymbol{\psi})\boldsymbol{x} + \sum_{i=1}^{N} \log(\pi(y_i|\boldsymbol{x},\boldsymbol{\psi})))
\end{aligned} \tag{4.4}$$

4.5 Integrated Nested Laplace Approximation

4.5.1 Nested Laplace approximation

By Bayes' theorem, the posterior of the latent effects is $\pi(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})\pi(\boldsymbol{x}|\boldsymbol{\psi})$, and it follows that

$$\log\pi(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{y}) = \log\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi}) + \log\pi(\boldsymbol{x}|\boldsymbol{\psi}) + c \tag{4.5}$$

where c is a constant. Now, since $\boldsymbol{x}|\boldsymbol{\psi}$ is Gaussian by definition, taking the logarithm yields a second order polynomial in x. Furthermore, using a second order Taylor expansion of $\log\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})$, it appears that $\log\pi(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{y})$ can be approximated by a second order polynomial for the case of Gaussian observations $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})$ is already second order. If $\log\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\psi})$ is a second order polynomial, then $\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\psi}$ is Gaussian. This is referred to as the Laplace approximation. Let $N(x;\mu,\sigma^2)$ denote the Gaussian density with mean μ and variance σ^2 at configuration x. Also, let $\pi_{LA}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\psi})$ denote the Laplace approximation of $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\psi})$. Then specifically,

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\psi}) \approx \pi_{LA}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\psi}) = N(\boldsymbol{x};\boldsymbol{x}^*,H(\boldsymbol{x}^*)) \tag{4.6}$$

due to the uniqueness of the Taylor expansion. Here, x^* denotes the mode of $\log\pi(\boldsymbol{x^*}|\boldsymbol{\psi},\boldsymbol{y})$ or equivalently $\boldsymbol{x}^* = \arg\max_x \pi(\boldsymbol{x^*}|\boldsymbol{\psi},\boldsymbol{y})$, and $H(\boldsymbol{x}^*)$ is the Hessian matrix of second derivatives of the log posterior evaluated at x^*, given by

$$\begin{aligned} H(\boldsymbol{x}^*) &= -[\frac{\partial^2\log\pi(\boldsymbol{x^*}|\boldsymbol{\psi},\boldsymbol{y})}{\partial\boldsymbol{x}^{*2}}]^{-1} \\ &= Q - -[\frac{\partial^2\log\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})}{\partial\boldsymbol{x}^2}], \end{aligned} \tag{4.7}$$

where we have used the Gaussianity of $\pi(\boldsymbol{x}|\boldsymbol{\psi})$ in Equation 4.5 to get Q.

As usual, the approximation is most accurate at the mode. Laplace approximations provide a computationally efficient way to approximate the posterior distribution, especially when analytical methods are not feasible. However, it relies on the assumption that the posterior distribution is approximately Gaussian, which may not hold in all cases, particularly for multimodal or highly skewed distributions. Therefore, it's important to assess the adequacy of the approximation, possibly through sensitivity analysis or comparison with alternative methods.

Then, the approximation of the full marginal posterior $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ is optimized as follows.

$$\begin{aligned} \pi(\boldsymbol{\psi}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\psi},\boldsymbol{y})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{x},\boldsymbol{\psi}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{y})} \\ &\propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})\pi(\boldsymbol{x}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\pi(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{y})} \end{aligned} \tag{4.8}$$

where $\pi(\boldsymbol{x},\boldsymbol{\psi}|\boldsymbol{y})$ is shown in Section 4.4 and $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\psi})$, and $\pi(\boldsymbol{x}|\boldsymbol{\psi})$and $\pi(\boldsymbol{\psi})$ are the assumption of the model. For a given $\boldsymbol{\psi}$, it is possible to obtain an approximation by replacing the denominator with its Laplace approximation from Equation (4.6), and evaluate it at its most accurate point, namely the mode $\boldsymbol{x}^*(\boldsymbol{\psi}) = argmax_x\pi(\boldsymbol{x}|\boldsymbol{\psi},\boldsymbol{y})$. Note that the mode depends on the hyperparameters. This gives the approximation

$$\pi(\boldsymbol{\psi}|\boldsymbol{y}) \approx \tilde{\pi}_{LA}(\boldsymbol{\psi},\boldsymbol{y}) = \frac{\pi(y|\boldsymbol{x},\boldsymbol{\psi})\pi(x|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\pi_{LA}(x|\boldsymbol{\psi},\boldsymbol{y})} \tag{4.9}$$

4.5.2 Integrated nested Laplace approximation

The Laplace approximation is performed around the mode of the posterior distribution. This means that it may fail to capture the global characteristics of the posterior distribution, especially in cases of multimodal or highly skewed distributions. The computational complexity of the Laplace approximation increases rapidly in high-dimensional parameter spaces. When dealing with large parameter spaces, it may result in inaccurate approximations or computational difficulties. Under the assumption of independent observations, INLA addresses these drawbacks by using efficient numerical integration techniques, allowing INLA to infer over a broader range of parameter space, and improving the accuracy of the approximation.

As mentioned earlier, INLA does not attempt to estimate the full joint posterior distribution but focuses on computing the marginal distributions of the latent effects and hyperparameters. The computation of the marginal distributions for the latent effects and hyperparameters can be accomplished by considering that

$$\pi(\psi_i|\boldsymbol{y}) = \int \pi(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}_{-i} \approx \int \tilde{\pi}_{LA}(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}_{-i} \tag{4.10}$$

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\psi}, \boldsymbol{y})\pi(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi} \approx \int \tilde{\pi}_{LA}(x_i|\boldsymbol{\psi}, \boldsymbol{y})\tilde{\pi}_{LA}(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi} \tag{4.11}$$

where $\tilde{\pi}_{LA}(\cdot)$ is the Laplace approximation as above.

The expressions in Equation (4.10) and (4.11) suggest that the algorithm needs to proceed in two steps. First a step where $\pi(\boldsymbol{\psi}|\boldsymbol{y})$ is approximated and then a second where $\pi(x_i|\boldsymbol{\psi}, \boldsymbol{y})$ is approximated. Note how in both expressions integration is done over the space of the hyperparameters and that a good approximation to the joint posterior distribution of the hyperparameters is required. Rue, Martino, and Chopin [25] use the Laplace approximation to compute the posterior marginal of the latent parameter x_i as:

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\psi_k, \boldsymbol{y}) \cdot \tilde{\pi}(\psi_k|\boldsymbol{y}) \cdot \Delta_k \tag{4.12}$$

Here, Δ_k are the weights associated with a vector of values ψ_k of the hyperparameters on a grid. This method transforms complex high-dimensional calculations into numerical integration.

4.6 Model selection and validation

4.6.1 Prediction errors

A common approach to evaluate a model's performance is by examining its predictive accuracy. Initially, a subset of the data is reserved as a validation set before any estimation is performed. The remaining data is then used for estimation. Let Y_i represent an observation in the prediction set, and define the prediction error as $\epsilon_i = Y_i - E[Y_i]$. It can also be beneficial to examine the normalized prediction errors, defined as $\epsilon_i^{\text{norm}} = \epsilon_i / E[Y_i]$. This allows for a comparison of models based on prediction quality, for instance, using the Mean Squared Error (MSE), which is defined as:

$$MSE = \frac{1}{N} \sum_i \epsilon_i^2 \tag{4.13}$$

where N is the number of observations in the validation set. This method provides a clear metric for assessing and comparing the predictive performance of different models. Moreover, the prediction errors should show no spatial pattern, indicating that the spatial structure has been successfully captured by the model.

4.6.2 Deviance information criteria

If it is interesting to compare the performance of different models, the deviance can be used [19]. The deviance is defined as

$$D(\boldsymbol{\theta}) = -2 \log \pi(\boldsymbol{y}|\boldsymbol{\theta}) \tag{4.14}$$

where, as usual, $\boldsymbol{\theta}$ identifies the parameters of the likelihood, i.e. $\boldsymbol{\psi}$ in the model and y is the observed data. In the Bayesian framework $\boldsymbol{\theta}$ is a random variable, and hence also $D(\boldsymbol{\theta})$. Typically, the posterior mean deviance

$$\bar{D} = E[D(\boldsymbol{\theta})] \tag{4.15}$$

is used to quantify the deviance as a measure of fit. However, with an increasing number of parameters, the fit will be better and hence $(\bar{D})$ smaller. Therefore it is necessary to introduce a penalty against $\bar{D}$ which reflects the model complexity. Such a measure was proposed by Spiegelhalter et al. [27] as the Deviance information criteria (DIC). It is a generalization of the well-known Akaike information criteria (AIC). The DIC is a sum of two components, first the deviance $\bar{D}$ which measures the model fit, and second the effective number of parameters. The effective number of parameters p_D is a measure of model complexity, defined as the difference between the deviance evaluated at the posterior mean of the parameters $\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta}]$ and the posterior mean of the deviance, defined as

$$p_D = \bar{D} - D(\hat{\boldsymbol{\theta}}) \tag{4.16}$$

Finally, DIC is defined as the sum of the posterior mean of the deviance and the effective number of parameters:

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\hat{\boldsymbol{\theta}}) \tag{4.17}$$

Here $\bar{D}$ decreases with better model fit, and p_D increases with added complexity. DIC is used to evaluate and compare the goodness of fit of statistical models while penalizing for model complexity. It aims to balance model fit and complexity, preferring models that provide a good fit to the data without being overly complex.

4.6.3 WAIC

In statistics, the widely applicable information criterion (WAIC)[28], also known as Watanabe–Akaike information criterion, is the generalized version of the Akaike information criterion (AIC) onto singular statistical models. Like DIC, WAIC also has two components. The first part is log pointwise predictive density (LPPD), which is computed as

$$LPPD = \sum_{i=1}^{N} \log E[p(y_i|\boldsymbol{\theta})] \tag{4.18}$$

And the penalty term of WAIC is fully Bayesian and can be expressed as

$$p_{WAIC} = \sum_{i=1}^{N} Var[\log p(y_i|\boldsymbol{\theta})] \tag{4.19}$$

Finally, WAIC is defined as

$$WAIC = -2(LPPD - p_{WAIC}) \tag{4.20}$$

Gelman et al. [29] recommends WAIC because its results are closer in practice to the results of leave-one-out cross-validation (LOO-CV). Also, a lower value of the WAIC indicates that the model is better.

LOO differs from the aforementioned information criterion-based indices in that its computation requires no penalty term. Specifically, LOO is computed as

$$LOO = -2LPPD_{loo} = -2\sum_{i=1}^{N} \log \int p(y_i|\boldsymbol{\theta})p_{-i}(\boldsymbol{\theta})d\theta \tag{4.21}$$

where p_{-i} is the posterior distribution based on the data minus data point i. LOO only focuses on prediction. It provides an unbiased estimate of the model's predictive performance and helps to assess how well the model generalizes to new, unseen data. However, it requires rebuilding the model for each data point, making it computationally intensive and the performance estimate can have high variance, especially with small datasets, which can lead to unreliable estimates of model performance.

4.6.4 Moran's I test

In statistics, Moran's I is a measure of spatial autocorrelation developed by Patrick Alfred Pierce Moran [30]. Spatial autocorrelation is characterized by a correlation in a signal among nearby locations in space. Global Moran's I is a measure of the overall clustering of the spatial data, defined as

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4.22}$$

where N is a number of spatial units indexed by i and j, x is the variable of interest with mean $\bar{x}$ like the residual of the model, w_{ij} is the elements of a matrix of spatial weights with zeroes on the diagonal, and W is the sum of all spatial weights.

Note that "Moran's I = 0" indicates no spatial autocorrelation, suggesting a random spatial pattern. But it is not rigorous to judge spatial autocorrelation only from the numerical values of Moran's I. To determine if the observed Moran's I is significantly different from what would be expected under a null hypothesis of no spatial autocorrelation, it is often standardized and compared to a normal distribution.

The value of I can depend quite a bit on the assumptions built into the spatial weights matrix w_{ij}. The matrix is required because, in order to address spatial autocorrelation and also model spatial interaction, we need to impose a structure to constrain the number of neighbors to be considered. This is related to Tobler's first law of geography [9]. The law implies a spatial distance decay function, such that even though all observations have an influence on all other observations, after some distance threshold that influence can be neglected.

5 Results

In Section 5.1, the different types of linear predictors are specified. Then, Section 5.2 shows the simple method to choose covariates. In Section 5.3, the results from fitting housing prices in the Yangtze River Delta are presented. In Section 5.4, the prediction of housing prices will be compared with real values, which can be used to assess the quality of the model. And in Section 5.5, we will compare models among the three regions.

From Section 5.1 to Section 5.4, we will firstly built a spatial statistical model for the Yangtze River Delta region as inspiration for models in other regions.

5.1 The linear predictor

Four models have been fitted to housing prices in the Yangtze River Delta, all of which follow the form outlined in Equation (3.7). The models differ based on the inclusion of the co-variate effect $\boldsymbol{z\beta}$, the spatial effect u_i, and the unstructured effect v_i. The precision parameters follow the default prior distribution as specified by R-INLA. Table 5.1 presents and names these models for future reference.

Model Name	Linear Predictor
GLM	$\eta_{it} = \beta_0 + z_{it}\beta$
IID	$\eta_{it} = \beta_0 + z_{it}\beta + v_i$
BESAG	$\eta_{it} = \beta_0 + z_{it}\beta + u_i$
BYM	$\eta_{it} = \beta_0 + z_{it}\beta + u_i + v_i$

Table 5.1: The different structures in the linear predictor yield four different models

The data comprises 303 areas during four years. A random 20% of the data is selected as a prediction set, leaving 969 observations (calculated as $0.8 \cdot 303 \cdot 4$) for model fitting. During the modeling stages, this is done by setting those values as y_i = NA in R-INLA, which excludes them from contributing to the observation likelihood. However, R-INLA still makes predictions for these points, providing a straightforward method for predictive validation.

5.2 Choice of covariates

It is obvious that housing prices are related to many factors, and the establishment of the model depends on the perspective from which the modeler analyzes the problem. Therefore, we need to select appropriate covariates to explain the model.

5.2.1 Smooth

As assumption, the covariates have a basic linear relationship with the observations. However, the visual effect is not necessarily realistic, and we still have to consider whether there is a nonlinear relationship between them and then transform them to a computable model. Firstly, we can use Box-Cox transformation like log-transformation. Then if necessary, we can use more flexible model to explain the non-linear relationship such as spline, the "autoregressive model (AR)" or "random walk model (RW)". Moreover, numerous statistical applications necessitate a flexible model structure, accommodating non-linear relationships between the response variable and the covariates.

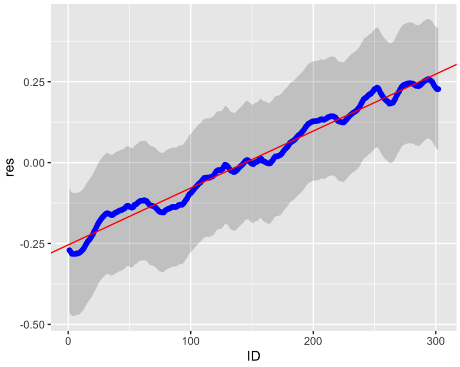


Figure 5.1: The effect of GDP on housing prices in 2019. An AR(1) model used for the relationship between log-transformed GDP and logPrice. The x-axis is the log-transformed GDP (*logGdp*) of each districts in 2019 and the y-axis shows the effect of *logGdp* on log-transformed housing prices (*logPrice*) in 2019.

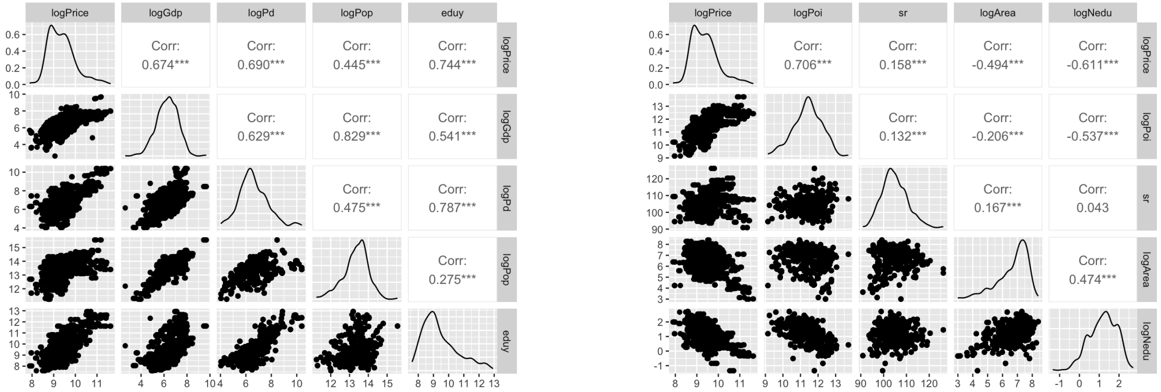
For example, we apply an AR(1) model to a factor log-transformed GDP "*logGdp*" and, using standard INLA formulation of the non-linear effect [31], we set

$$logPrice \sim f(logGdp, model =' ar1') + \boldsymbol{z}^{-\boldsymbol{G}}\boldsymbol{\beta}$$

where $\boldsymbol{z}^{-\boldsymbol{G}}$ is a vector containing all covariates except *logGdp*. We plot the results in Figure 5.1 and find that *logPrice* are linear on the *logGdp* scale, albeit less than perfectly. It implies that we do not need to do any more work to deal with the GDP data in 2019 after log-transformation. Repeat this process, we obtained some possible covariates in Section 5.2.2.

5.2.2 Correlation

Intuitively, because it is a linear predictor, covariates, or their transformations, should have a linear correlation with the variable $\eta_i t$. As above, we use log-transformed housing prices to build a model so that we also need to transform covariates to find a suitable linear relationship. In this case, we apply log-transformation to some covariates. In Figure 5.2 displays that sex ratio (based on females) has the weakest relationship with housing prices the (correlation coefficient is 0.158) while the plots where the covariates are sex ratio (sr), log-transformed population (logPop), area (logArea) and illiteracy rate (logNedu) show nonlinear patterns. Hence, it is reasonable to suspect that these covariates will have a negative impact on the model, but due to their high correlation coefficients, we cannot directly remove them.



(a) The correlation between log-transformed housing prices and education yaear (eduy), log-transformed GDP (logGdp), population density (logPd) and population (logPop).

(b) The correlation between log-transformed housing prices and sex ratio ((based on females)), log-transformed number of points of interest (logPoi), area and illiteracy rate (logNedu).

Figure 5.2: The correlation between log-transformed housing prices and possible covariates.

5.2.3 Confidence Interval

Approximate credibility intervals and p-values can be obtained from the estimated posterior distributions of each β-coefficient. This allows for determining whether all covariates are significant at the 95% level. It is a general method to determine whether a variable is significant. However, because the data are limited, there are fewer covariates to choose from. We can roughly decide whether to use a covariate based on its confidence interval. If its confidence interval contains 0, it is considered not needed and should be deleted from the model. If it is a large model, we can check the significance of all parameters β_i, remove β-coefficients which are not significant anymore, and refit the model. Then compare criteria such as DIC to choose the better one. If a more rigorous approach is desired, there exists a substantial body of literature on variable screening in linear regression, exemplified by Gelman and Hill [32].

Firstly, we use all possible covariates to build a model and find that there are some covariates that are not significant. Then prioritize removal of covariates discussed in the previous section if they are not significant. Repeat this process and the final covariates are shown in Table 5.2. It is worth mentioning that the education level and population density among the selected covariates are less affected by time, while GDP changes significantly every year.

Selected covariates			
Covariate	Estimate	C.I.	Explanation
logGDP	0.168	(0.117,0.218)	log-transformed GDP
logPd	0.121	(0.076,0.166)	log-transformed population density
year	0.044	(0.040,0.048)	4 years from 2019 - 2022
eduy	0.173	(0.132,0.214)	Average year of education for people over 15 years old
Removed covariates			
sr	0.009	(-0.005,0.023)	sex ratio (based on females)
logNedu	-0.033	(-0.075,0.010)	log-transformed illiteracy rate
logPoi	0.195	(-0.002,0.388)	log-transformed number of points of interest
logArea	-0.046	(-35.847,35.754)	log-transformed area of each district
logPop	-0.029	(-35.830,35.771)	log-transformed population

Table 5.2: Significant covariates after checking correlation and confidence interval. 'C.I.' is confidence interval.

5.3 Housing price model

In this section, a housing price model in the Yangtze River Delta will be built. And the common part (GLM part) of the model is given by

$$logPrice_{it} \sim 1 + logGdp_{it} + logPd_i + year_t + eduy_i \tag{5.1}$$

Emphatically, η_{it} is modeled as Gaussian-distributed and the mean is linked to the linear predictor by

$$E[y_{it}|\eta_{it}] = \eta_{it}$$

where y_{it} is the log-transformed housing prices in each area and each year conditioned on η_{it} and η_{it} is the linear predictor outlined in Table 5.1. The MSE using training data, DIC, WAIC, and Moran's I test are reported in Table 5.3. We use Moran's I test to test the residuals to make sure that a model can contain and explain all spatial effects in the model. The GLM and IID model still remain spatial effects in the residuals without structured effect $\boldsymbol{u}$ hence we can not choose these two models.

Model	MSE	DIC	WAIC	Moran's I test
GLM	0.1312	988	988	The residuals contain spatial effects
IID	0.0006627	-2653	-2637	The residuals contain spatial effects
BESAG	0.004330	-2603	-2594	The residuals do **not** contain spatial effects
BYM	**0.0006117**	**-2665**	**-2649**	The residuals do **not** contain spatial effects

Table 5.3: MSE, DIC, WAIC, and Moran's I test from fitting the housing price model.

It is obvious that the models incorporating the unstructured effect v_i (IID and BYM) are preferred in terms of the MSE, DIC, and WAIC. Meanwhile, the BESAG model behaves better than the GLM model, and the BYM model also behaves better than the IID model, which implies that the existence of a spatial effect u_i is reasonable. Besides, the BESAG and BYM model do not have significant spatial correlation in the residuals, which implies they can explain spatial effects in the model perfectly. Thus the BYM model is the best specification for the linear predictor given by $\eta_{it} = \beta_0 + z_{it}\beta + u_i + v_i$.

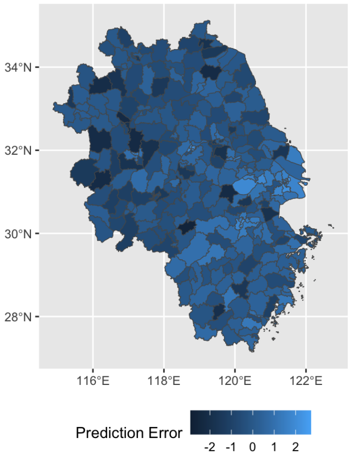


Figure 5.3: Prediction errors plotted according to their coordinates in 2019.

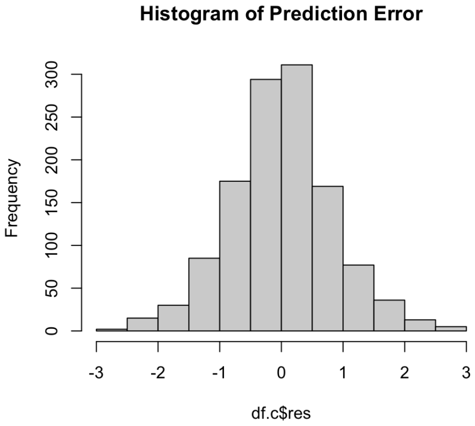


Figure 5.4: Histogram of prediction errors

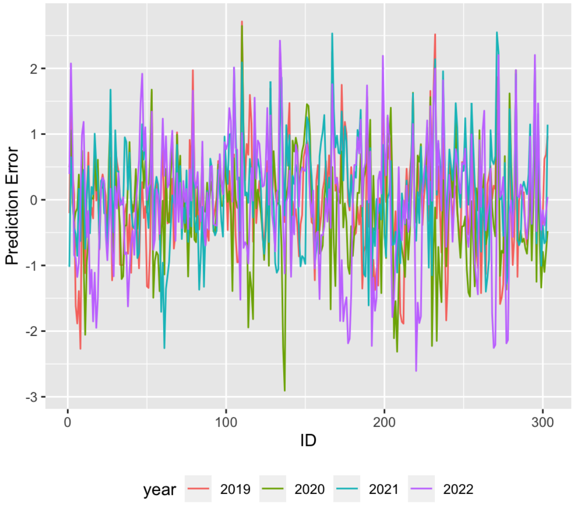


Figure 5.5: Prediction error separated by year and indexed by locations

5.3.1 Residual analysis for the BYM-model

Figure 5.3 depicts the prediction errors plotted based on their geographical location, indicating the absence of any discernible spatial trend in the data. This suggests that the BYM model effectively removes spatial effects from the data. Meanwhile, Figure 5.4 shows that the prediction errors essentially follow a normal distribution with a mean value of 0, consistent with our modeling assumptions. Analyzing the prediction errors from another perspective, in Figure 5.5 the residuals in different years are roughly the same, while the residuals in different regions vary greatly. Therefore, temporal effects have less impact on the model and predictions, and spatial effects play a major role in the model.

5.4 Prediction

The **R-INLA** package provides very convenient linear prediction functions which includes linear prediction and marginal linear prediction as well as their confidence intervals. In Figure 5.6, the upper and lower 95% bounds of the linear predicted values in 2019 definitely contain the true value which implies the BYM model is reasonable. It is straightforward to compute estimates of housing prices in their original scale by transforming the estimates of the logarithm of housing prices. First, use the function **inla.tmarginal** to obtain the marginals of the prices as exp(log(price)). Then, use the function **inla.zmarginal** to obtain the summaries of the marginals such that original scale prediction is obtained.

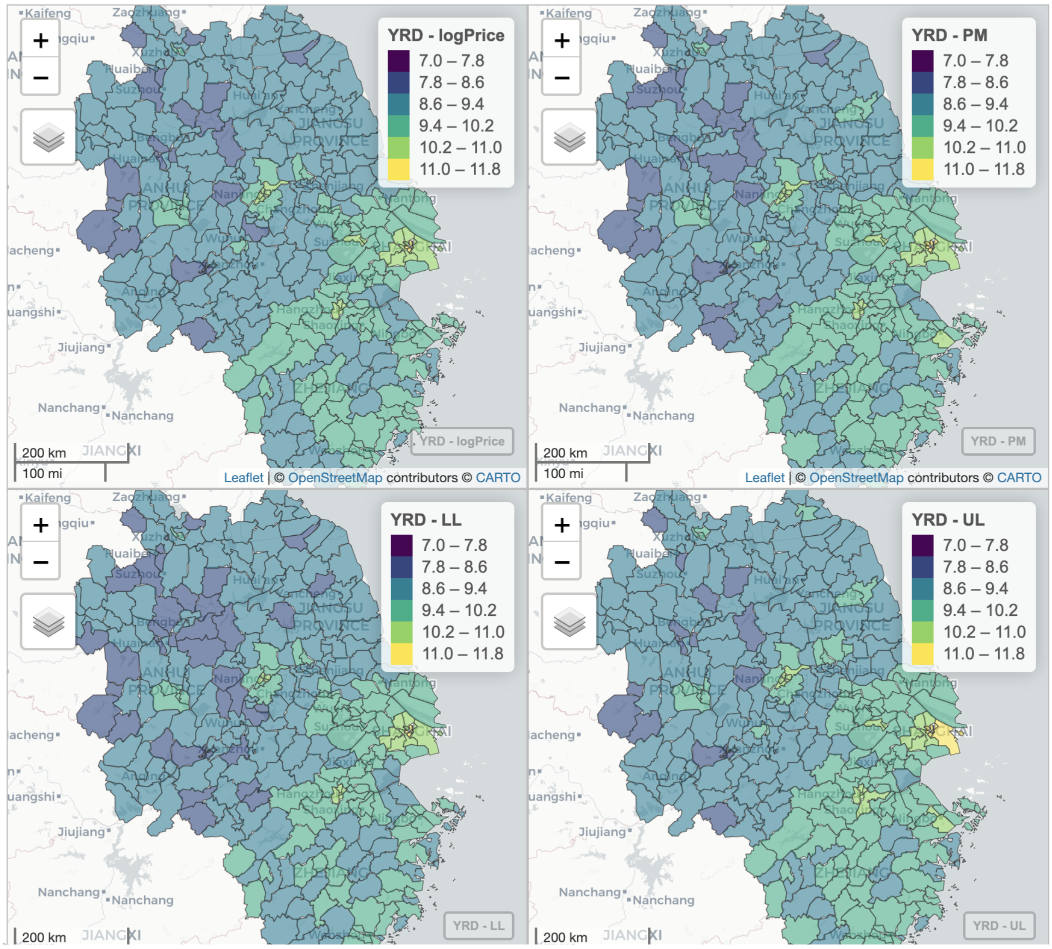


Figure 5.6: Log-transformed housing prices (YRD-logPrice), predictions (YRD-PM), and 95% lower (YRD-LL) and upper (YRD-UL) bounds in 2022.

Then, we check how the training and test sets behave. In Figure 5.7, it shows that the model works well since predictions in both training and test sets are consistent with the real observations. A correlation test between real observations and predictions for the test set obtained a P-value of less than 0.01 and a correlation coefficient of 0.99. However, the correlation test is weak in this case and we prefer to use MSE and RMSE. The MSE and RMSE of the training set are 3.7×10^{-3} and 6.1×10^{-2} while The MSE and RMSE of the test set are 7.0×10^{-3} and 8.4×10^{-2}. The very small MSE and RSE of the training and test set support that the BYM model performs well.

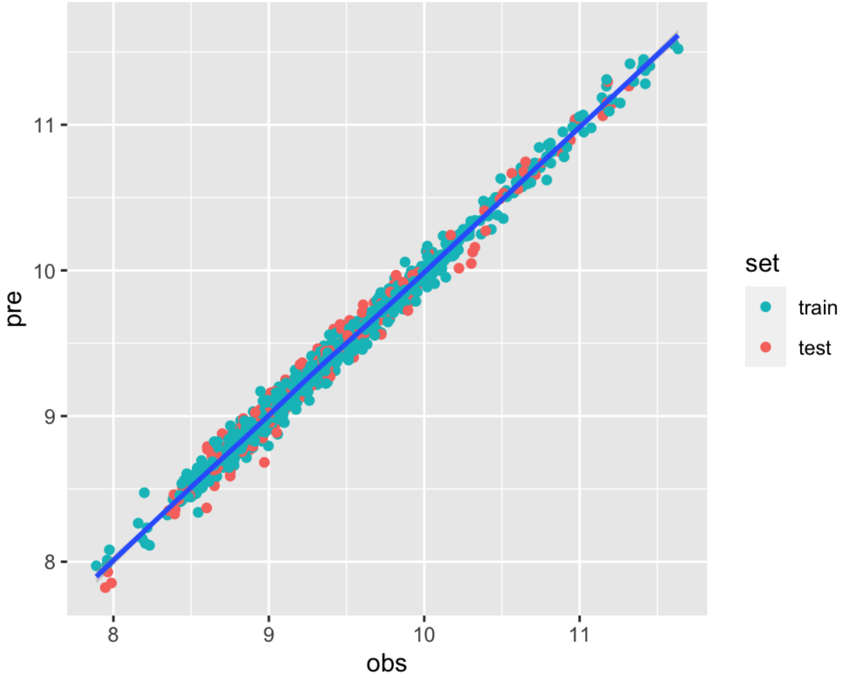


Figure 5.7: The prediction of training set and test set against the real observations.

5.4.1 Spatial effect

Recall that the BYM model is a union of the besag model $\boldsymbol{u}$ (spatial part) and an i.i.d. model $\boldsymbol{v}$ given by Equation (3.2). Plotting the estimations against geographical location clearly reveals spatial trends, which are shown in Figure 5.8. In most areas, there is no spatial trend that can be distinguished by the naked eye in the GLM part, but there are still individual bright spots. This can be explained by the provincial capital. The i.i.d. and besag part have very similar performances, such that their sum underscores the point that housing prices south of the Yangtze River are generally higher than those north of the Yangtze River.

5.4.2 Prediction into the future

However, the current data can only verify the validity and quality of the model. People and real estate companies are more concerned about future changes in housing prices. In our model, population and average years of education for people over 15 years old change sightly since we can regard them as constant across years. But GDP always has different changes in different regions and years. However, Modeling GDP separately and then forecasting housing prices using estimations of GDP is too cumbersome and increases the error. **R-INLA** provides an advanced function - copy model specified in Equation (3.9), which can share an effect that is estimated from two or more parts of the dataset so that all of them provide information about the effect when fitting the model. Hence, we can rewrite the model as

$$logPrice_{it} \sim 1 + logPd_i + eduy_i + year_t + f(logGDP_{it}) + u_i + v_i$$

where $f(logGDP_{it})$ will be modeled for $logGDP_{it}$ and shared.

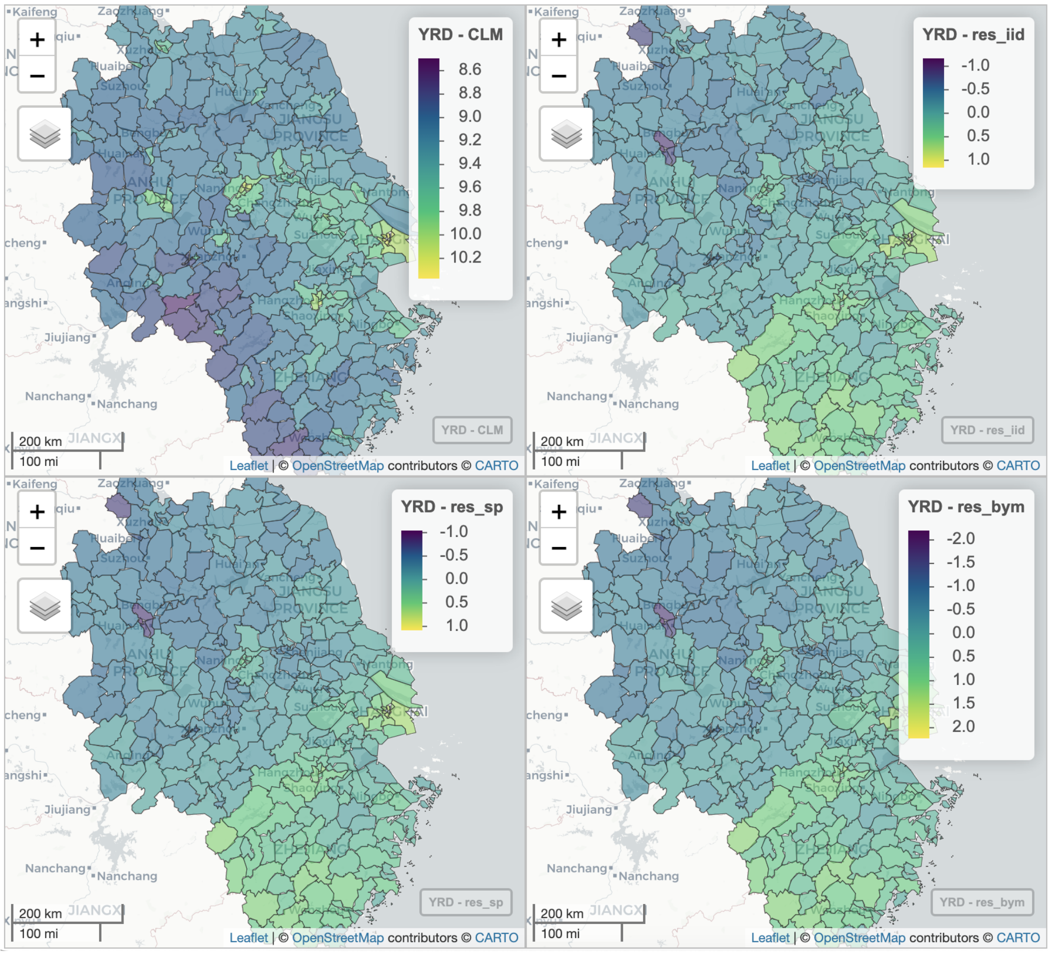


Figure 5.8: The plot in the upper left corner is estimates of the GLM-part, $\boldsymbol{z}\boldsymbol{\beta}$, in the BYM model (YRD-GLM); the upper right corner is the estimate of the i.i.d. part, $\boldsymbol{v}$ (YRD-res_iid); the lower left corner is the estimate of the besag part, $\boldsymbol{u}$ (YRD-res_sp); the lower right corner is the estimate of all the spatial part $\boldsymbol{u} + \boldsymbol{v}$ (YRD-res_bym).

The housing prices in 2023 are only collected at 231 prices. There still are 72 missing observations, where the model may have significant errors. In Figure 5.9 and Figure 5.10, the plots seem to show a curvilinear (non-linear) trend using the copy model since the prediction of GDP increases the error of the model. The prediction in 2023 is not as close to the observations as before. But most of their confidence intervals include the function $y = x$. Judging from the current situation, this model is credible. Although the model has some shortcomings, its advantages cannot be ignored.

The copy feature adds flexibility to model building. It allows us to construct hierarchical and multi-level models where random effects are naturally shared across different levels or components of the model. When different components of the model are driven by the same random effects, it is easier to understand the relationships and dependencies within the model. This can be crucial for maintaining consistency. For computers, random effects are estimated only once and used repeatedly, saving time and computing resources. We believe that GDP and time are related which implies there is a random effect between time and GDP, while education level (eduy) and population structure (logPd) change slightly at the macro level which implies there is no random effect between them and time. Therefore, when predicting future housing prices, it is also necessary to predict future GDP. In this case, the copy model can give full play to its advantages and the results obtained are more feasible.

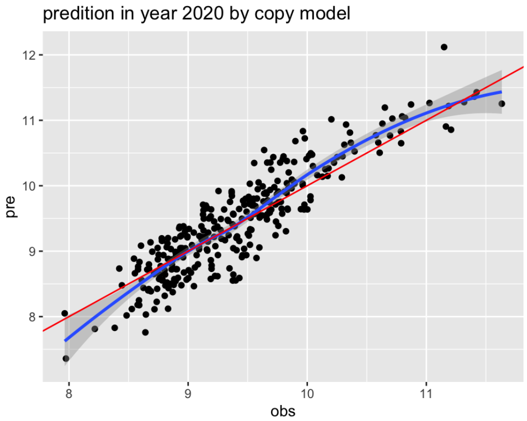


Figure 5.9: The prediction in 2020 using a copy model against the real observations.

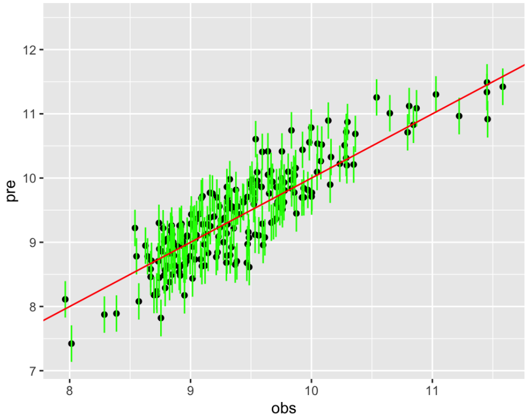


Figure 5.10: The prediction in 2023 of copy model. Green line is the error bar.

5.5 Comparison with two other regions

5.5.1 Comparison by temporal effect

Repeating the previous process, the spatial models of housing prices in the Pearl River Delta and Beijing-Tianjin-Hebei regions are built. For consistency, three models share the form given by Equation 5.1. And an '$AR(1)$' model is applied to *year* in all three models to compare temporal effect. Figure 5.11 shows that during these four years, housing prices in the PRD and YRD have exhibited a positive upward trend. Specially, housing prices in the YRD have grown the fastest. In contrast, the housing price trend in the BTH region has been relatively stable, with growth rates hovering around zero. This stability can be attributed to macro-control measures implemented by the government, which have effectively curbed rapid price increases and maintained market stability. Meanwhile, in the PRD housing prices increase regularly over time (more linear), indicating that economic development and macroeconomic control are working simultaneously in the region. Clearly, this kind of housing market is positive and healthy.

Overall, these temporal comparisons highlight the varying dynamics of housing price growth across different regions. The strong growth in the YRD underscores the region's robust economic development, while the stability in the BTH region reflects the effectiveness of governmental intervention in stabilizing the housing market. The PRD is somewhere in between. Understanding these regional differences is crucial for policymakers and stakeholders aiming to address housing affordability and market sustainability.

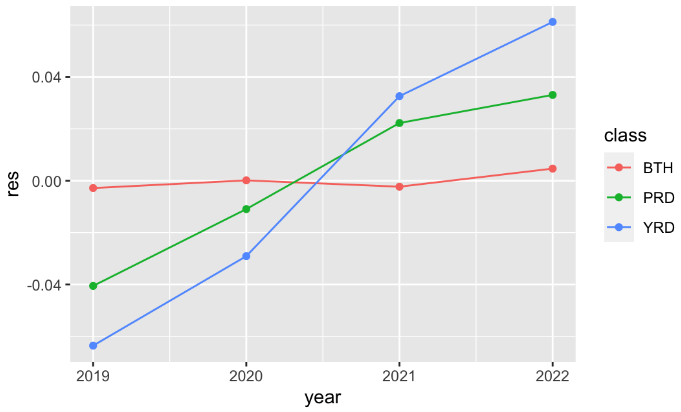


Figure 5.11: Temporal effects in the housing price model. An AR model used for year in the three regions.

5.5.2 Comparison by spatial effect and map

In Figure 5.12, the Beijng-Tianjin-Hebei (BTH) region exhibits a clear spatial trend, with housing prices increasing towards the center. However, while there is a noticeable highlight in the Pearl River Delta (PRD), the expected spatial effect is not evident on the map. This observation raises two potential explanations: Firstly, due to the region's comprehensive economic development, housing prices may have experienced fluctuations over the past few years without forming a distinct spatial trend. Alternatively, it's possible that the current model is not sufficiently robust to capture the underlying spatial dynamics. The latter explanation holds more credibility, primarily due to the limitations posed by the available data. As mentioned, housing prices in the YRD show north-south differences.

The GLM part of the PRD model in Figure 5.13 shows the east-west differentiation in housing prices. The GLM part of the BTH model has a similar spatial correlation as the spatial effects but it hints at the impact of transportation and economic activity on housing prices, as housing prices in areas along Beijing's route to the port are relatively high. These two patterns mainly depend on covariate GDP. In the PRD, due to large-scale population migration to central areas, housing prices in peripheral areas are significantly lower than those in central areas.

Then, in Figure 5.14, housing prices in three regions all have a significant spatial effect - rising towards one or more centers. Especially, in the YRD, housing prices in areas along the Yangtze River are higher than elsewhere and reach the highest in Shanghai. Meanwhile, housing prices in areas south of the Yangtze River are generally higher than in the north. There is only one obvious center in the BTH. At the same time, Tianjin has the largest port in the BTH region and the developed foreign trade economy makes its housing prices slightly higher than those in other regions (excluding Beijing). There are two gathering points in the Pearl River Delta region – the provincial capital city Guangzhou, and Hong Kong. These two gathering points represent two different development strategies of the Pearl River Delta, namely developing the inland economy (locally) with Guangzhou as the representative, and developing foreign trade (internationally) with Hong Kong as the representative. Additionally, the YRD stands for the region affected by geographical factors, where gathering points

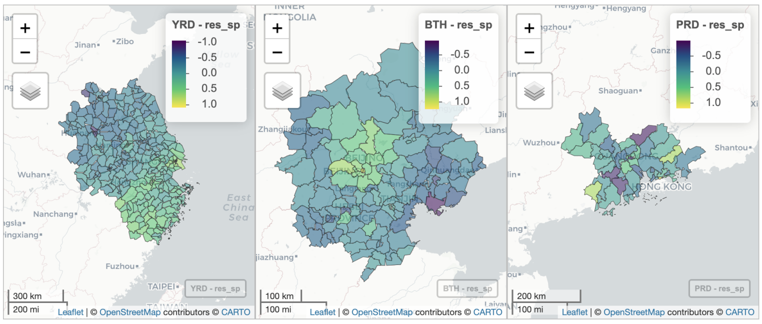


Figure 5.12: Spatial effects $\boldsymbol{u}$ in the YRD, BTH, and PRD model.

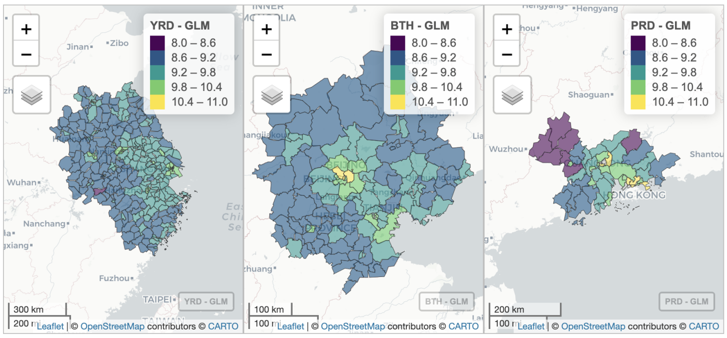


Figure 5.13: The GLM part, $\boldsymbol{z\beta}$, of the YRD, BTH, and PRD model in 2019.

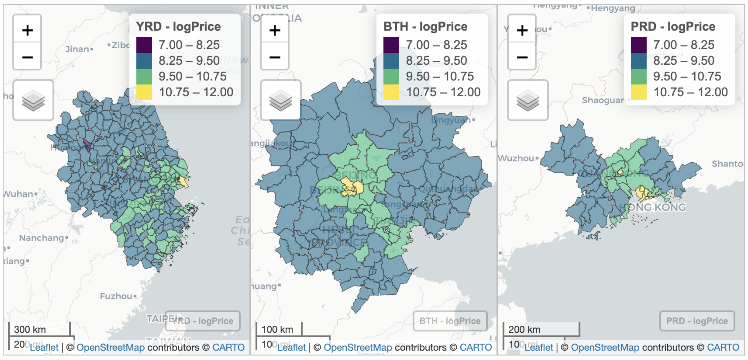


Figure 5.14: Housing prices in the YRD, BTH, and PRD in 2019.

appear in strips or blocks; the BTH represents the region with a special characteristic (like capital); the PRD represents the region affected by policies (like economic development strategies). Hence, the models in these three regions can be a good inspiration for models in other regions in China.

5.5.3 Comparison by model

Table 5.4 shows the β-coefficients in all three models. Beijing-Tianjin-Hebei is the political and cultural center of China, with a long history and cultural heritage. Due to policy reasons, population mobility in the BTH has stabilized and is much lower than that of the PRD and YRD. Hence, the effect of population is lowest. Additionally, due to the Beijing-Tianjin-Hebei (BTH) region's emphasis on cultural and political development rather than economic development, the β-coefficient of GDP in this region is smaller compared to other regions. However, the weight of education increases significantly. Also, Figure 5.15 shows that "logPd" is not significant in the BTH model because its confidence interval includes 0 and is uncertain for the PRD model because it has a large range of confidence interval and the lower bound is close to 0. Section 2.1.4 gives the reason: the population structure in the BTH is already stable while migration in the PRD is numerous and unstable. For the covariate "year", the estimates are consistent with the results in Section 5.5.1 – the temporal effect is most significant in the YRD while we might consider removing the covariate "year" in the BTH model because the temporal effects has minimal impact in this region.

Region	Intercept	logGDP	logPd	eduy	year
YRD	6.409	0.168	0.122	0.173	0.044
BTH	6.340	0.082	0.033	0.441	0.005
PRD	3.483	0.096	0.136	0.188	0.03

Table 5.4: β-coefficient in the three different regions.

The Yangtze River Delta is the earliest established and most developed economic and trade center in China. Hence the indicator representing the economy "GDP" plays the most important role in the model and the starting point (intercept) of the model is the largest. At the same time, the importance of population density due to prosperous economy has increased in housing price models compared to the BTH. Massive population movements have led to fluctuations in people's educational levels which explains the reduced importance of education in the model.

The Pearl River Delta has been a newly established economic development zone in the past two decades. Hence the starting point of housing prices is relatively low which is shown by the intercept of the model. This region's economic development level has surpassed BTH and is catching up with the Yangtze River Delta. In the past ten years, people have immigrated here in large numbers in search of wealth and business. The increase in population and its density further promotes changes in housing prices. Like the YRD, they have very similar performance in terms of population and education.

5.5.4 Summary

A spatial statistical modeling of housing prices in the Yangtze River Delta (YRD), when compared to the Pearl River Delta (PRD) and the Beijing-Tianjin-Hebei (BTH) coordinated development regions, reveals several notable insights.

In the YRD, housing prices have remained consistently high, particularly in cities like Shanghai, Nanjing, and Hangzhou. This can be attributed to their advantageous geographical location, robust economic development, and the increasing demand for quality housing and comfortable living conditions. For instance, Shanghai saw a 2.7% year-on-year increase in housing prices in 2021, with an average price reaching 60,000 yuan per square meter. This trend is expected to continue due to the region's economic growth and improving living standards.

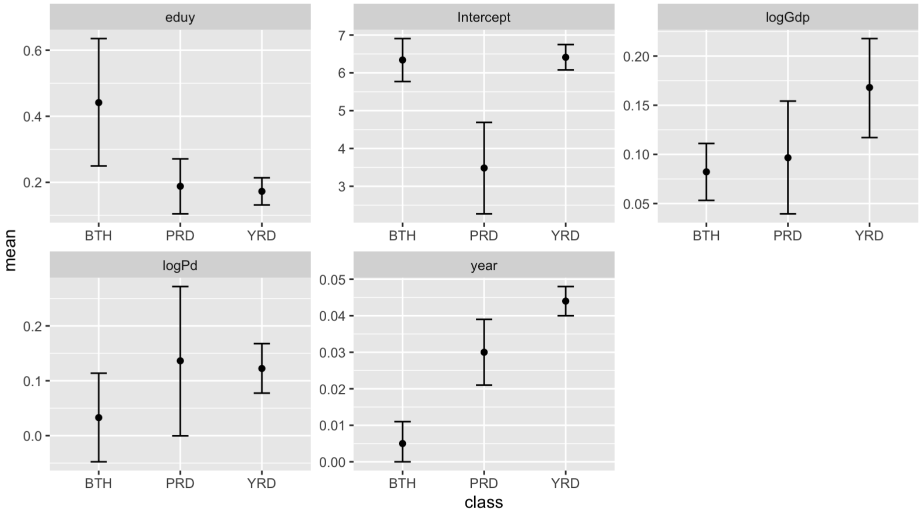


Figure 5.15: β-coefficient and confidence interval.

In contrast, the PRD region exhibits a more stable real estate market. Characterized by a high number of commercial and residential buildings, the PRD, notably Guangzhou, has a vibrant commercial real estate sector. Additionally, the region's real estate market is maturing, with an increasing number of mid-term and other long-term property projects available for purchase. When comparing the YRD and PRD with the BTH coordinated development region, differences in economic policies, infrastructure investments, and population dynamics likely play a significant role in shaping housing price trends. However, the spatial statistical modeling reveals that the YRD maintains a distinct lead in terms of housing prices, while the PRD offers a more balanced market with a diverse range of housing options.

In summary, the spatial statistical modeling of housing prices in the YRD, compared to the PRD and BTH regions, highlights the unique economic and geographic dynamics that shape real estate markets in these vital economic hubs of China.

6 Discussion and conclusions

6.1 Review

This study focuses on the spatial statistical modeling of housing prices in the Yangtze River Delta (YRD) region and its comparative analysis with the Pearl River Delta (PRD) and Beijing-Tianjin-Hebei (BTH) coordinated development area. The main objectives are to identify spatial patterns and determinants of housing prices in YRD and to assess how these patterns differ from PRD and BTH.

In Chapter 2, an overview of three regions and available data is presented. In Chapter 3, the statistical model for housing prices, Y, incorporating spatial dependence through spatial auto-regression, was introduced.

In Chapter 4, the inference method utilizing Integrated Nested Laplace Approximation (INLA) was elaborated upon. The models were presented within a fully Bayesian framework. And some metrics, such as Mean Squared Error (MSE), Deviance Information Criterion (DIC), Widely Applicable Information Criterion (WAIC), and Moran's I test, were proposed for general model selection.

The first part of Chapter 5 addressed the methods of selecting covariates like linear correlation and confidence intervals. The second part involved a comparison of models, revealing that the ordinary Generalized Linear Model (GLM) model could be enhanced by incorporating both the unstructured effect (v) and the spatial effect (u), resulting in the BYM model. In Section 5.4, it was shown that the results of the models for housing prices are in compliance with the observations and facts. The predictions are satisfactory. However, because of the lack of observations, the quality of future predictions is difficult to judge. Overall, the results obtained so far are acceptable.

In section 5.5, the models of the three regions are slightly different due to their respective characteristics. But the spatial model effectively captures the spatial and temporal effects of housing price changes in these three regions, as all of them exhibit significant positive spatial autocorrelation. Housing prices in YRD exhibit significant spatial variation, with higher prices concentrated in urban centers and along transportation corridors. Meanwhile, Beijing, the capital of China, is the center of BTH, and Shenzhen and Guangzhou become the center of the Pearl River Delta by economic development strategy. Economic development, population density, education resources, and geographical factors were found to be important determinants of housing prices in YRD. Compared to PRD and BTH, YRD showed a stronger correlation between housing prices and economic indicators, suggesting a more market-driven housing market. PRD and BTH also exhibited distinct spatial patterns and determinants, reflecting their unique economic, social, and policy contexts. Lessons learned from the comparative analysis can be applied to improve policy coordination among YRD, PRD, and BTH, enhancing the effectiveness of regional development strategies.

6.2 Model structure

By specifying the model as shown in Equation (3.7), simultaneous estimation of the covariate effect ($\boldsymbol{z_i \beta}$), the spatial effect (u_i), and the random effect (v_i) is achieved. The distribution of strength across each component becomes data-driven. An alternative approach involves initially conducting a pure GLM analysis, and subsequently utilizing this estimate as an offset when estimating u_i and v_i. In this scenario, u_i cannot be directly interpreted as a spatial effect but is instead viewed as an additional factor. This method consequently raises statistical concerns.

6.3 Future Study

Unfortunately, due to the late start in statistical analysis of the real estate industry in China and the lack of public data, the available data is limited. It is regrettable that the amount of data collected is relatively small. This scarcity of data poses a challenge in thoroughly analyzing the subject matter and drawing comprehensive conclusions. The limited dataset restricts the depth of our investigation and may hinder the accuracy of our findings. Considering the possibility of increasing the total sample size by adding more years of data or appropriately incorporating data from surrounding areas could help address this issue. In future research endeavors, efforts should be made to gather a more extensive and diverse dataset to provide a more robust foundation for analysis and interpretation.

6.3.1 Model improvement

An important aspect of model improvement involves selecting appropriate hyperpriors. However, this issue was not explored in this thesis, and thus, the impact of different hyperprior choices remains unknown. In practical applications, it is advisable to at least consider this question to some extent. For the housing prices data analyzed in this thesis, no specific prior beliefs were held regarding the hyperparameters in Equation (3.7). Therefore, one could argue that opting for a flat prior was a fair choice.

Then, the method of choosing covariates is sketchy. In fact, there are only a few possible covariates can be chosen such that covariates may not be enough to explain the model. Maybe some missing covariates can play a significant role in this model. For example, in this thesis, one variable is selected in each of the three aspects of economy, population, and culture, but in fact, these three indicators can select more variables. Further research could explore the dynamic interactions between housing prices and other economic indicators over time, providing a more nuanced understanding of the spatial evolution of housing markets. The integration of additional datasets, such as land use, environmental factors, and social media data, could offer new perspectives on the drivers of housing prices. As mentioned above, politics should also be reflected in the housing price model, but this article did not collect appropriate relevant data. Moreover, it's plausible that certain variables are not best represented by linear effects but rather exhibit non-linear patterns, potentially warranting the consideration of spline effects. Also, considering the spatial autocorrelation structure more comprehensively could enhance the spatial modeling aspect. Techniques such as spatially varying coefficient models or spatial autoregressive models with varying spatial weights matrices could capture spatial heterogeneity more effectively. Incorporating additional spatial predictors or exploring alternative spatial weighting schemes may also improve the spatial predictive performance of the model. Taking these nuances into account could lead to improved model accuracy. The observed discrepancies in prediction accuracy depicted in Figure 5.9 and Figure 5.10 might stem from suboptimal regression modeling. Nonetheless, enhancing the regression component is unlikely to alter the fundamental conclusions drawn in this thesis. It's improbable that refining the covariate analysis would entirely eliminate the spatial effect. However, there remains room for model enhancement at this stage.

6.3.2 Extending the model to a smaller area

Obviously, the division used in this article is not precise enough. The division adopted in this article is at the district and county level, with an average area of about 1170 km^2 and including many cities. This is not necessary for most house buyers. People are more interested in housing prices in a specific and small area, such that reducing the overall area and using grids to refine the area is reasonable. This method will greatly increase the amount of data, making the inference problem computationally intensive, and making it more urgent to use deterministic inference methods such as INLA. But as long as the number of regions remains within the order of 10^5 [25], the problem can still be solved within a feasible time using INLA.

Furthermore, we can reduce the overall area, limit the model to a certain city, and divide the area to the suburbs based on the city center. However, data collection is a difficult problem with this method. At the same time, more stable covariates should be selected to reduce errors caused by statistics and data collection. For individuals, this approach is almost impossible.

6.3.3 Extending the model to a larger area

However, the subject of this thesis is observing and modeling house prices at the macro level, rather than targeting an individual city. Therefore, extending the model to a larger area is a more reasonable direction, for example, modeling the eastern coastal area of China. This is more conducive to the national macro-control of regional housing prices. This method will also greatly increase the data set, and as mentioned above, INLA solves this problem very well.

6.3.4 More temporal data

In fact, the lack of public data related to housing prices before 2015 is very serious. But adding more years of data can still improve the accuracy of the model. By increasing the temporal scope of the dataset, we can capture a broader range of trends and variations in housing prices over time. This expanded time allows for a more comprehensive analysis of market dynamics and better prediction of future trends and overall performance.

6.4 Conclusion

Although this thesis has initially constructed a general hypothesis model of integrated regional housing price growth and spatial differentiation, there is no doubt that the spatial differentiation phenomenon of regional urban housing prices is characterized by diversity and complexity. As an exploratory work, there is still a lot of work that needs to be improved and deepened.

Inspecting Figure 5.14, house prices appear to be higher in densely populated and economically prosperous areas. This is an intuitive result because the market relationship in these areas is more biased toward supply exceeding demand, which is a seller's market. The strong predictive performance of the BYM model for housing prices in Section 5.3 suggests that spatial dependence should be considered in future house price modeling.

From a macro perspective, spatial smoothing of the final pricing factors is necessary. This helps the country regulate housing prices, avoid falsely high housing prices in certain places, and curb real estate speculation.

From a national perspective, these three metropolitan areas represent a higher stage of regional integration development, and their urban housing price growth and spatial differentiation patterns reflect, to a certain extent, the general patterns and development trends of the spatial and temporal evolution of housing prices in other urban agglomerations.

To wrap things up, accounting for spatial dependence when modeling housing prices yields a better model fit and the results are positive from a macro perspective.

Bibliography

[1] Wang Yang, Wang De-li and Wang Shao-jian. 'Spatial Differentiation Patterns and Impact Factors of Housing Prices of China's Cities'. In: *SCIENTIA GEOGRAPHICA SINICA* 33.10, 1157 (2013), pp. 1157–1165. DOI: `10.13249/j.cnki.sgs.2013.010.1157`. URL: `http://geoscien.neigae.ac.cn/CN/10.13249/j.cnki.sgs.2013.010.1157`.

[2] Alfred Weber. 'The Theory of The Location of Industries'. In: *The University of Chicago Press, Chicago & London.* (1909).

[3] Walter Christaller. *Die zentralen Orte in Süddeutschland.* Jena: Gustav Fischer, 1933.

[4] Homer Hoyt. *The Structure and Growth of Residential Neighborhoods in American Cities.* Washington, D.C.: Federal Housing Administration, 1939.

[5] Georges Matheron. *Principles of Geostatistics.* Originally published as "Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature" in 1965. Paris: Paris School of Mines, 1963.

[6] Danie G. Krige. 'A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand'. In: *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52 (1951). This paper introduces the concept of making best linear unbiased predictions (BLUP) for spatially correlated data, which later became known as kriging., pp. 119–139.

[7] Peter Whittle. 'On Stationary Processes in the Plane'. In: *Biometrika* 41.3/4 (1954). This paper deals with the spatial prediction of random fields and has been influential in the development of spatial statistics., pp. 434–449.

[8] William Alonso. 'Location and land use'. In: *Cambridge: Harvard University Press* (1964).

[9] Waldo R Tobler. 'A computer movie simulating urban growth in the Detroit region'. In: *Economic geography* 46.sup1 (1970), pp. 234–240.

[10] Yehua Dennis Wei. 'Urbanization and sustainable urban development'. In: *Journal of Planning Education and Research* 34.2 (2014), pp. 233–247.

[11] Simon Elias Bibri. 'Urban sustainability and digital technology'. In: *Sustainability* 9.2 (2017), p. 261.

[12] Sajal Ghosh and Kakali Kanjilal. 'Long-term equilibrium relationship between urbanization, energy consumption and economic activity: empirical evidence from India'. In: *Energy* 66 (2014), pp. 324–331.

[13] Hooi Hooi Lean and Russell Smyth. 'Regional house prices and the ripple effect in Malaysia'. In: *Urban Studies* 50.5 (2013), pp. 895–922.

[14] Scott Orford. 'Modelling spatial structures in local housing market dynamics: A multilevel perspective'. In: *Urban Studies* 37.9 (2000), pp. 1643–1671.

[15] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer Science & Business Media, 2002.

[16] *The Seventh National Population Census of the People's Republic of China*. National Bureau of Statistics of China, 2021.

[17] National Bureau of Statistics of China. *China Statistical Yearbook*. China Statistical Publishing House, 2019,2020,2021,2022.

[18] Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu and Tsutomu Watanabe. 'On the evolution of the house price distribution'. In: (2011).

[19] Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

[20] Julian Besag. 'Spatial Interaction and the Statistical Analysis of Lattice Systems'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225.

[21] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[22] Alan E Gelfand, Peter Diggle, Peter Guttorp and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

[23] Thiago G Martins, Daniel Simpson, Finn Lindgren and Håvard Rue. 'Bayesian computing with INLA: new features'. In: *Computational Statistics & Data Analysis* 67 (2013), pp. 68–83.

[24] Elias T. Krainski, Virgilio Gómez-Rubio, Amanda Lenzi Haakon Bakka, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren and Håvard Rue. *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman & Hall/CRC Press., 2019.

[25] Håvard Rue, Sara Martino and Nicolas Chopin. 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.2 (2009), pp. 319–392.

[26] Finn Lindgren and Håvard Rue. 'Bayesian spatial modelling with R-INLA'. In: *Journal of statistical software* 63.19 (2015).

[27] David J Spiegelhalter, Nicola G Best, Bradley P Carlin and Angelika Van Der Linde. 'Bayesian measures of model complexity and fit'. In: *Journal of the royal statistical society: Series b (statistical methodology)* 64.4 (2002), pp. 583–639.

[28] Sumio Watanabe. 'A widely applicable Bayesian information criterion'. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 867–897.

[29] Andrew Gelman, Jessica Hwang and Aki Vehtari. 'Understanding predictive information criteria for Bayesian models'. In: *Statistics and computing* 24 (2014), pp. 997–1016.

[30] Patrick AP Moran. 'Notes on continuous stochastic phenomena'. In: *Biometrika* 37.1/2 (1950), pp. 17–23.

[31] Virgilio Gómez-Rubio. *Bayesian inference with INLA*. Chapman & Hall/CRC Press., 2020.

[32] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[33] Song Weixuan, Chen Yanru, Sun Jie and He Miao. 'Spatial differentiation of urban housing prices in integrated region of Yangtze River Delta'. In: *Acta Geographica Sinica* 75.10, 2109 (2020), pp. 2109–2125. DOI: 10.11821/dlxb202010006. URL: https://www.geog.com.cn/CN/10.11821/dlxb202010006.

[34] Yu Jiachen. 'Prediction on Housing Price Based on the Data on Kaggle'. In: *2022 3rd International Conference on E-commerce and Internet Technology (ECIT 2022)*. Atlantis Press. 2022, pp. 627–634.

[35] Cao Wenyan and Qiao jian. 'Research on the Housing Prices in Beijing Based ON Spatial Econometric Model'. In: *Statistics and Application* 12 (2023), p. 1034.

[36] Moraga Paula. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC Biostatistics Series, 2019.

[37] Yong Zhao. 'The origin and application history of statistical analysis of spatial data'. In: *GEOGRAPHICAL RESEARCH* 37.10, 2058 (2018), pp. 2058–2074. DOI: 10.11821/dlyj201810014. URL: https://www.dlyj.ac.cn/EN/10.11821/dlyj201810014.

[38] Alastair S Adair, Jim N Berry and W Stanley McGreal. 'Hedonic modelling, housing submarkets and residential valuation'. In: *Journal of property Research* 13.1 (1996), pp. 67–83.

[39] Ana L Papoila, Andrea Riebler, Antónia Amaral-Turkman, Ricardo São-João, Conceição Ribeiro, Carlos Geraldes and Ana Miranda. 'Stomach cancer incidence in Southern Portugal 1998–2006: A spatio-temporal analysis'. In: *Biometrical Journal* 56.3 (2014), pp. 403–415.

[40] Julian Besag, Jeremy York and Annie Mollié. 'Bayesian image restoration, with two applications in spatial statistics'. In: *Annals of the institute of statistical mathematics* 43 (1991), pp. 1–20.

Master's Theses in Mathematical Sciences 2024:E24
ISSN 1404-6342

LUNFMS-3127-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/