

Geostatistical and Spatial Econometric Analysis for Regional and Real Estate Economics

March 30th, 2024

Sachio Muto
CREI, University of Tokyo



I. Forecasting Regional Economic Statistics

Forecasting the housing vacancy rate in Japan using dynamic spatiotemporal effects models
Japanese Journal of Statistics and Data Science (JJSD), Vol. 6, 21–44, (2023), with S. Sugasawa and M. Suzuki

I-1. Problem of “Vacant Housing” in Japan

<Examples of possible problems>

Deterioration of disaster resistance

- Collapse, collapse, fall of roof/exterior walls, possible fire

Decrease in crime prevention capability

- Inducement of crime

Illegal dumping of garbage

Deterioration of sanitation, generation of bad odor

- Mosquitoes, flies, rats, stray cats, concentration

Deterioration of scenery and landscaping

Other

- Overgrowth of tree branches, weeds, scattering of fallen leaves

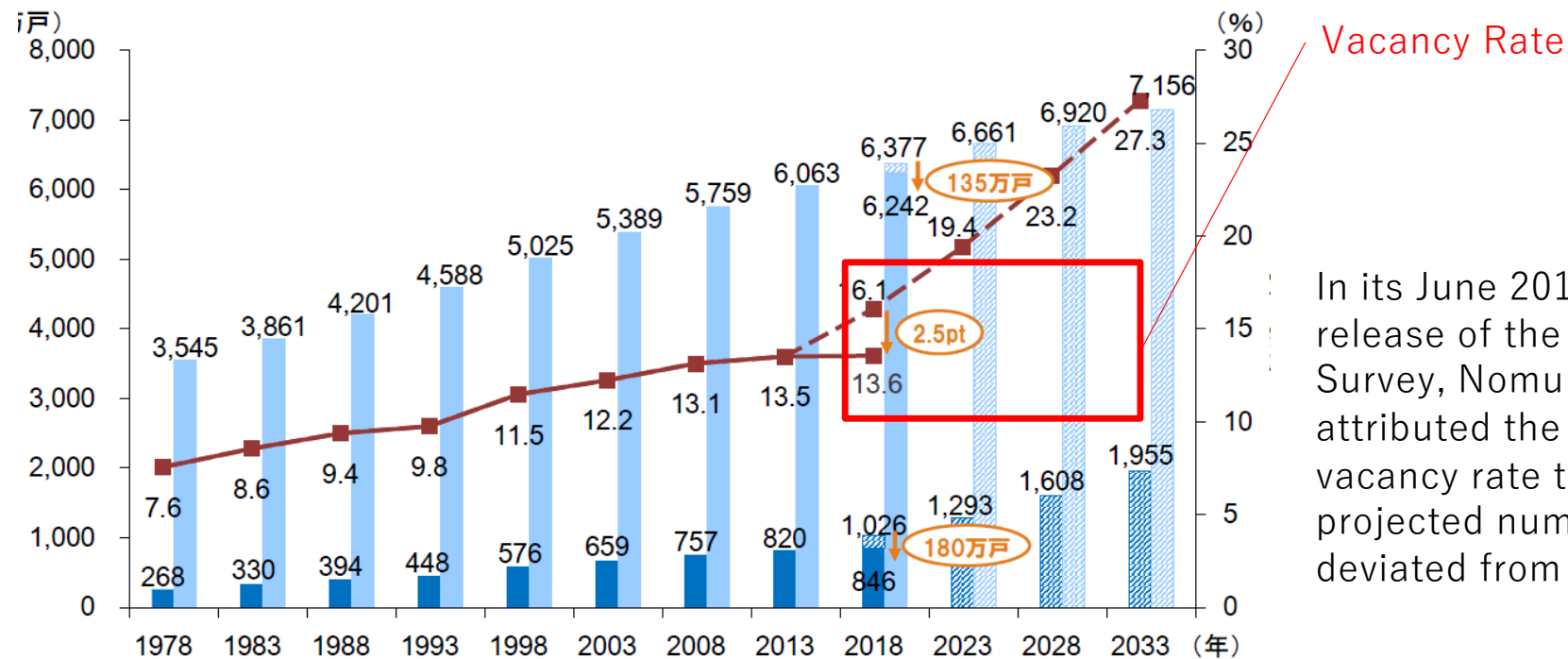


(Source: MLIT)

I-2. Data regarding "Vacant House Rate" in Japan

In the 2008 Housing and Land Survey, the "vacant house rate" was released, and it became a hot topic of conversation among those concerned that the national figure fell far short of the prediction, coming in at 13.6%, compared to the previously well-known estimate by Nomura Research Institute (16.1%).

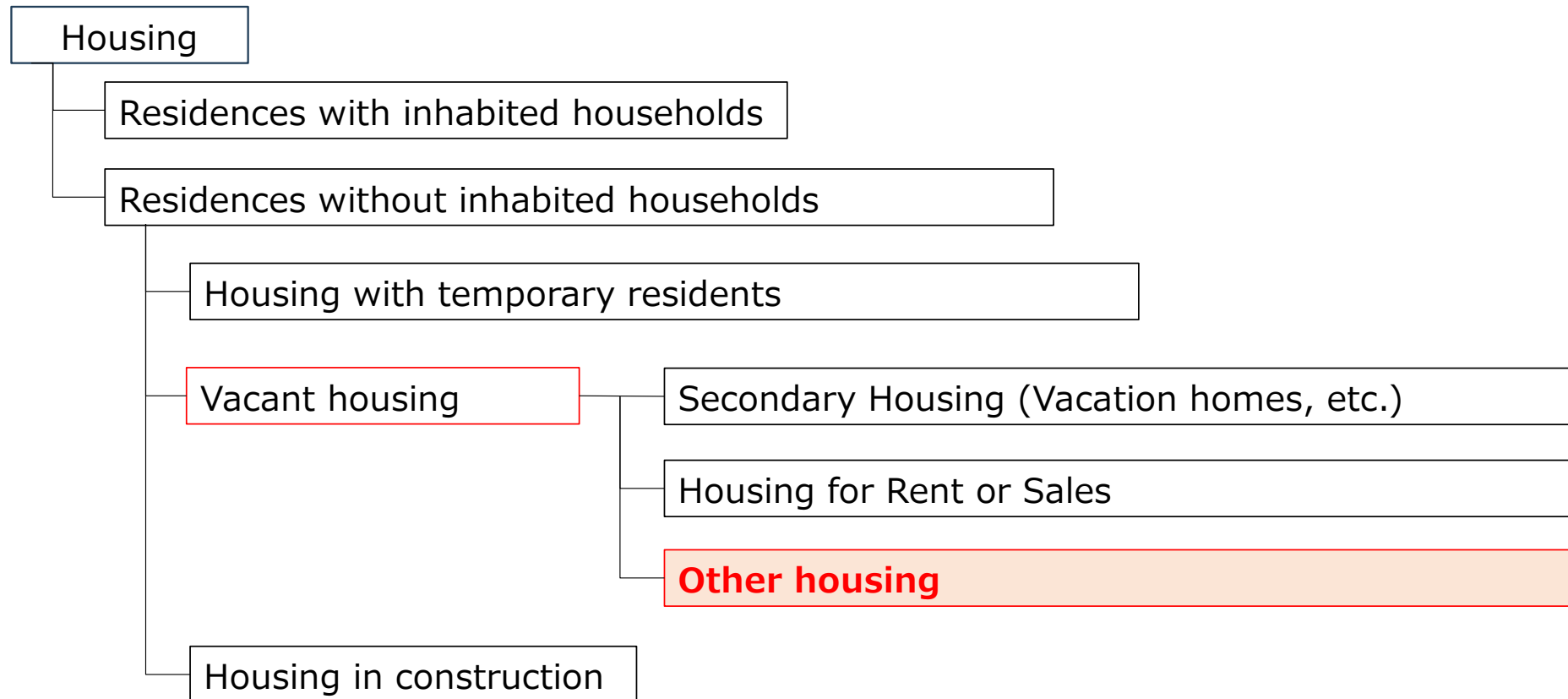
Total number of housing units, number of vacant housing units, and actual and projected vacancy rates



In its June 2019 report following the release of the Housing and Land Survey, Nomura Research Institute attributed the error in estimating the vacancy rate to the fact that the projected number of removals deviated from the actual.

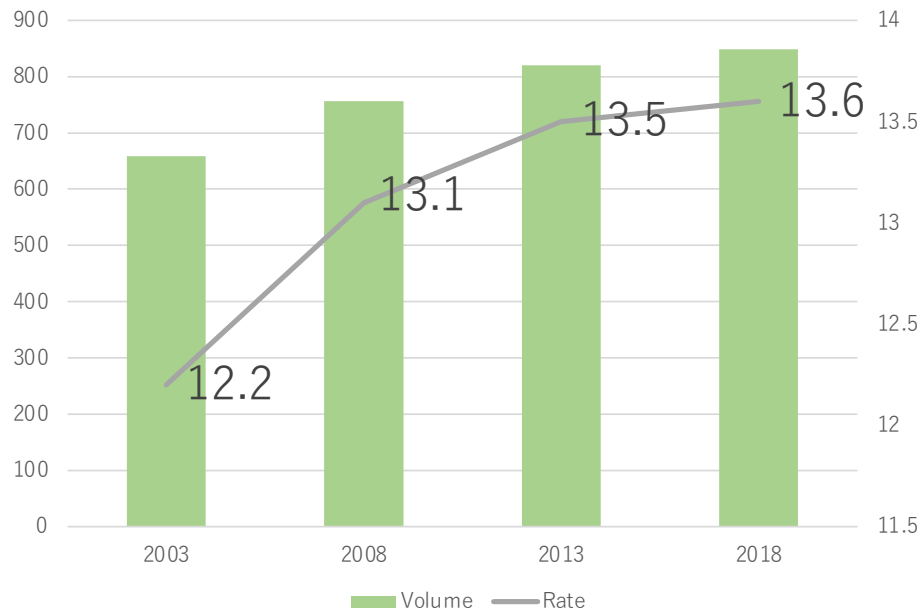
The so-called "vacant house" rate includes those that are vacant as houses for rent or sale, vacation homes, etc., and of policy importance are the so-called "other" vacant housing units (also called "other vacant" housing units).

"Housing" in National Survey of Housing and Land Statistics

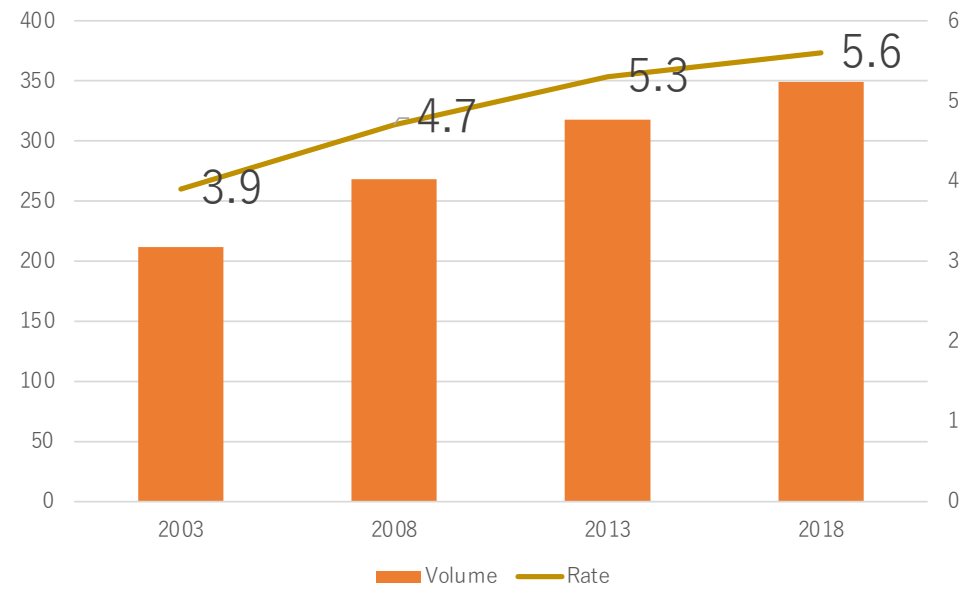


Increase of "Other Vacant Housing"

Total Vacant Housing

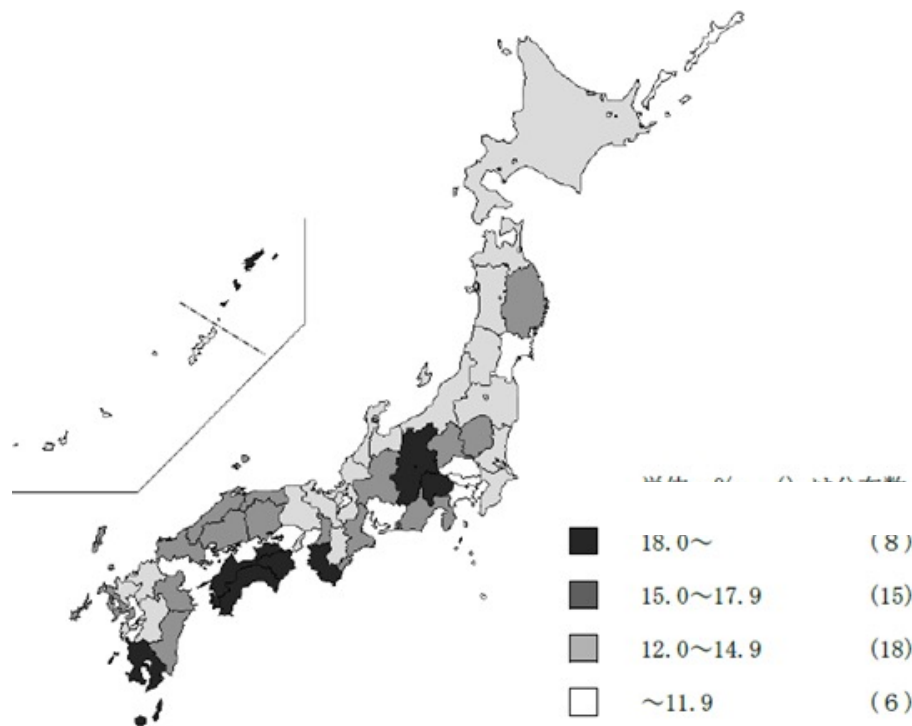


"Other" Vacant housing

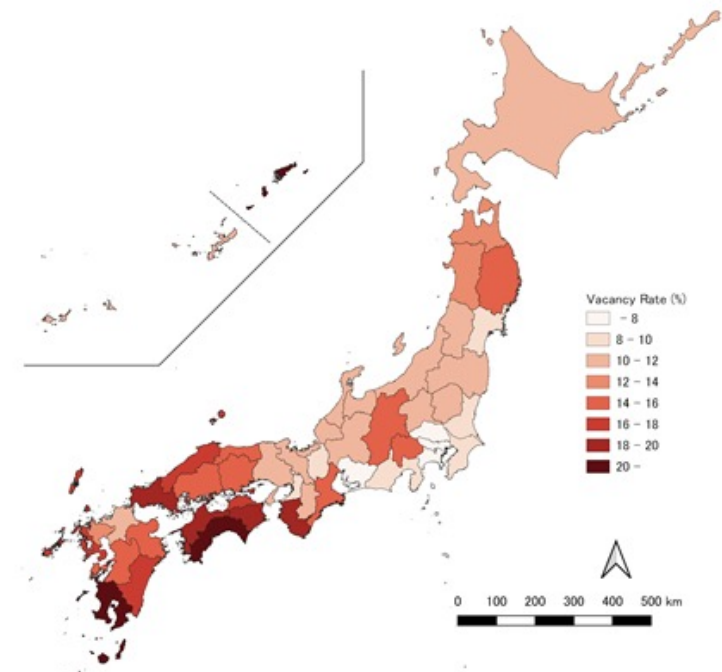


The "vacant house rate" is higher in Nagano and Shizuoka, where there are many vacation homes, etc., while the Rate of "other" vacant housing is in western Japan. (In any case, they are geographically unevenly distributed.)

Rate of Total Vacant Housing (2018)

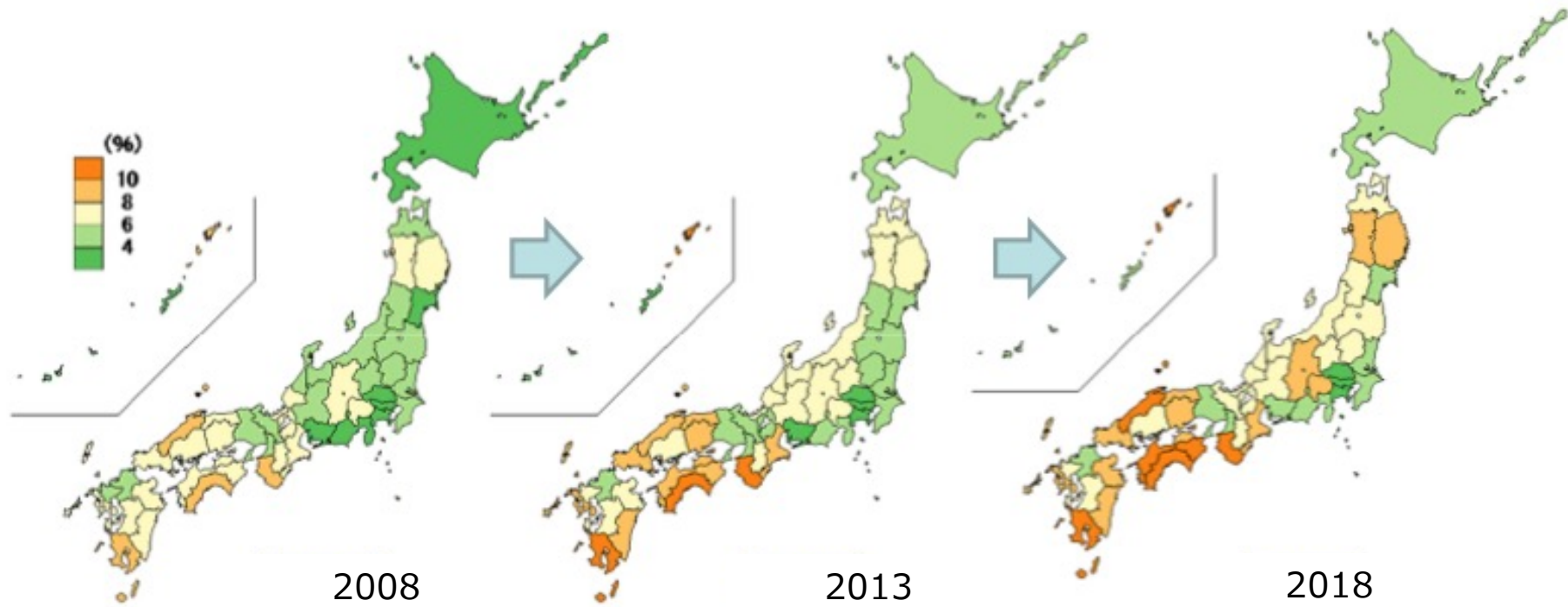


Rate of "Other" Vacant Housing (2018)



Source: Housing and Land Survey (MIC, Japan), 2018

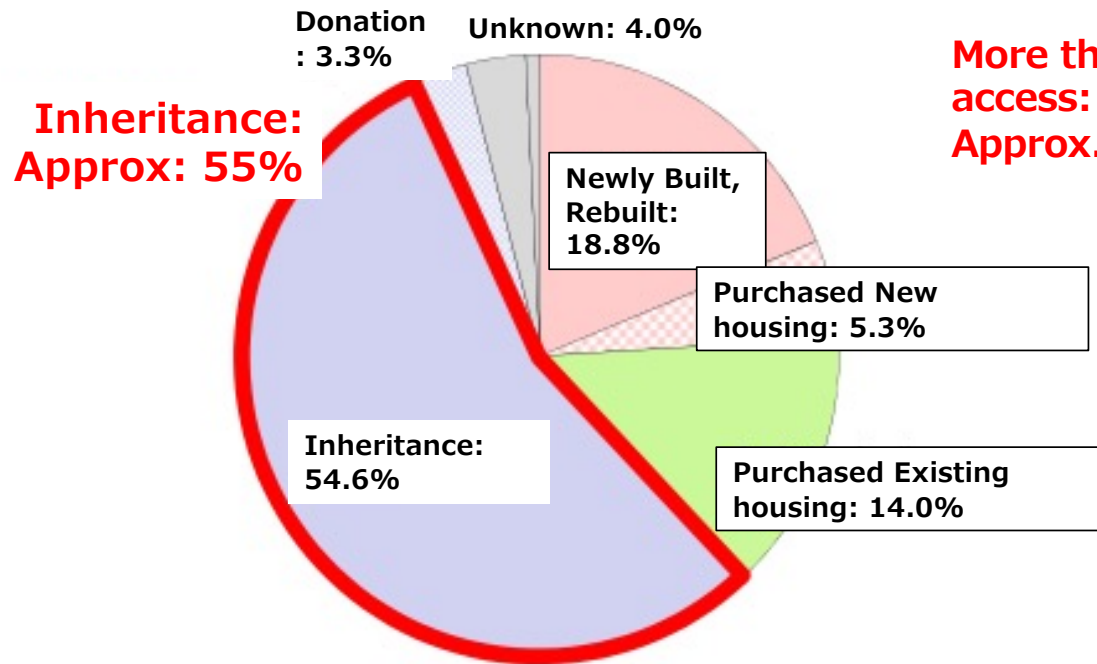
Transition of the Rate of “Other” Vacant Housing



Source: Housing and Land Survey (MIC, Japan), 2018

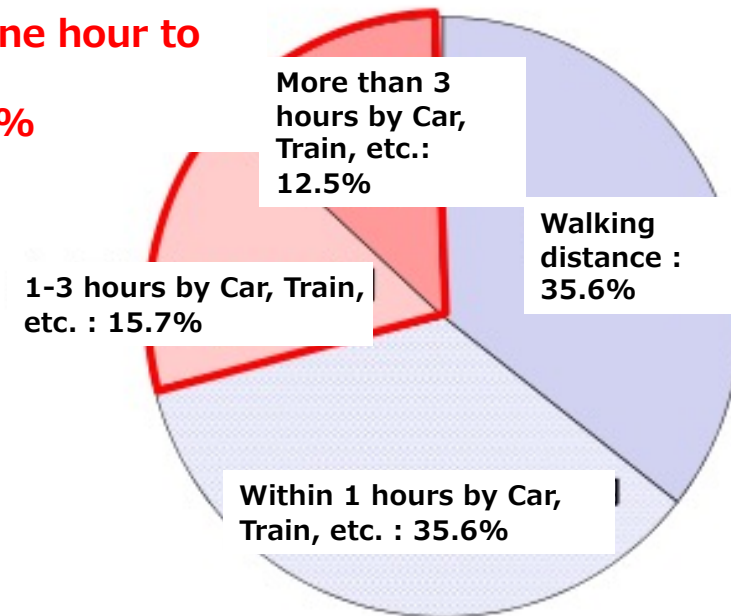
Role of Inheritance and location of Vacant housing (Japan, 2018)

Acquisition history of vacant houses



Location and distance between vacant housing and owners' residence

More than one hour to access: Approx.: 30%



Source: MLIT

I-2. Analytical Strategy

Build a model to forecast the "other vacant house rate" in each prefecture based on future population estimates, etc.

Explanatory variables that may affect the occurrence of "other" vacant houses

i) Age cohort in each region

- a. Number of changes in the number of elderly (65 years old and over, 75 years old and over) in each prefecture
→ The increased probability of mortality and gives away or inherits the house.
- b. Population aged 18 and under 65 in each prefecture
→ Possibility of inheritance (including inheritance in absentia)

a. and b. occur in the same prefecture
→ Likely to be used as housing with residence

a. and b. occur in distant prefectures
→ High probability that the house will be used as other vacant house



When the balance between a and b is disrupted in a prefecture ($a \gg b$), the number of other vacant houses will increase.

ii) Other attributes (May be spatially correlated)

I-3. Spatial Econometrics and Geostatistical Model

➤ Spatial Autocorrelation (Spatial Durbin)

Spatial Durbin Model LeSage and Pace (2009), Elhorst (2014)
Kawabata and Abe (2018)

$$y = \rho W y + a i_n + X \beta + W X \theta + \epsilon$$

$$y = (I_n + \rho W)^{-1} (a i_n + X \beta + W X \theta + \epsilon)$$

$$\epsilon \sim N(0, I_n)$$

$$\times (I_n + \rho W)^{-1} = I_n + \rho W + \rho^2 W^2 \dots$$

W (Weight Matrix) as explanatory variable and introduce spatial correlation as a coefficient ρ

➤ Spatial Autocorrelation (Geostatistics)

$$Y = X \beta + \omega + \epsilon,$$

where $\epsilon \sim N(0, \tau^2 I_n)$ and $\omega \sim N(0, \sigma^2 H(\phi))$
(ϵ and ω are mutually independent)

or

$$Y | \beta, \sigma, \phi, \tau \sim N(X \beta, \Sigma), \text{ where } \Sigma = \sigma^2 H(\phi) + \tau^2 I_n$$

Generates spatial autocorrelation in the error term

→ Provides flexibility to fit to data

It cannot be solved analytically and estimated with the Bayesian MCMC

I-4. Estimation Model

1) Basic Structure

$$Y = X\beta + \omega + \varepsilon,$$

where $\varepsilon \sim N(0, \tau^2 I_n)$ and $\omega \sim N(0, \sigma^2 H(\phi))$
 (ε and ω are mutually independent)

or
 $Y|\beta, \sigma, \phi, \tau \sim N(X\beta, \Sigma)$, where $\Sigma = \sigma^2 H(\phi) + \tau^2 I_n$

y : The rate of “Other Vacant Houses”

$$X = \left[\underbrace{\frac{Pop(Age\ 0-29)}{Total\ Pop}, \frac{Pop(Age\ 30-64)}{Total\ Pop}, \frac{Pop(Age\ 65-74)}{Total\ Pop}}_{\text{Prefectural (Current Period} \cdot 1 \text{ term (5 years lag))}}, \frac{Pop(Age\ 75\ and\ over)}{\text{総人口}}, x_0 \right]$$

x_0 : Prefectural Fixed Effect Dummy

$H(\phi) = I_n \otimes h(\phi)$: $h(\phi)$ Location matrix (based on longitude and latitude information)

Matrix is developed through (i, j) element of $h(\phi)$ to be $\exp(-\|s_i - s_j\|^2 / 2\phi)$

Denoting the data point as n , $H(\phi)$ is $n \times n$ matrix, where $n = 47$ prefectures

2) Introduction of time correlation and the Dynamic Spatiotemporal Effects (DSE) model

For modeling y_t with X_t and spatial information while accounting for spatial and temporal correlation, we employ the following regression model with unobserved spatiotemporal effects:

$$y_t = X_t\beta + \omega_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \tau^2 I_n), \quad t = 1, \dots, T, \quad (1)$$

where I_n is an $n \times n$ identity matrix, ε_t and ω_t are mutually independent error terms and spatial effects at time t , respectively, and τ^2 is an error variance. For the unobserved spatial effects ω_t , we assume the following autoregressive (AR) model:

$$\omega_t = \delta \omega_{t-1} + u_t, \quad u_t \sim N(0, \sigma^2 H(\phi)), \quad \omega_0 = 0, \quad (2)$$

where u_t is an n -dimensional innovation, δ is an autocorrelation parameter, σ^2 is a variance parameter of the spatial effects, and $H(\phi)$ is a correlation matrix whose (i, j) -element is $\rho(s_i - s_j; \phi)$. Here, ρ is a valid isotropic correlation function indexed by an unknown range parameter ϕ . We used the Gaussian correlation ρ_G , given as follows:

$$\rho_G(s_i - s_j; \phi) = \exp(-\|s_i - s_j\|^2 / \phi^2). \quad (3)$$

3) DSE-AR model and DSE-RW model

Under model (1), the conditional distribution of y_t given ω_{t-1} (spatial effects in the previous time point) is $N(X_t\beta + \delta\omega_{t-1}, \tau^2I_n + \sigma^2H(\phi))$, such that spatial correlation is introduced in y_t . The distribution depends on ω_{t-1} , which introduces temporal correlation among y_t s. Specifically, the marginal covariance of y_t and y_s ($t \neq s$) is $\text{Cov}(y_t, y_s) = \delta^{t-s}\sigma^2H(\phi)/(1 - \delta^2)$ for $|\delta| < 1$ and $\text{Cov}(y_t, y_s) = (t - s)\sigma^2H(\phi)$ for $\delta = 1$. Hence, δ controls the strength of the serial correlation, and the spatial effects in different time points are independent when $\delta = 0$. Here, we call the model (1) with (2) DSE model with AR spatial effects (DSE-AR); this model captures the temporal correlation of y_t .

As a representative sub-model, we set $\delta = 1$ instead of estimating from the data, leading to a RW model for the spatial effects, which would be useful to capture potential non-stationarity. This model is called the DSE model with RW spatial effects (DSE-RW). It may be possible to consider using a

4) Complete Conditionals for Bayesian MCMC

We draw β and ω from complete conditionals:

$$\beta \mid \omega, \sigma, \phi, \tau, y \sim N(D_\beta d_\beta, D_\beta),$$

where $D_\beta = (X\Sigma^{-1}X' + V_\beta^{-1})^{-1}$, $d_\beta = X\Sigma^{-1}y + V_\beta^{-1}\beta_0$

$$\beta_{prior} = N(\beta_0, V_\beta)$$

$$\omega \mid \beta, \sigma, \phi, \tau, y \sim N(D_\omega d_\omega, D_\omega),$$

where $D_\omega = (\Sigma^{-1} + V_\omega^{-1})^{-1}$, $d_\omega = \Sigma^{-1}(y - X\beta) + V_\omega^{-1}\omega_0$

$$\omega_{prior} = N(\omega_0, V_\omega)$$

Then, we execute Metropolis Hasting algorithm using joint posterior density to obtain σ , ϕ and τ ,

Posterior density of σ, ϕ, τ : $p(\sigma, \phi, \tau \mid \beta, y) \propto p(\sigma)p(\tau)p(\phi)p(Y \mid \beta, \sigma, \tau)$

where $\log(p(Y; \beta, \sigma, \phi, \tau)) = \text{const.} - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(y - X\beta)' \Sigma^{-1}(y - X\beta)$

I-5. Monte Carlo Experiment

i) Scenario Setting

Scenario I

Without serial correlation

Scenario II

Moderate serial correlation

Scenario III

Strong serial correlation

	Scenario I	Scenario II	Scenario III
	Spatial Autoregressive Error	Spatiotemporal Effects (Moderate Serial Autocorrelation)	Spatiotemporal Effects (Strong Serial Autocorrelation)
Data generating Process	$y_t = \alpha + \beta x_t + u + \varepsilon_t$, where $u \sim N(0, cH(\phi))$, $\varepsilon_t = \rho W \varepsilon_t + v_t$, $v_t \sim N(0, \tau^2 I_n)$	$y_t = \alpha + \beta x_t + u + \xi_t + \varepsilon_t$ where $u \sim N(0, cH(\phi))$, $\xi_t \xi_{t-1} \sim N(\rho \xi_{t-1}, \sigma^2 H(\phi))$, $\varepsilon_t \sim N(0, \tau^2 I_n)$.	$y_t = \alpha + \beta x_t + u + \xi_t + \varepsilon_t$ where $u \sim N(0, cH(\phi))$, $\xi_t \xi_{t-1} \sim N(\rho \xi_{t-1}, \sigma^2 H(\phi))$, $\varepsilon_t \sim N(0, \tau^2 I_n)$.
Regions	i ($i = 1, \dots, n$ with $n = 47$)		
Time Periods	t ($t = 1, \dots, T$ with $T = 5$)		
Fixed Parameters	$\alpha = 2$, $\beta = 5$, $c = 9$, $\rho = 0.5$, and $\tau = 0.3$		
Geographic Weight Matrix	$H(\phi)$: Correlation matrix where each element is $\rho_G(s_i - s_j; \phi) = \exp(-\ s_i - s_j\ ^2 / \phi^2)$, where $\phi = 1$ and $\ s_i - s_j\ $ is a geographical distance between region i and j . $W = H(h)$ with $h = 2$		
Explanatory Variable	x_{it} is generated by uniform distribution on $[t - 1, t + 1]$. The mean value of x_{it} is t ($t = 1, \dots, T$).		

ii) Monte Carlo Experiment Result

	(Scenario I) Spatial Autoregressive Error			(Scenario II) Spatiotemporal Effects ($\delta=0.5$)			(Scenario III) Spatiotemporal Effects ($\delta=1$)		
	Coefficient (True: $\beta=5$)	Standard Deviation (True: $\tau=0.3$)	Forecasting MSE	Coefficient (True: $\beta=5$)	Standard Deviation (True: $\tau=0.3$)	Forecasting MSE	Coefficient (True: $\beta=5$)	Standard Deviation (True: $\tau=0.3$)	Forecasting MSE
Ordinary Panel (OP)									
Mean	5.0000	0.3274	0.1345	5.0028	0.5518	0.4595	5.0013	0.5714	0.6411
Standard Error	0.0112	0.0453	0.0754	0.0473	0.0396	0.1230	0.0625	0.0470	0.1871
SAC									
Mean	4.9968	0.3049	0.1368	4.9998	0.5109	0.4648	5.0017	0.5285	0.6491
Standard Error	0.0300	0.0349	0.0756	0.0537	0.0320	0.1271	0.0648	0.0388	0.2038
SDM									
Mean	4.9995	0.2709	0.1384	5.0014	0.4575	0.4615	5.0038	0.4737	0.6349
Standard Error	0.0302	0.0295	0.0778	0.0538	0.0286	0.1252	0.0661	0.0347	0.1851
DSE-AR									
Mean	5.0001	0.2335	0.1463	5.0010	0.3604	0.4152	5.0007	0.3454	0.4181
Standard Error	0.0218	0.0103	0.0854	0.0360	0.0238	0.1110	0.0403	0.0222	0.1128
DSE-RW									
Mean	5.0001	0.2383	0.1843	5.0005	0.3768	0.4469	5.0006	0.3490	0.4061
Standard Error	0.0264	0.0109	0.1117	0.0393	0.0261	0.1229	0.0405	0.0235	0.1085

I-6. Estimation Results

Table 4. Estimation Results, Prediction and Forecast Performance (Time Periods in the Model: 1988, 1993, 1998, 2003, 2008, 2013)

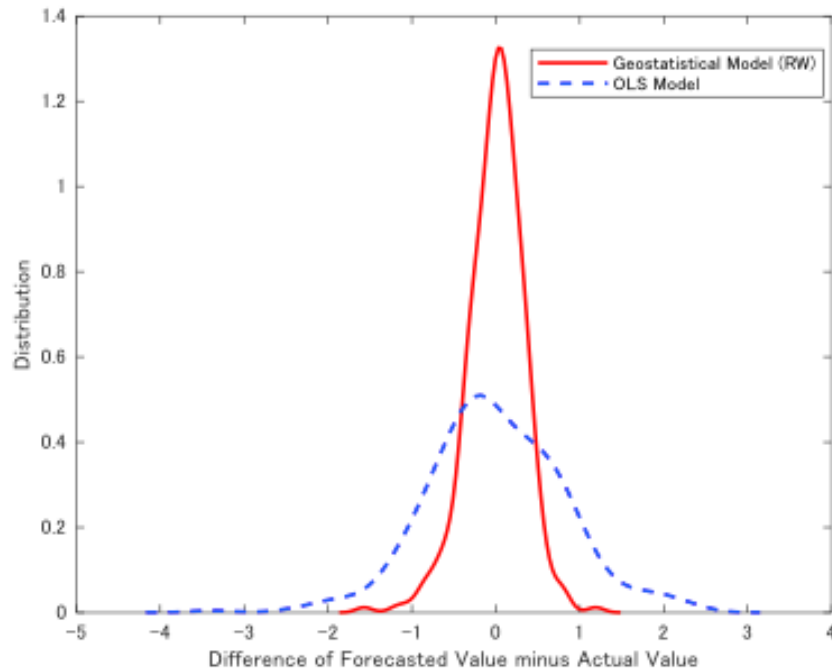
Explanatory Variables	OP		SAC		SDM		DSE-AR		DSE-RW	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Age 0-29	0.9402	0.5485	0.8239	0.4552	1.0422	0.4555	0.3455	0.5053	0.3704	0.5026
Age 30-64	0.8528	0.5602	0.7115	0.4745	0.7628	0.4764	0.4392	0.5182	0.4625	0.5153
Age 65-74	0.9797	0.5522	1.0689	0.4642	1.2597	0.4630	0.6388	0.5117	0.6461	0.5087
Age 75-	1.1766	0.6027	1.1486	0.5292	1.2421	0.5418	0.5005	0.5675	0.5110	0.5649
Age 0-29 (5-year lag)	0.8288	0.5603	0.9904	0.4695	1.1081	0.4676	0.6459	0.5033	0.6315	0.5081
Age 30-64 (5-year lag)	0.5926	0.5753	0.7208	0.4868	0.9645	0.4906	0.4036	0.5146	0.4022	0.5159
Age 65-74 (5-year lag)	0.1911	0.5684	0.0764	0.4909	0.1045	0.4846	0.2457	0.5207	0.2526	0.5218
Age 75- (5-year lag)	1.2489	0.6057	1.2655	0.5271	1.3721	0.5289	1.4290	0.5584	1.4314	0.5612
Time Trend ^a	0.1465	0.0610	0.1693	0.0648	0.2264	0.0711	-0.1521	0.1787	-0.1487	0.1777
Parameters for the SAC and SDM										
ρ^b	-	-	0.0529	0.1185	0.3423	0.1806	-	-	-	-
λ^b	-	-	0.7627	0.1373	-	-	-	-	-	-
Parameters for Error Structure										
τ	0.7485	-	0.4721	0.0336	0.3699	0.0312	0.5524	0.0423	0.5519	0.0429
σ	-	-	-	-	-	-	0.8088	0.0589	0.7939	0.0579
ϕ	-	-	-	-	-	-	0.8111	0.0987	0.7809	0.1022
δ	-	-	-	-	-	-	0.9828	0.0159	-	-
MSE for Future Values (2018)	2.7177		2.7827		2.6859		1.3959		1.3486	

^a Time Trend is the ordered set of natural numbers, which is $t = (1, 2, 3, \dots, 6)$, that measures the time span between observations. the Time Trend value is advanced from 6, which represents the most recent year in the data period (2013), to 7, which represents the time period (2018) when we forecast one period ahead of data period.

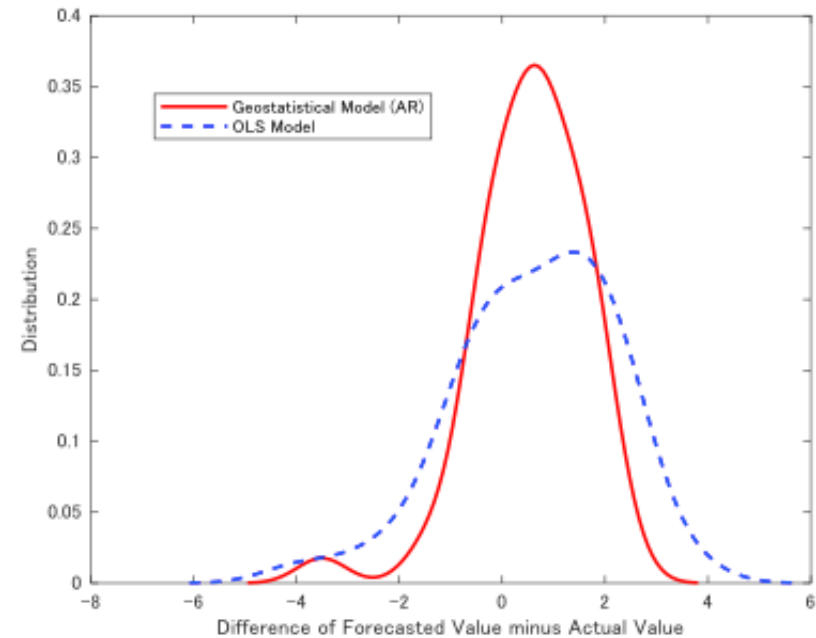
^b ρ and λ are defined in the spatial econometric models (the SAC and SDM) specified in Appendix 2.

Comparison of the distribution of the difference in actual values and predicted values between the ordinary the geostatistical panel estimation

Performance of Current Value Prediction (DSE-RW)



Performance of Future Value Forecasting (DSE-AR)



I-7. Conclusion

What the Geostatistical Panel Model brings to regional economic analysis

* There is nothing (much) surprising in the estimates of the coefficients, etc.

If the root causes of the regional structure (mainly those that are difficult to measure) continuously affect the explained variables, it will lead to a significant improvement in the current model fit (Prediction) as well as in the future predicted values (Forecast).

* Possibility of minimizing the consideration of regional (spatial) dummy variables that have no clear theoretical underpinnings, etc.

Therefore, it may be useful to construct various forecasting models for variables affected by region and space.

In situations where spatial correlation is always assumed, such as "real estate prices," geostatistical model including the DSE, which can explicitly estimate the correlation of spatial error terms, can be used to create data-efficient estimation models, especially in situations where there is not a large number of data.

8. Expected Application Areas of DSE models

Estimation and Forecasting of:

- Regional economic and social analysis using panel data with a small number of time points and locations

e.g. utilizing regional demographic forecast as explanatory variables where the national forecast is available.
- Real estate prices in areas without a large number of data points
- Variables for which spatial correlation and direct effects can be mixed (e.g., external economic effects of vacant houses)

II. Geostatistical Model for Real Estate Price Estimation

Hedonic real estate price estimation with the spatiotemporal geostatistical model

Journal of Spatial Econometrics Vol. 4, 10, (2023), with S. Sugasawa, M. Suzuki

II-1. Motivation for research

Improve the accuracy of forecasting for hedonic pricing models with a transparent method

<Major Limitation>

Large price forecast errors, especially where rural areas where transaction data is limited

Zestimate (Zillow Estimate (50 U.S. states)): 115.5 million data (2020)

Median error (off-market transactions) 7.5%
(Zestimate error rate for homes on the market is 1.9%)

II-2. Strategy for Modeling

1) Why Considering Spatial Autocorrelation

Ordinary regression analysis model

→ In regions with a large sample size, a certain level of performance can be achieved, but in regions with a small sample size (region A), it is difficult to improve forecasting accuracy.



Considering spatial autocorrelation

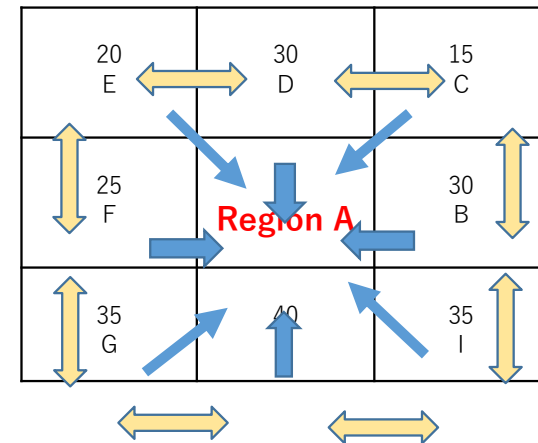
→ By introducing the spatial correlation that "region A has a close relationship with neighboring regions," can we make highly accurate forecasts not only for regions with a certain sample, but also for regions where the sample is scarce?



What model should we employ?

(Ordinary AI price estimation: Blackbox)

20 E	30 D	15 C
25 F	5 Region A	30 B
35 G	40 H	35 I



Numbers: # of Transaction

2) Spatio-temporal Estimation model

$$y = X\beta + \omega + \varepsilon, \text{ where } \varepsilon \sim N(0, \tau^2 I_n), \omega \sim N(0, \sigma^2 H(\phi, \delta)) \quad (1)$$

where $y = (y_1, \dots, y_n)^T$ is the n -dimensional vector of real estate price; X denotes the $n \times p$ matrix of the explanatory variables; I_n is the $n \times n$ identity matrix; ε and ω are mutually independent error terms and spatiotemporal effects, respectively; and $H(\bullet)$ is the spatiotemporal correlation matrix in which the (i, j) -element includes a spatial correlation of $\rho_S(s_i - s_j; \phi)$ as well as temporal correlation $\rho_T(t_i - t_j; \delta)$.

Spatial Correlation

$$\rho_{SGij}(s_i - s_j; \phi) = \exp\left(-\frac{\|s_i - s_j\|^2}{\phi^2}\right), \rho_{SEij}(s_i - s_j; \phi) = \exp\left(-\frac{\|s_i - s_j\|}{\phi}\right) \quad (2)$$

Temporal Correlation

$$\rho_{TGij}(t_i - t_j; \delta) = \exp\left(-\frac{|t_i - t_j|^2}{\delta^2}\right), \rho_{TEij}(t_i - t_j; \delta) = \exp\left(-|t_i - t_j| \frac{1}{\delta}\right) \quad (3)$$

Space * Time: Spatio-Temporal Correlation

3) Performance of the Estimation Model

Current Sample Data is almost fully explained by the model.


 Is it performing good enough for “forecasting”?

Table 3 (continued)

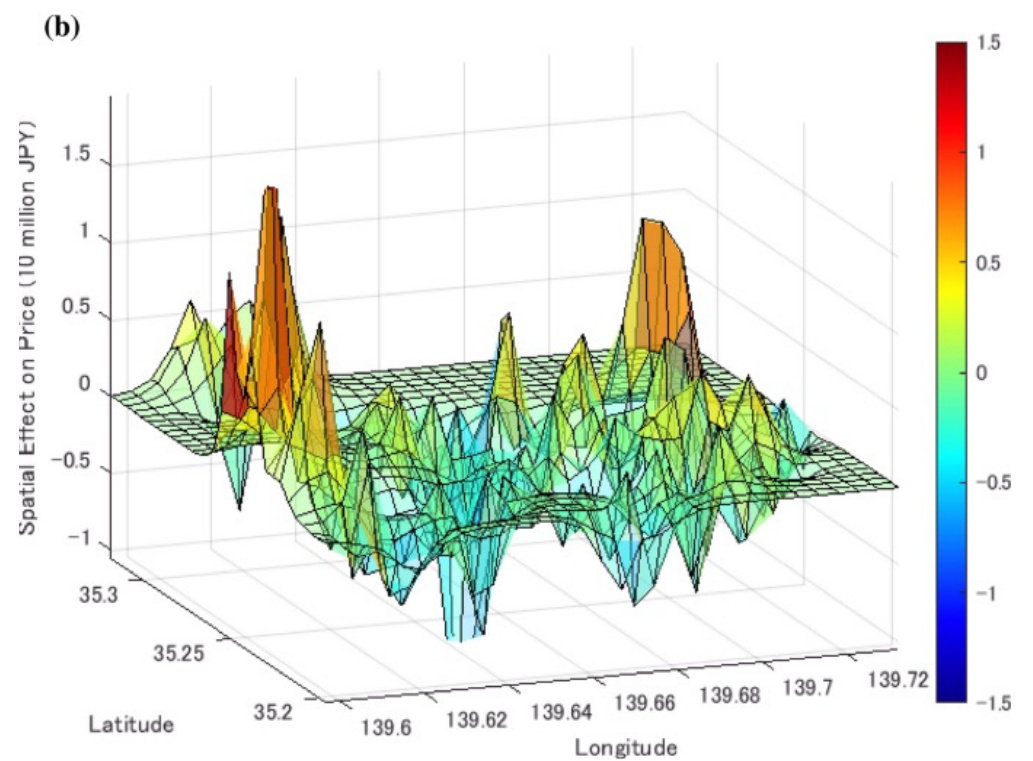
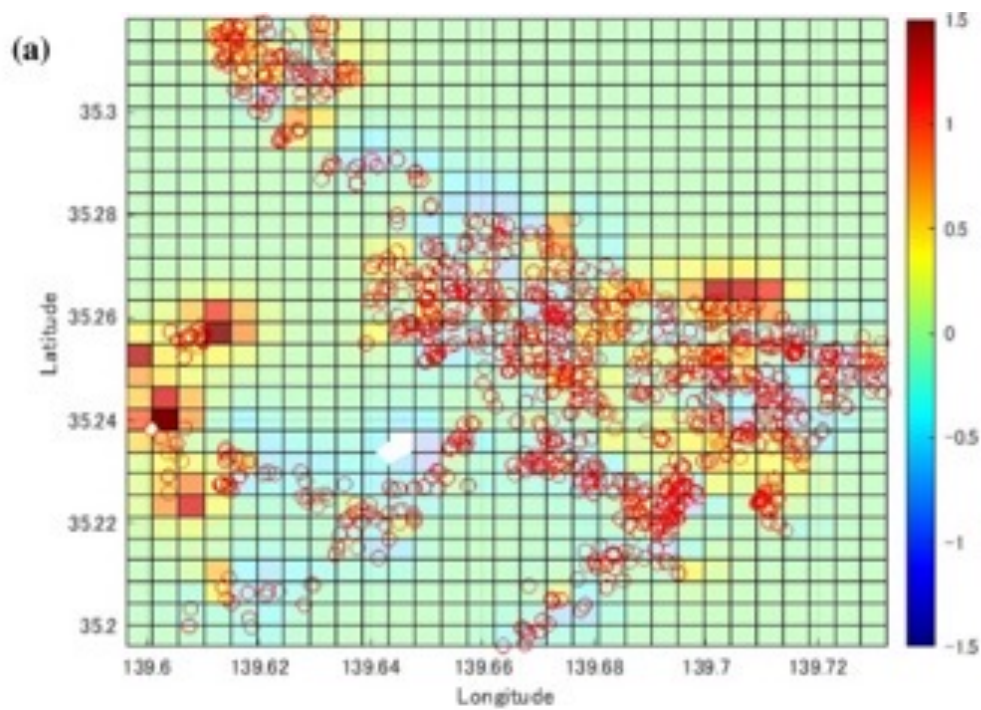
Dependent variable: Transaction price [Ten thousand JPY]

Explanatory variables	(1)				(2)				(3)															
	No time dummy, No regional dummy								No time dummy & No regional dummy								No time dummy & With regional dummy							
	With less property-level spatial variable								With full property-level spatial variable								With full property-level spatial variable							
	OLS		Geostatistical Model		OLS		Geostatistical Model		OLS		Geostatistical Model		OLS		Geostatistical Model									
Estimate	S. E	Estimate	S. E	Estimate	S. E	Estimate	S. E	Estimate	S. E	Estimate	S. E	Estimate	S. E	Estimate	S. E									
δ	-	-	16.558	(9.143)	-	-	12.227	(5.408)	-	-	2.392	(0.736)												
<i>Performance of predicted values</i>																								
R^2 (in-sample)	0.603		0.987		0.653		0.989		0.861		0.997													
MSE (in-sample)	0.474		0.015		0.414		0.013		0.165		0.004													
R^2 (out-of-sample)	0.647		0.728		0.694		0.734		0.682		0.695													
MSE (out-of-sample)	0.432		0.333		0.375		0.326		0.389		0.373													

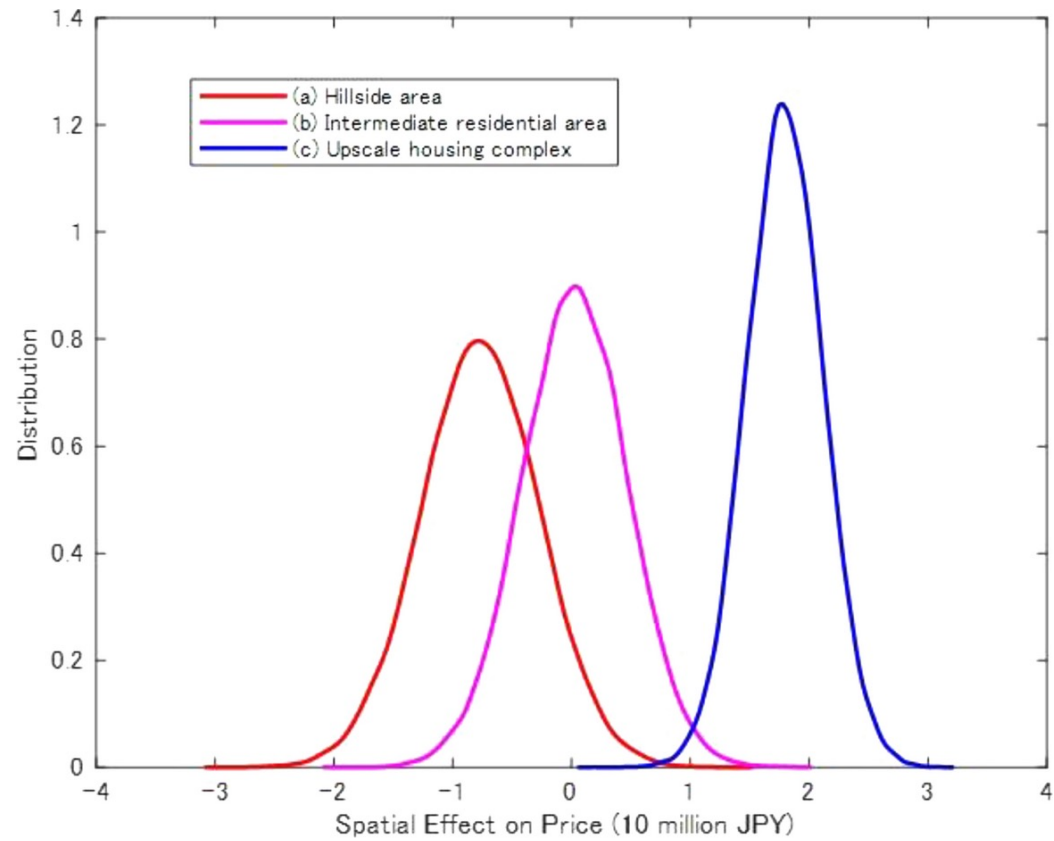
Slightly better performance than throwing in all the district dummy variables!

II-3 “Byproducts” of the model

i) Real Estate Price Surface



ii) Real Estate Price Distribution (Bayesian Posterior Density)



Upscale Housing Area



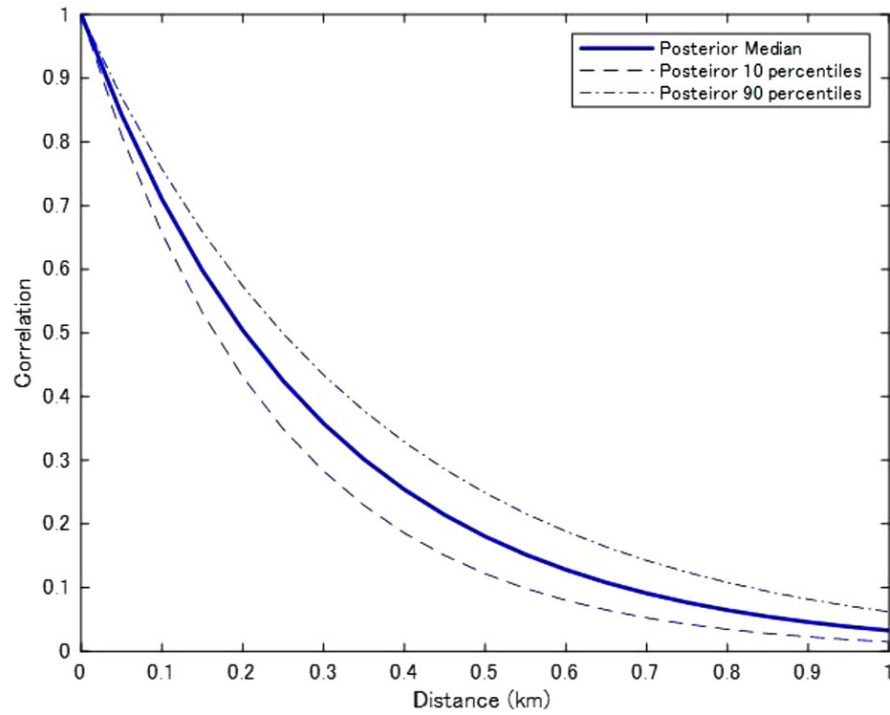
Source: Google Street View

Hillside Area (Yato)



Source: Kanagawa Newspaper (Kanaroko)

iii) Spatial Decay Function



Distance (km)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
10%tile	0.656	0.431	0.283	0.186	0.122	0.080	0.053	0.034	0.023	0.018
Median	0.710	0.504	0.358	0.254	0.180	0.128	0.091	0.065	0.046	0.039
90%tile	0.757	0.573	0.434	0.329	0.249	0.188	0.143	0.108	0.082	0.071

II-4. Limitation

Computing requirements of M-H algorithm

$$\log(p(Y; \beta, \sigma, \phi, \tau)) \propto \frac{1}{2} \log(|\Sigma \cdot s_{tu}^2|) - \frac{1}{2} ((y - X\beta) \cdot s_{tu})^T (\Sigma \cdot s_{tu}^2)^{-1} ((y - X\beta) \cdot s_{tu}) \quad (7A)$$

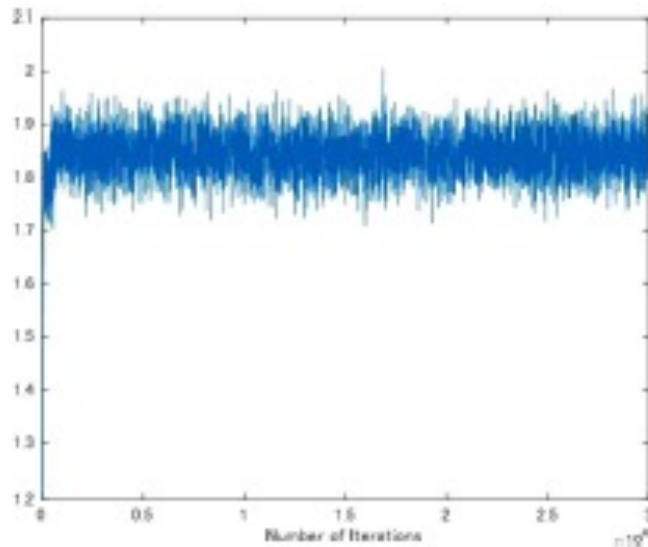


Fig. 5 The chain value of x_u for the geostatistical model estimation

“Data scale sensitive” (good for less than about 1,000)

→ We need to employ other methods for larger data sets:

Predictive Gaussian process

(Banerjee et al. 2008 and Latimer et al. 2009)

Nearest-neighbor Gaussian process

(Datta et al. 2016).

Conclusion

- When considering data on local economies, real estate, etc., it is useful to take geographic and spatial relationships and correlations among specific points in time.
- In theory, such an understanding may not have been rare, but until recently, at least, the calculations have been impossible or very time consuming.
- Nowadays, the theory and practice of spatial econometrics and spatio-temporal Geostatistics, as well as improvements in computer performance, have made such analyses possible.
- However, such computational techniques are seemed to be quite limited, at least in Japanese practitioners, and their correct understanding and application can greatly improve analytical techniques.
(More and more customized code will be available by R, Python, etc.)