Spatial Ensemble Learning for House Price Prediction

Shonosuke Sugasawa

(Faculty of Economics, Keio University)

CREI International Forum

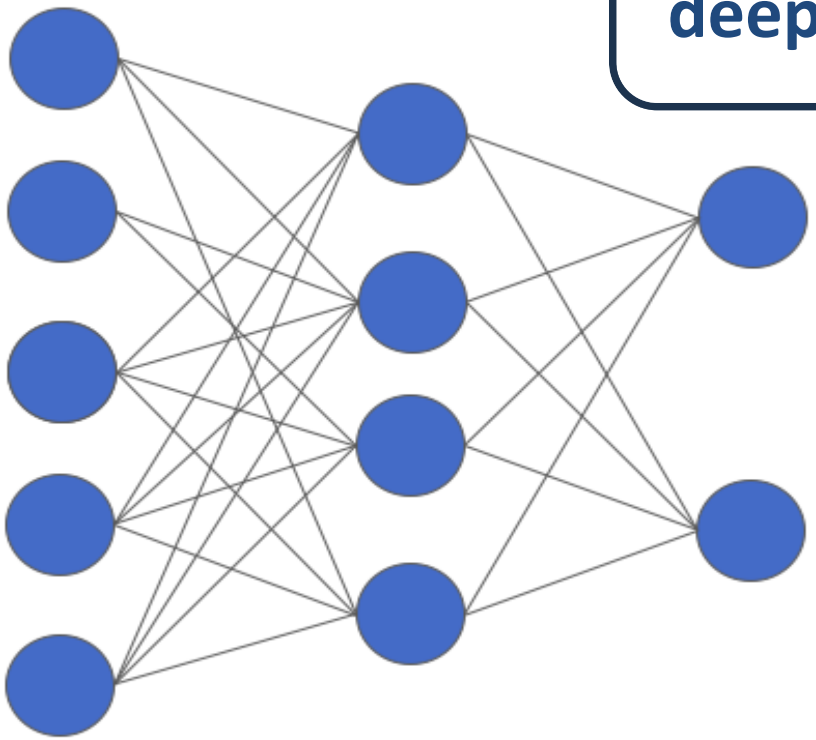
Real Estate Price Prediction by AI

- **(Existing) AI technology for evaluating real estate price**
  - Learning underlying structures based on large-scale data

  ⇨ Predict price of new real estate
- **Advantage**
  - High prediction accuracy can be achieved.
- **Disadvantage**
  - Prediction mechanism is too complex to understand.
  - It does not take account of spatial information.

Input
(Characteristics
of real estate)

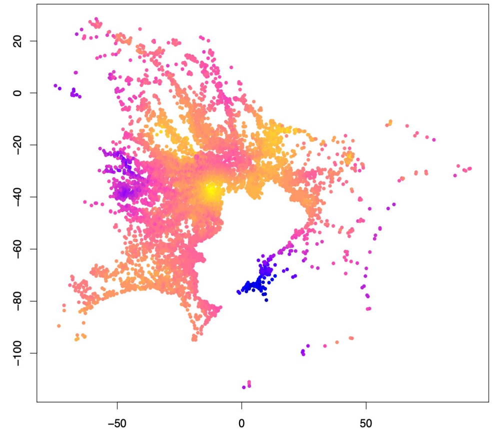
Image of deep learning

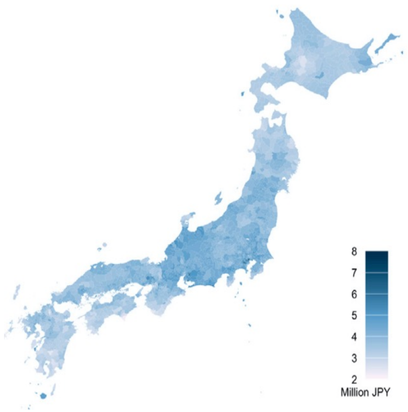


Output
(Price)

Spatio-temporal data

- **Spatio-temporal data**
  - Real estate data is observed over space and time
    - ⇨ Spatio-temporal data
  - Key characteristics:
    - ➢ **Spatial correlation / spatial heterogeneity**
    - ➢ **Temporal trend**





- (Typical) machine learning (AI) algorithm
  - Assuming independence between observations
    - ⇨ **Need (probabilistic) statistical models handling spatio-temporal correlation**

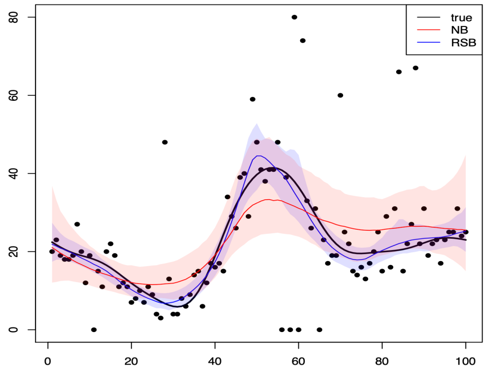
- Advantages of statistical models (e.g. hedonic models) for prediction
  - We can obtain point prediction as well as uncertainty information
  - They tend to be more interpretable than machine learning algorithm (e.g. neural network)

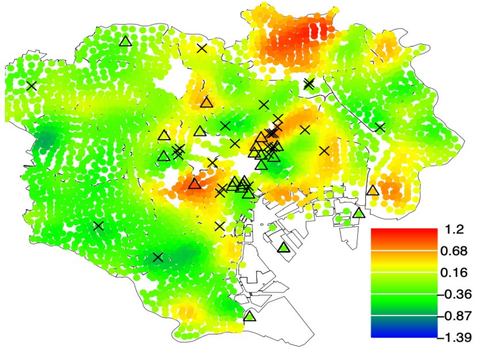
Prediction synthesis: Background

There are variety of statistical models…

- Hedonic models (linear regression)
- Additive models
- Spatial regression

etc





Question: Which models should we use for prediction?

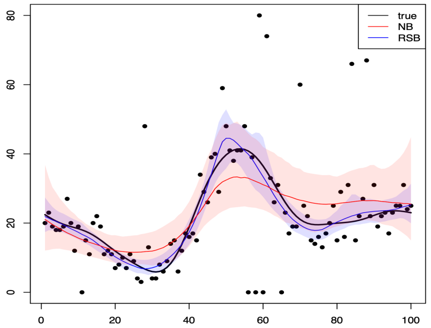
Answer: Combine all the models to get better prediction!

Spatial Ensemble Learning

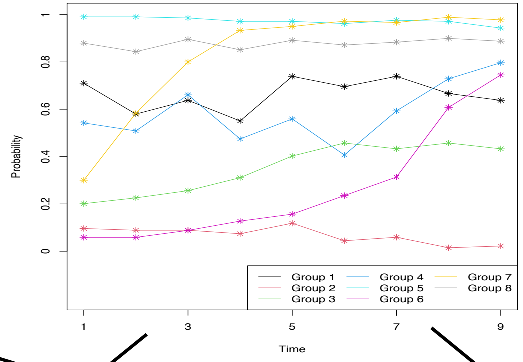
Combine multiple (whitebox) models

(Extension of existing ensemble learning to handle spatial data)

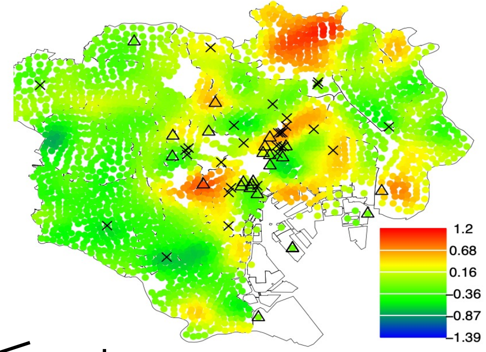
Model 1



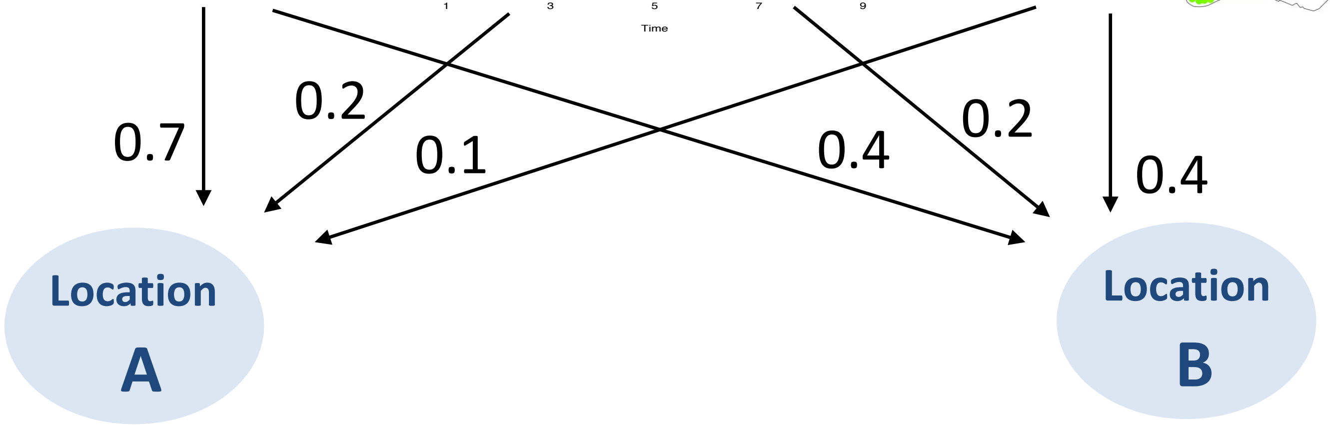
Model 2



Model 3



Model weight can vary over space.
⇨ Flexible prediction!!



Model weight is automatically learned from the data

Bayesian Spatial Predictive Synthesis (BSPS)

General form of Bayesian predictive synthesis

Prediction models at site s

$$p(y(s) \mid \Phi(s), F(s)) = \int \alpha(y(s) \mid F(s), \Phi(s)) \prod_{j=1}^{J} h_j(f_j(s)) \mathrm{d}f_j(s)$$

Apply the idea of BPS to spatial prediction!

Spatially varying coefficient model (for continuous response)

$$y(s) = \beta_0(s) + \sum_{j=1}^{J} \beta_j(s) f_j(s) + \varepsilon(s), \quad \varepsilon(s) \sim N(0, \sigma^2)$$

$$\beta_j(s) \sim \mathrm{GP}(\tau_j, h_j), \quad \text{independently for } j = 0, 1, \ldots, J$$

The model weight can vary over space
(Significant difference from existing ensemble methods)

BSPS: learning

- Settings (sampled data)
  - $s_1, \ldots, s_n \in \mathcal{S}$: sampled locations
  - Simplified notation: $y_i = y(s_i), f_{ji} = f_j(s_i),\ \varepsilon_i = \varepsilon(s_i),\ \beta_{ji} = \beta_j(s_i)$
  - **Synthesis model for sampled data**

$$y_i = \beta_{0i} + \sum_{j=1}^{J} \beta_{ji} f_{ji} + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, \ldots, n$$

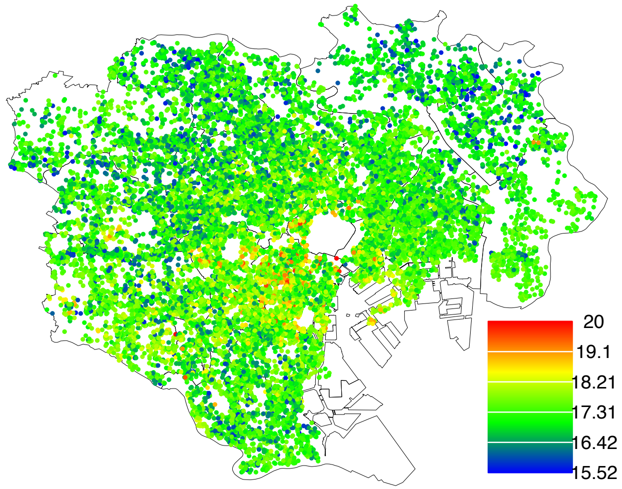
$$\beta_j \equiv (\beta_{j1}, \ldots, \beta_{jn})^\top \sim N(0, \tau_j H(h_j)), \quad j = 0, \ldots, J$$

  - $\left(H(h_j)\right)_{ii'} = \rho(s_i - s_{i'}; h_j), \quad \rho$: correlation function

⇨ Use **Markov Chain Monte Carlo algorithm** for learning model weights

Real data example

- Prediction of apartment house prices
  - Prediction of apartment prices in 23 wards in Tokyo in 2017
  - About 22,000 samples
    ⇨ leave 2000 samples as test data
  - Auxiliary information: characteristics of apartment (11 dimension) and location information (2 dimension)



- Models to be synthesized
  - Tokyo-level model : fit additive models for all the training samples
  - Ward-level mode: fit additive models ward-by-ward
  - Station-level model: fit linear regression station-by-station

model size & sample size

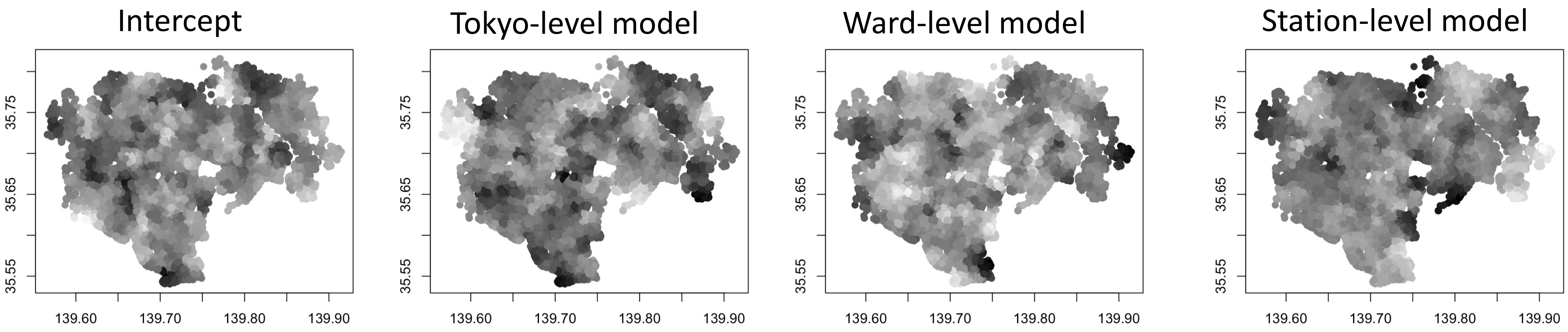
Large

Small

⇨ **Synthesize the three models through BSPS**

Real data example

Model contribution

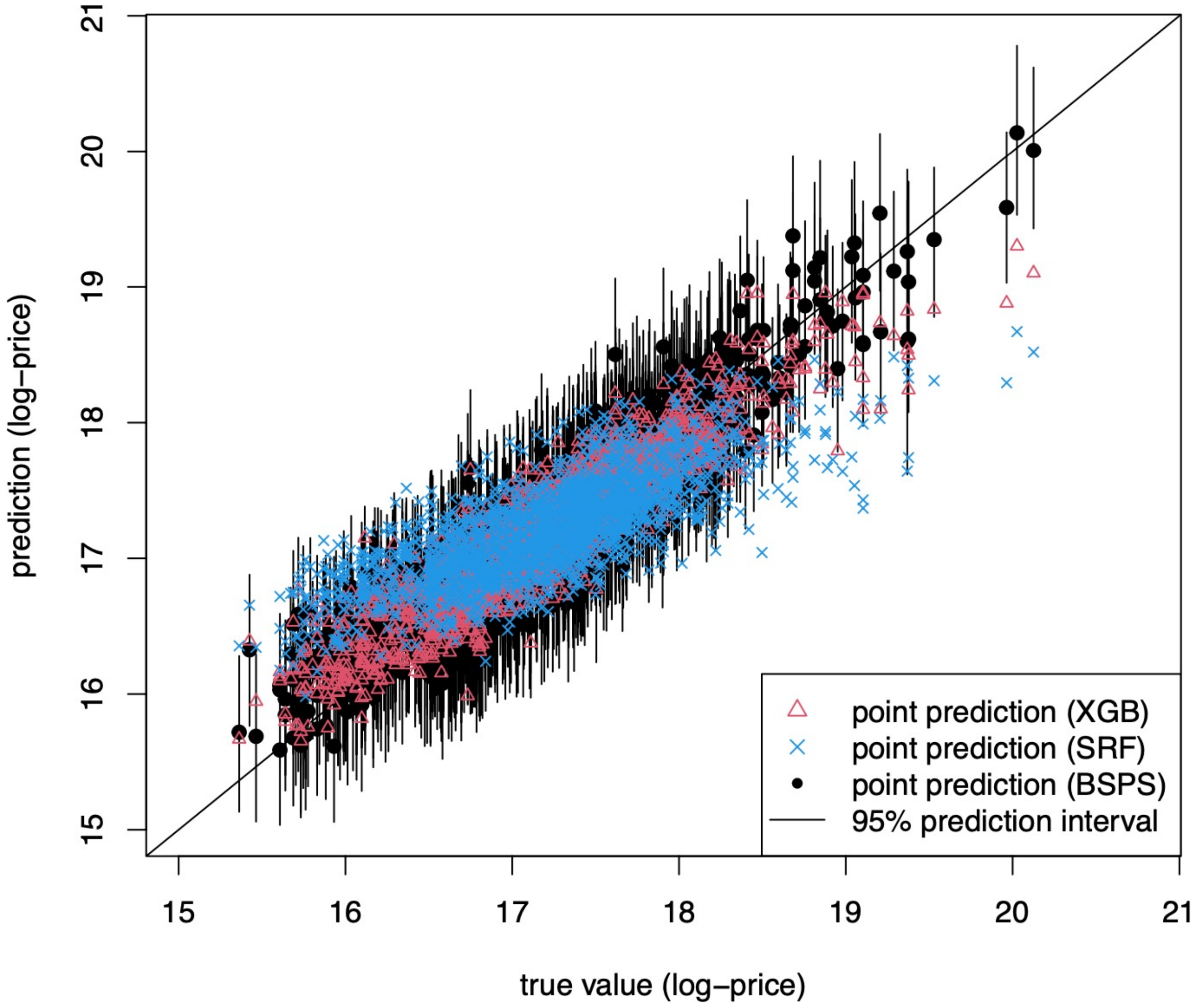


- Prediction accuracy (MSE of log-price of 2000 test samples)
  - **(proposed) SBPS: 0.240**
  - Gradient boosting tree: 0.257
  - Random forest: 0.424
  - Spatial regression: 0.268

✓ SBPS is highly interpretable

✓ Prediction process of tree boosting and random forest is complicated.

Real data example



- BSPS can also provide prediction intervals

- Coverage rate for 2000 test samples is 97.1%

Summary

- **Prediction synthesis is useful for real estate price prediction**
  - High accuracy & high interpretability
  - Applicability of wide range of prediction problems

- **Future works**
  - More scalable algorithm (which can handle 1 million observations)
  - Develop R or Python package

- **Working paper & code**
  - Cabel, D., Sugasawa, S., Kato, M., Takanashi, K. and McAlinn, K. (2022). Bayesian spatial predictive synthesis. *arXiv:2203.05197*
  - R code: https://github.com/sshonosuke/BSPS

Shonosuke Sugasawa **Email: sugasawa@econ.keio.ac.jp**