Residential price indices using different sources of information

Paulo Picchetti (FGV)

Sixteenth Meeting of the Ottawa Group, May 2019

Motivation

Need for robust hedonic house prices coefficient estimates, to plug into price index formulas (e.g.):

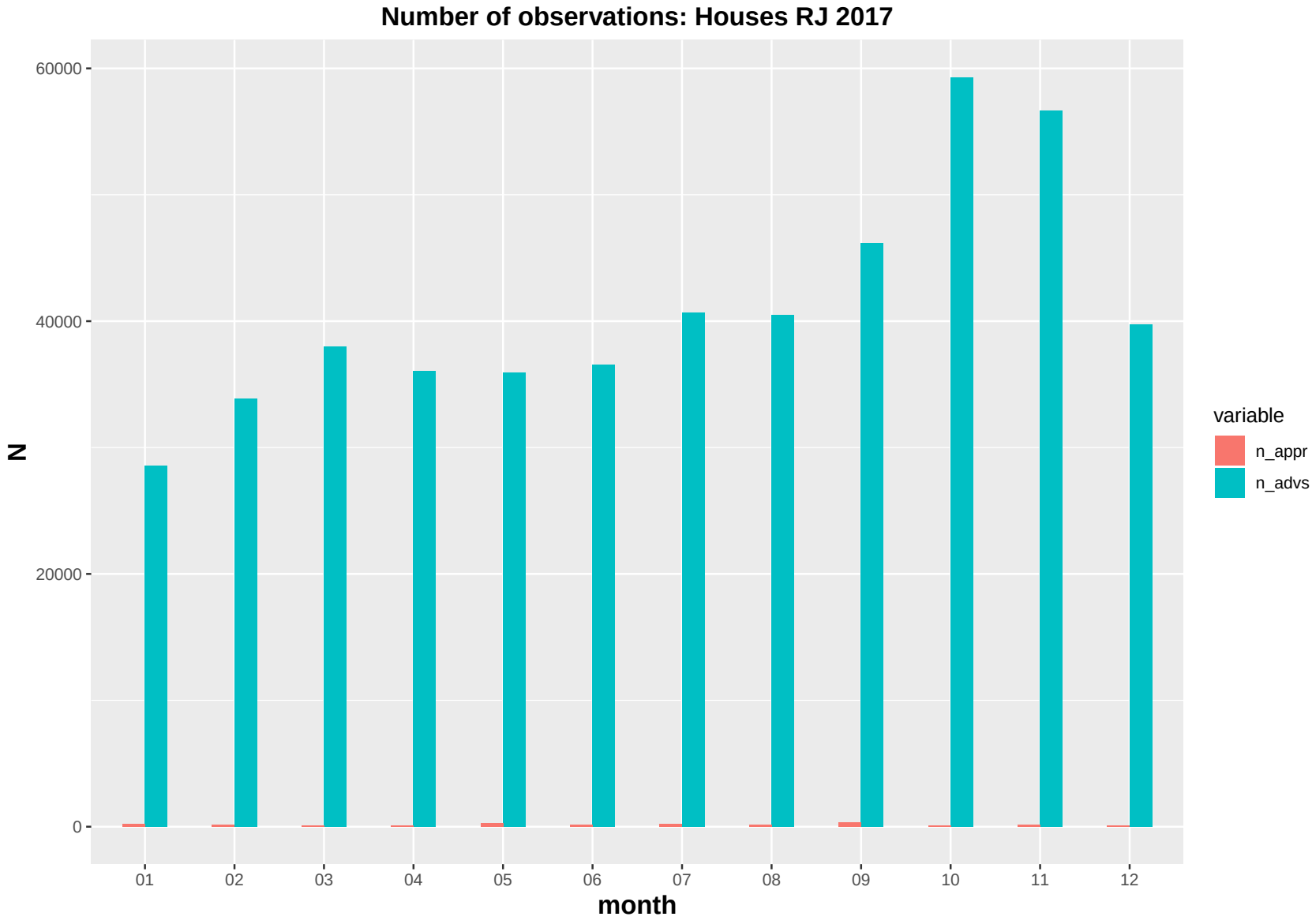
$$P_{HDIL}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t \bar{z}_k^0}{\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 \bar{z}_k^0}$$

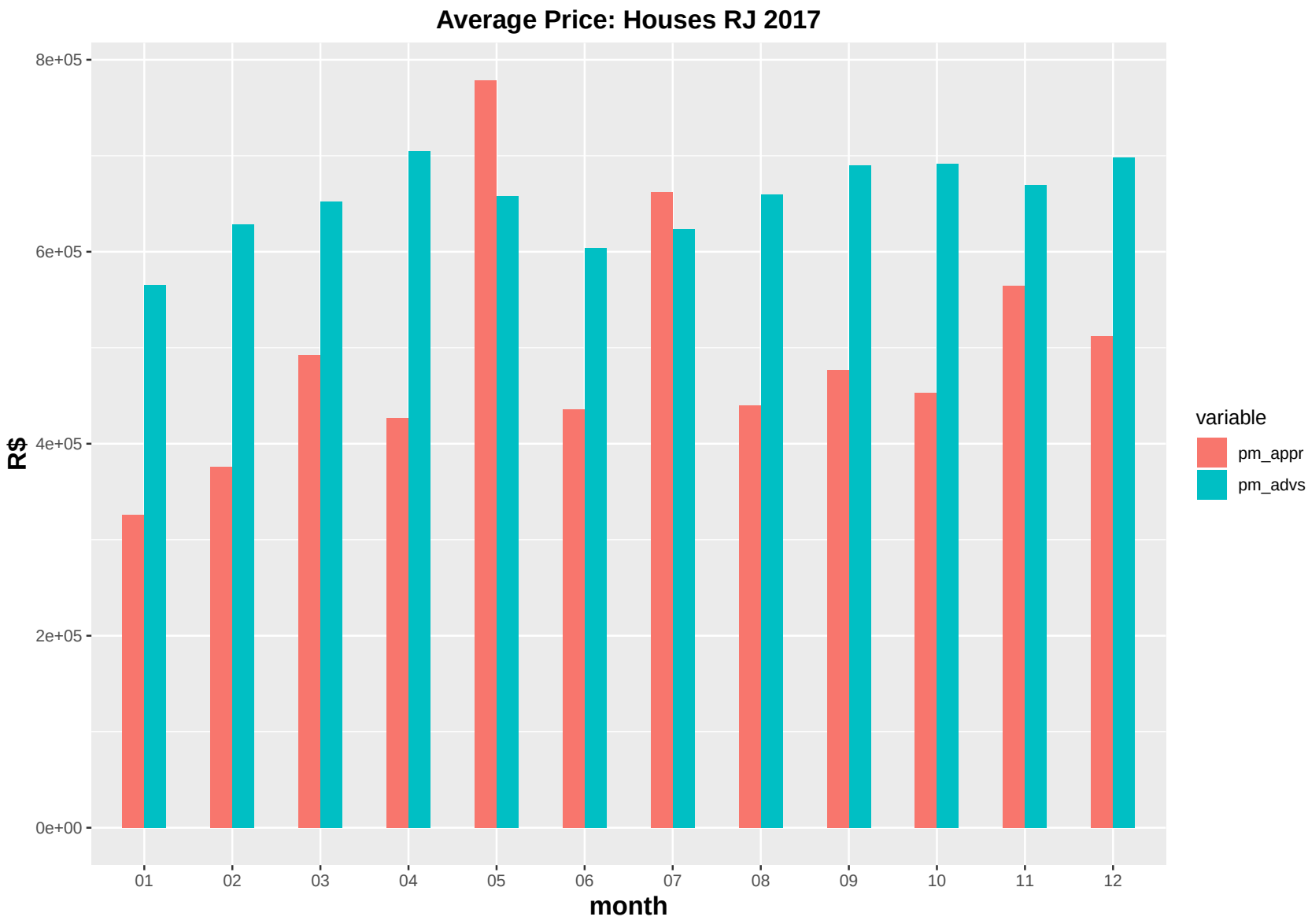
Data

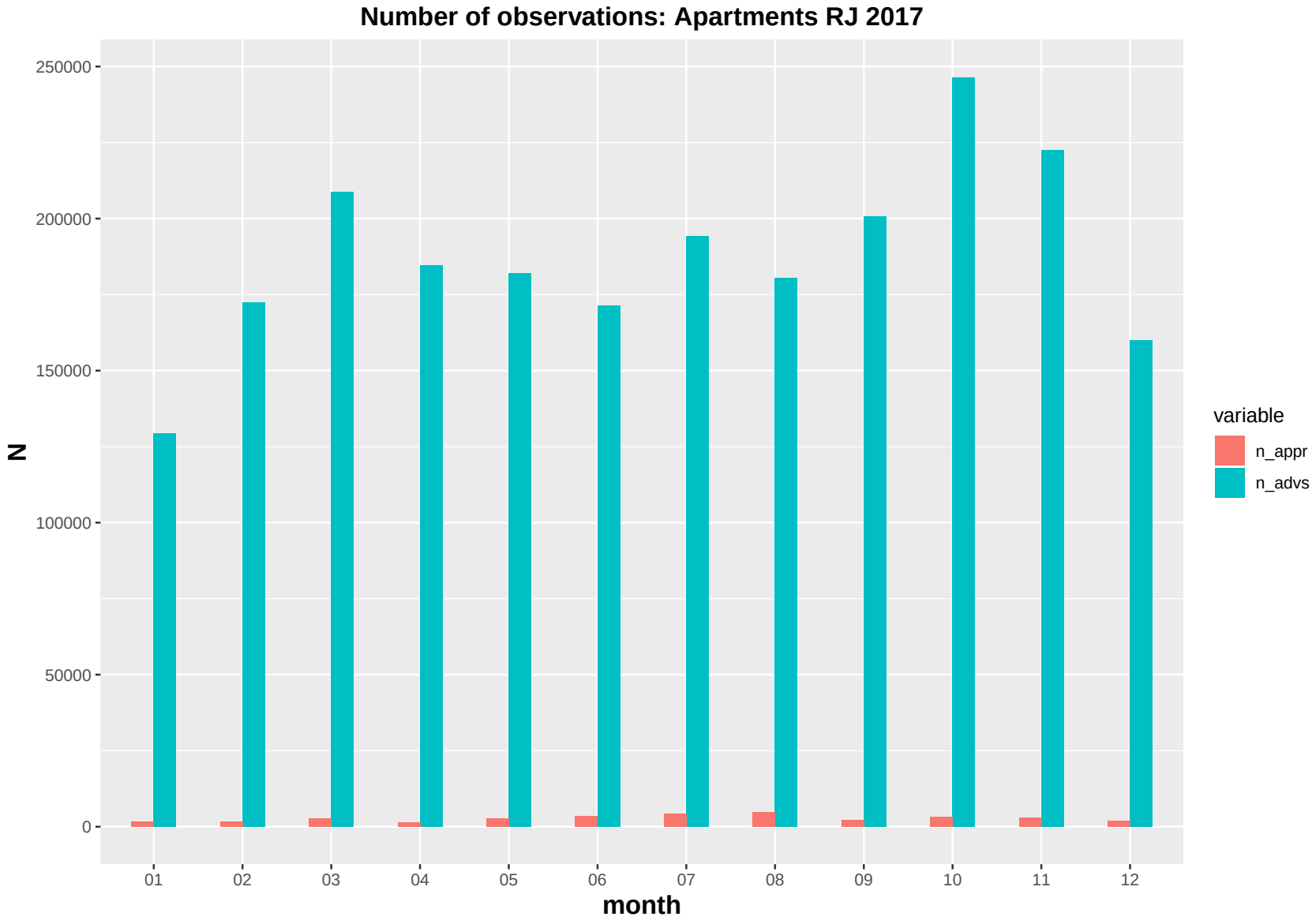
- Appraisals for loans
- Online advertisements

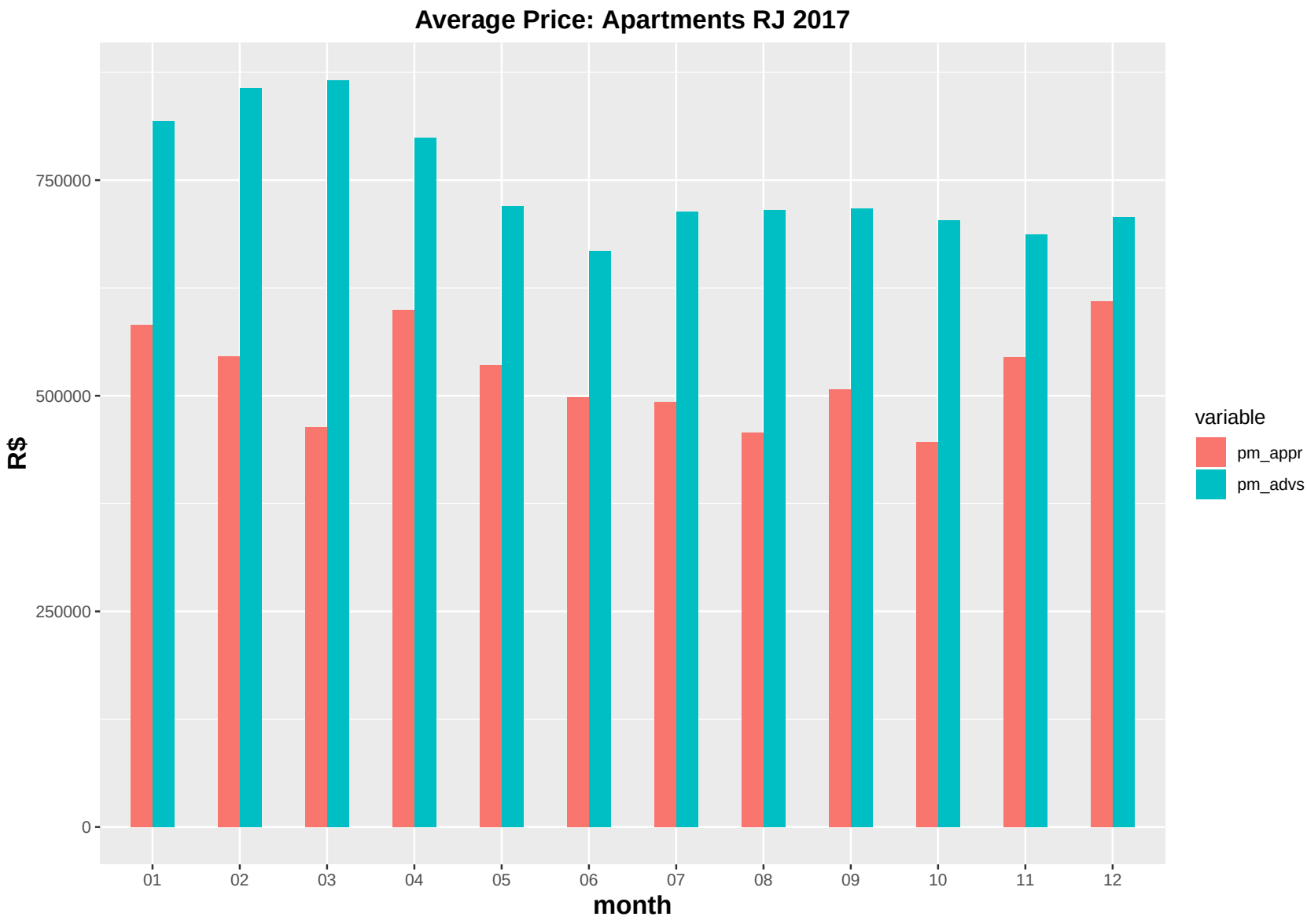
from "Handbook on Residential Property Price Indices (Eurostat(2013) - Ch 9)":

. . . although a house price index based on surveys of asking prices may be more timely, the difficulties in determining exactly how the survey information was compiled and the uncertain relationship between asking price and selling price mean that care should taken if such an index is to be used as a barometer of house prices.

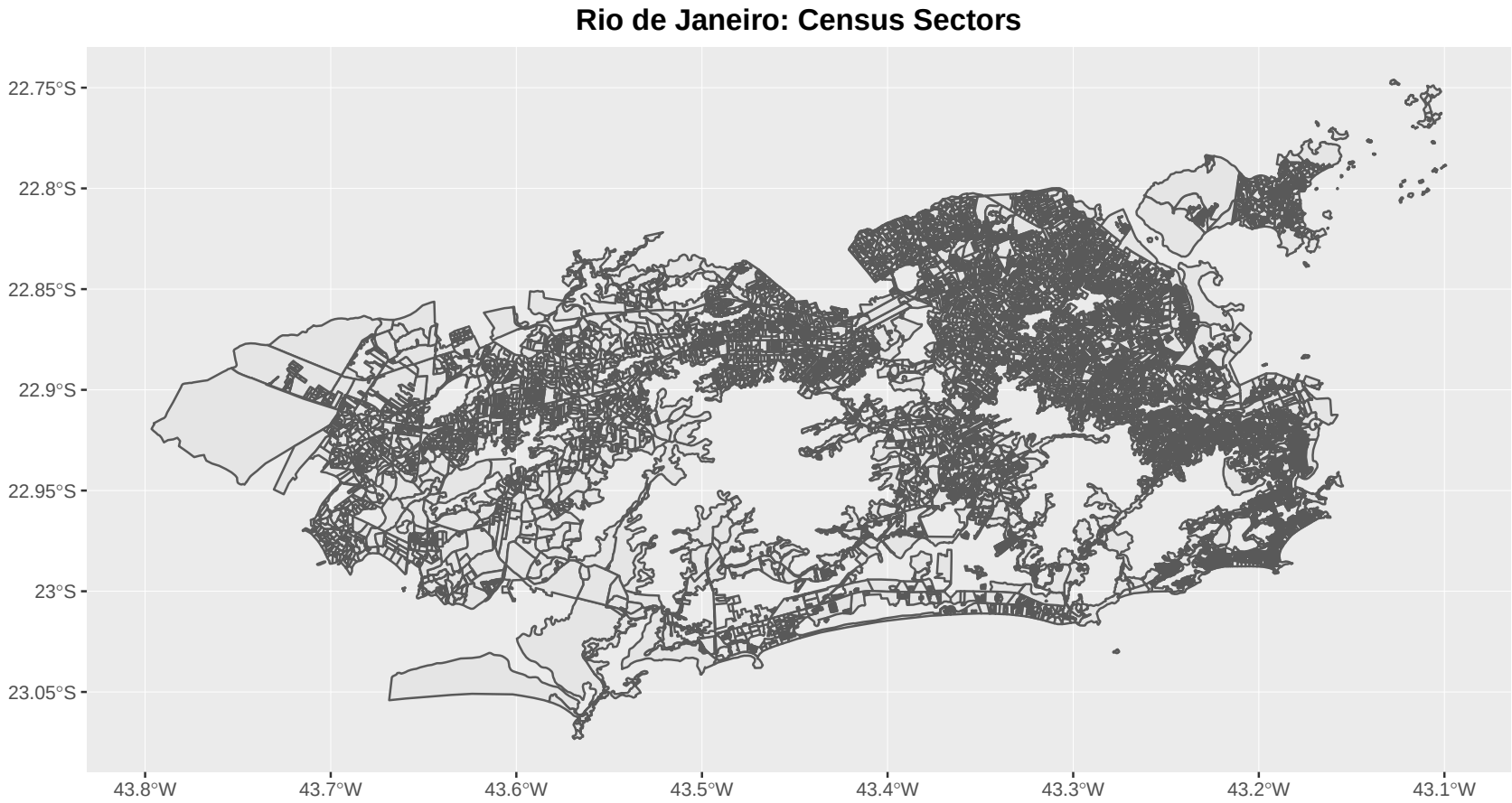

Number of observations: Houses RJ 2017
60000
40000
20000
0
N
01
02
03
04
05
06
07
08
09
10
11
12
month
variable
n_appr
n_advs


Average Price: Houses RJ 2017
8e+05
6e+05
4e+05
2e+05
0e+00
R$
01
02
03
04
05
06
07
08
09
10
11
12
month
variable
pm_appr
pm_advs

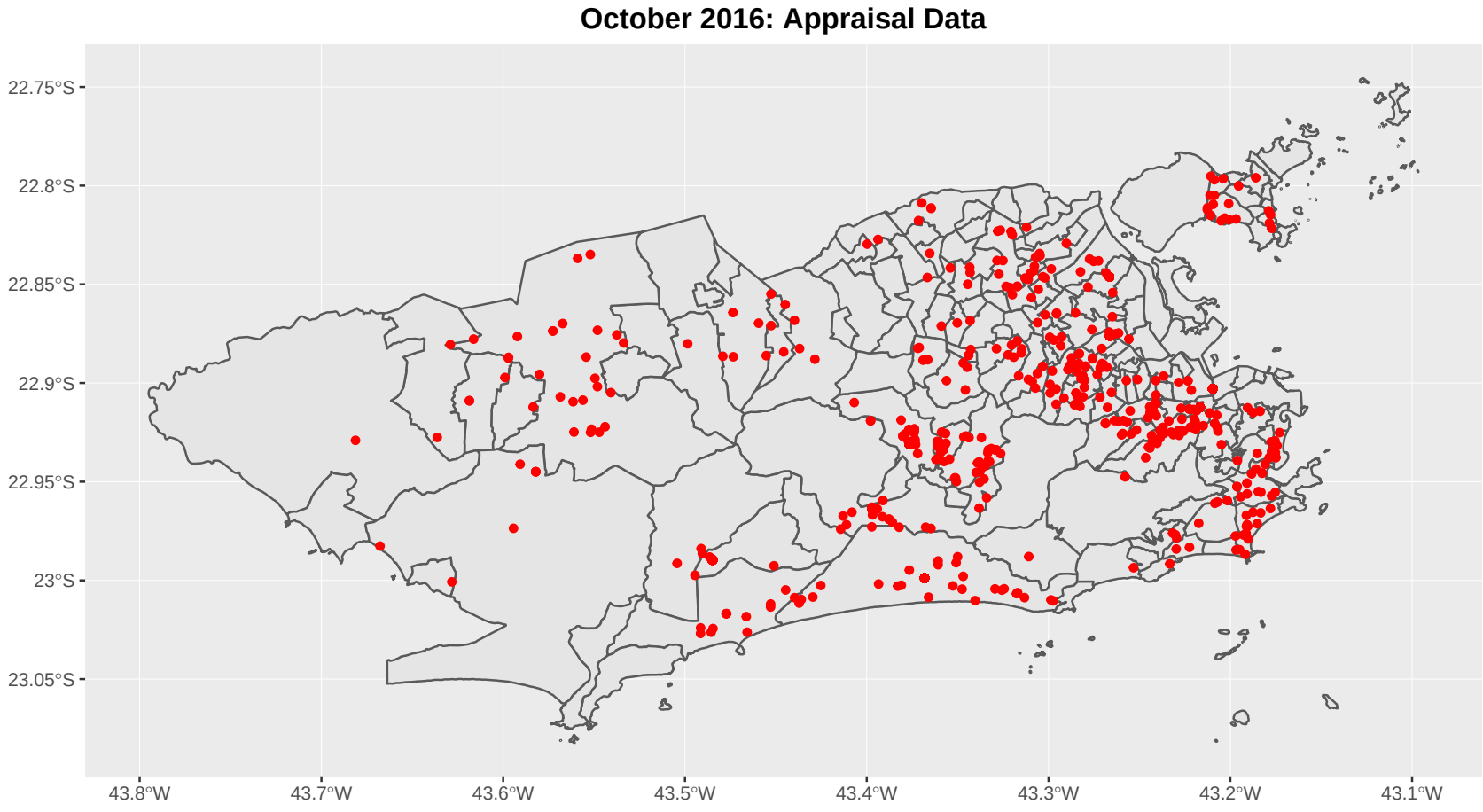

Number of observations: Apartments RJ 2017
N
250000
200000
150000
100000
50000
0
01
02
03
04
05
06
07
08
09
10
11
12
month
variable
n_appr
n_advs


Average Price: Apartments RJ 2017
R$
750000
500000
250000
0
01
02
03
04
05
06
07
08
09
10
11
12
month
variable
pm_appr
pm_advs

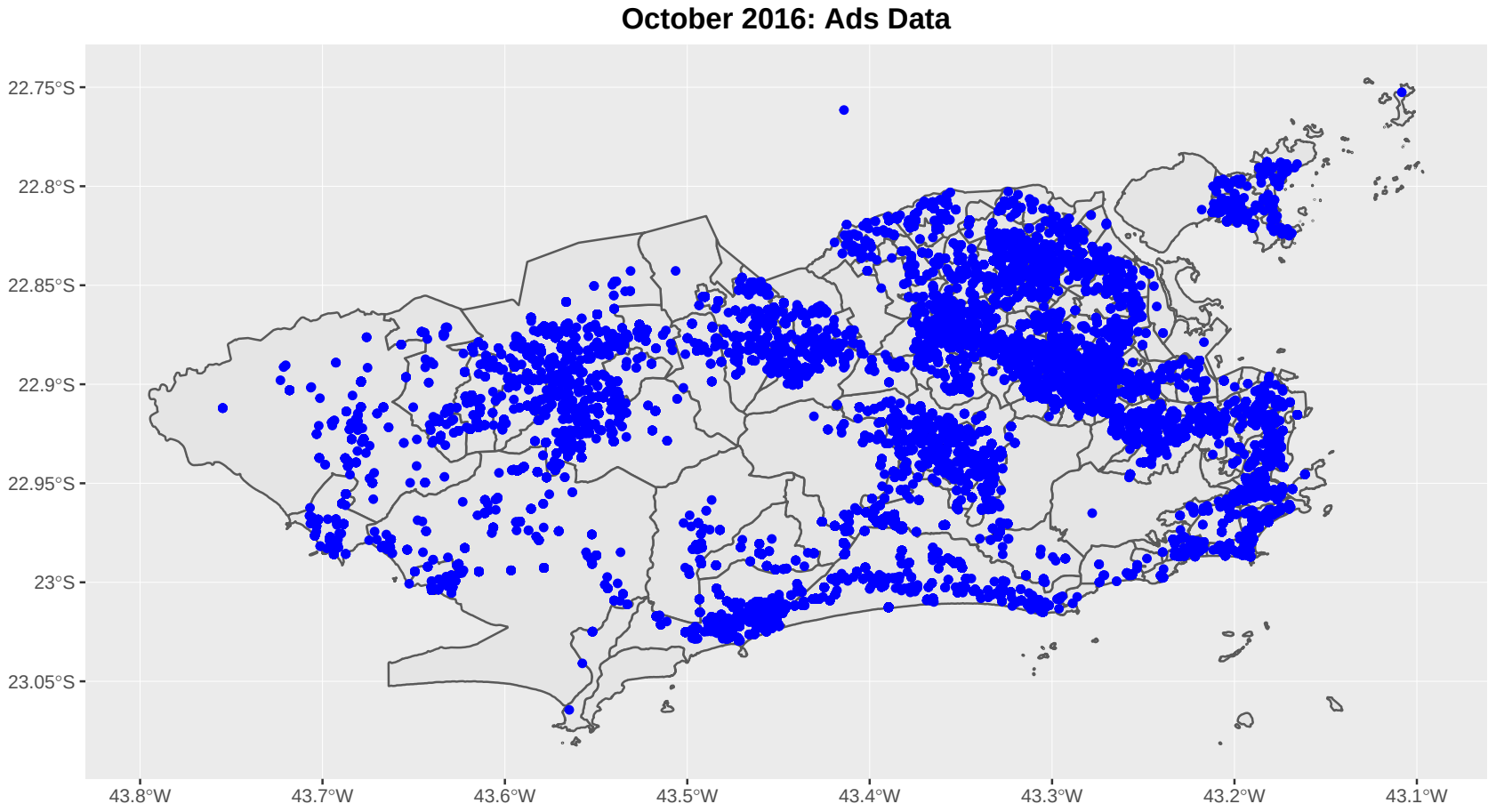
Spatial Coverage



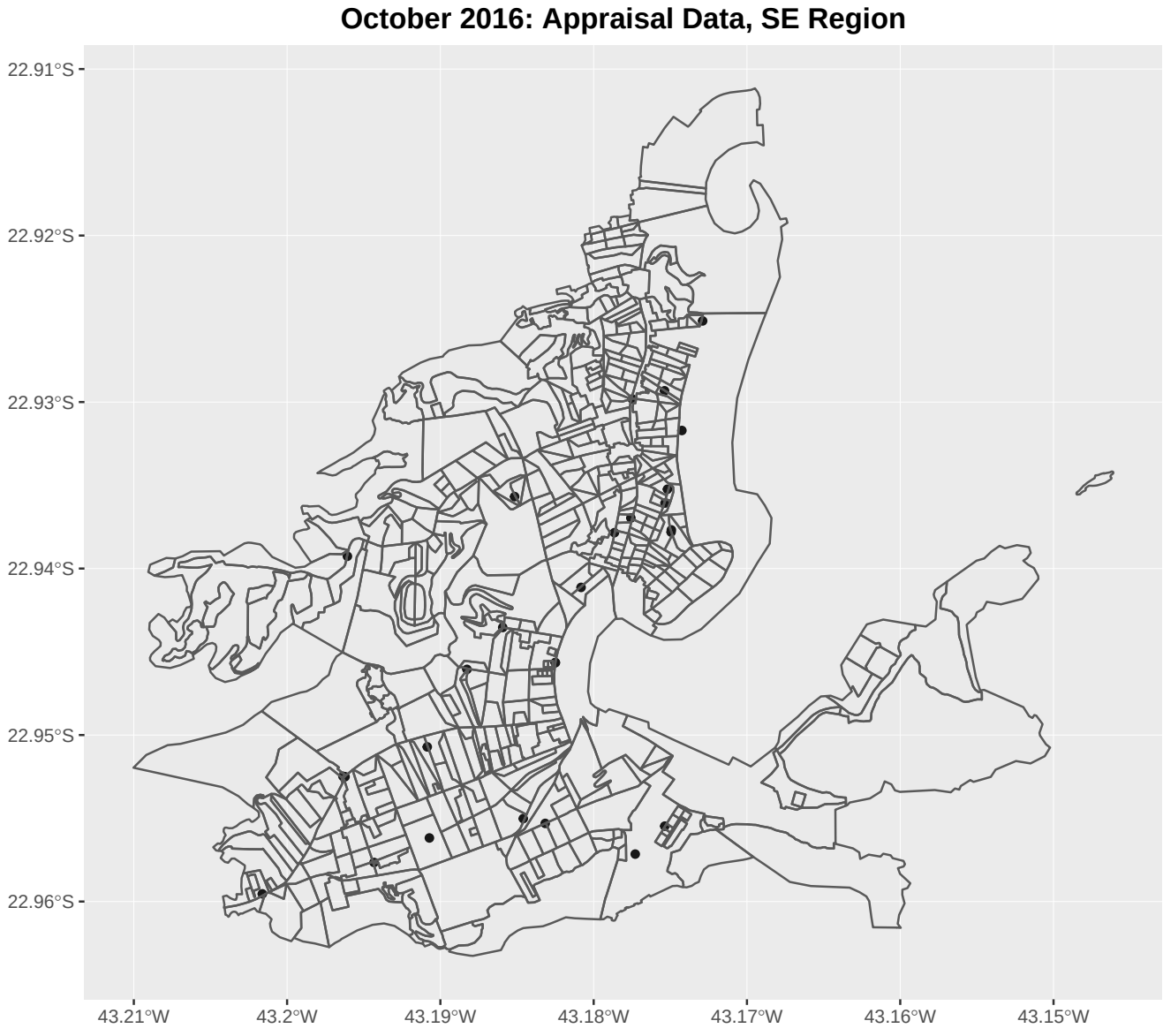
Spatial Coverage



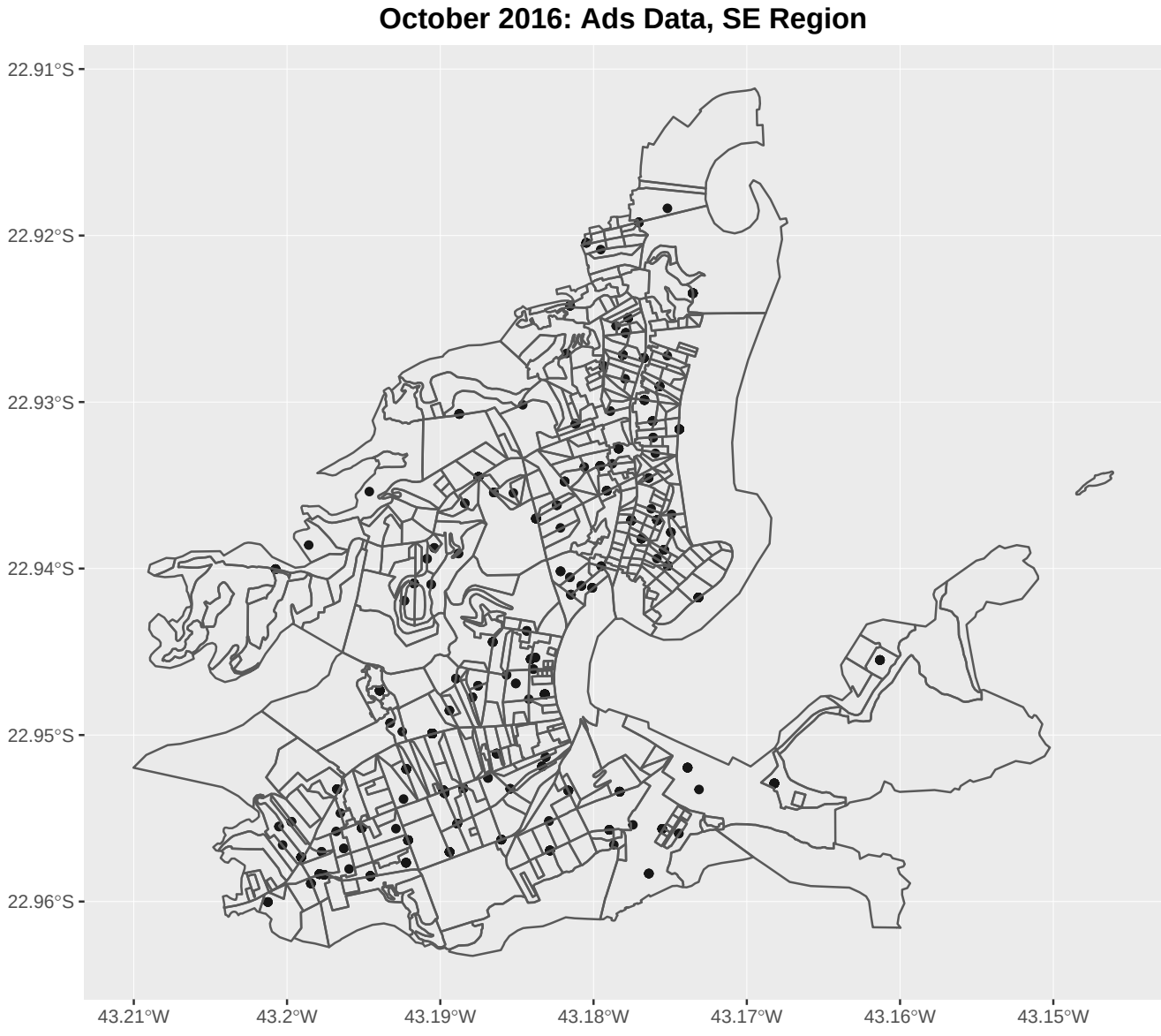
Spatial Coverage



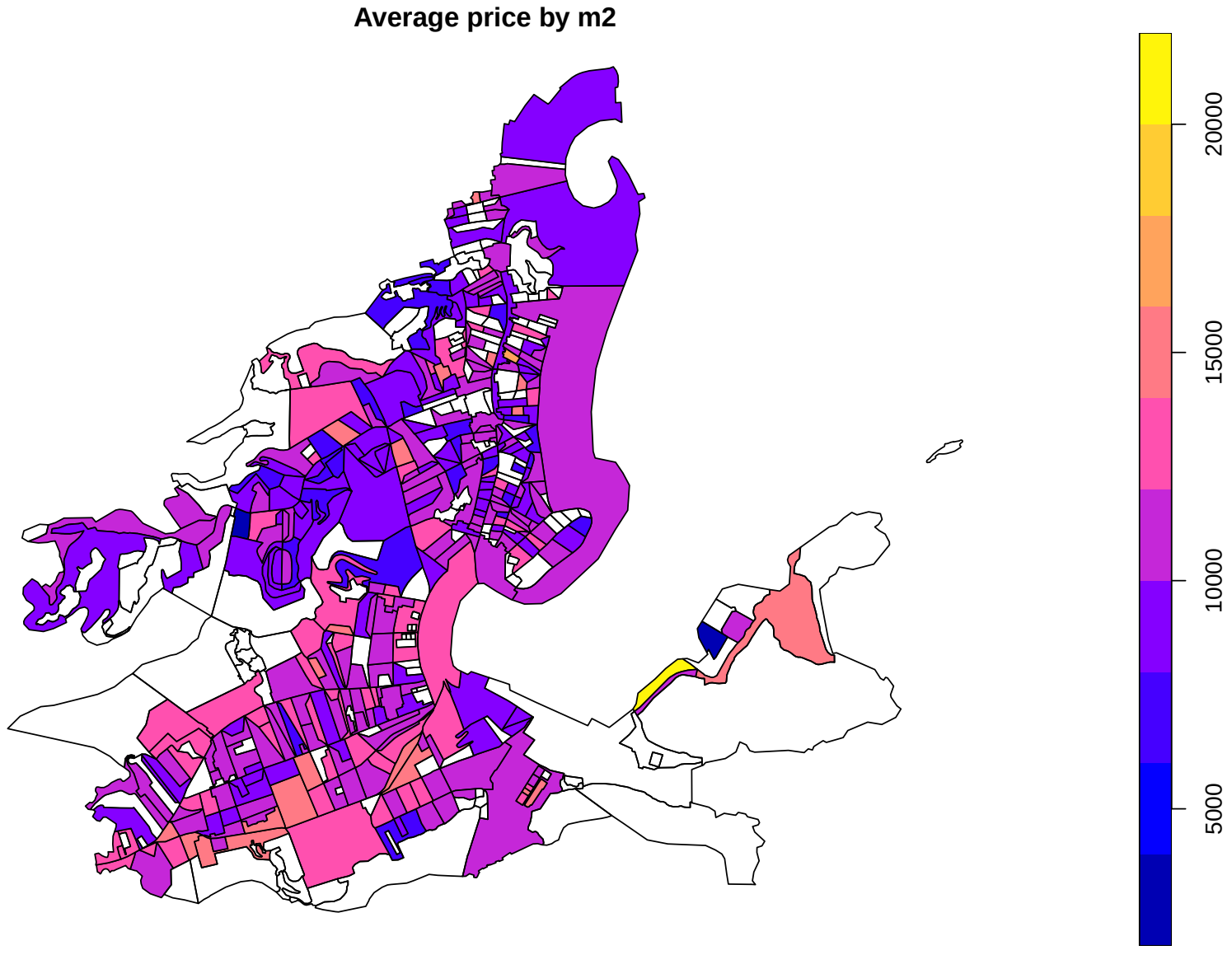
Spatial Coverage: Zooming in



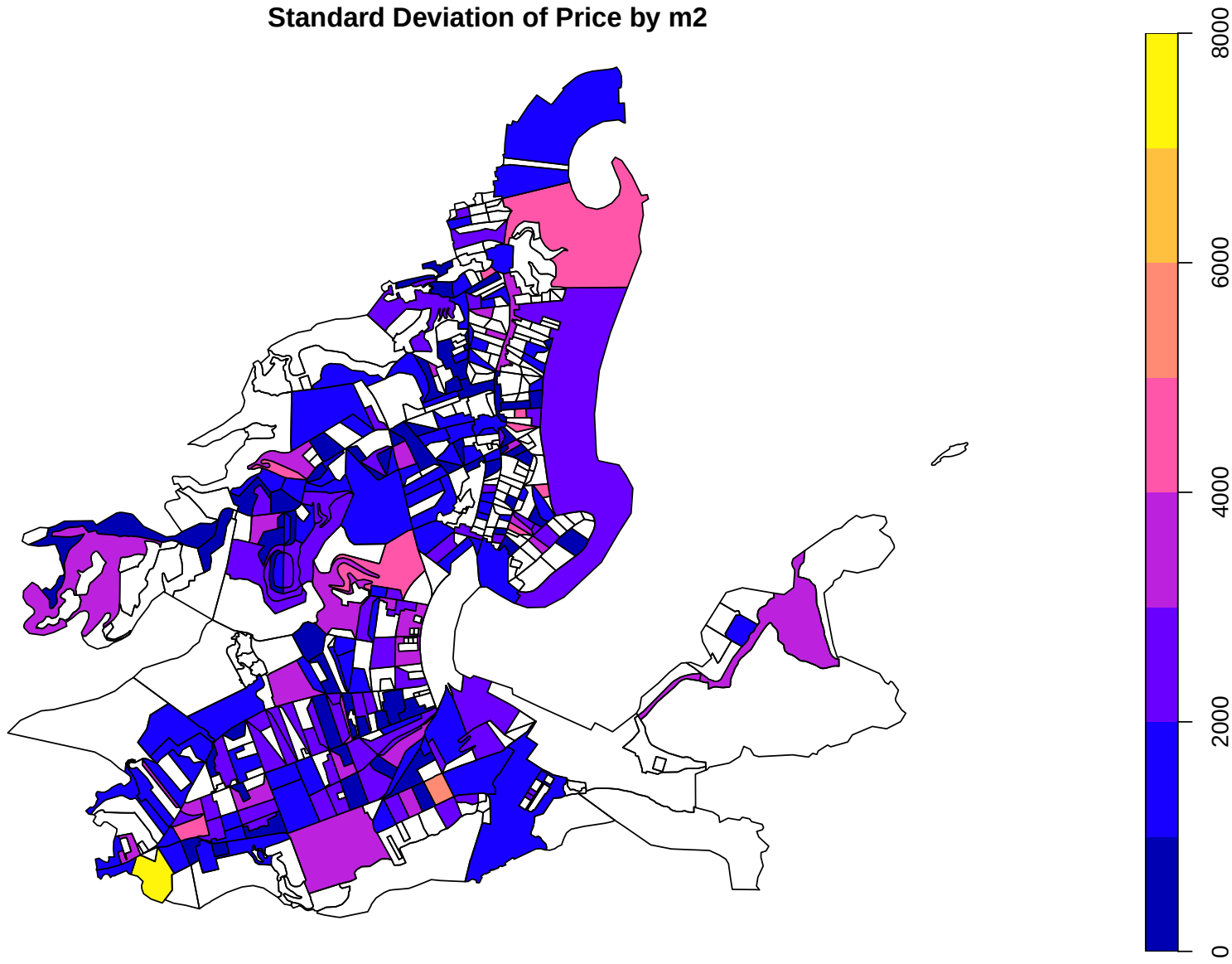
Spatial Coverage: Zooming in



Aggregation ?



Aggregation ?



Data Misalignment Model

$$Y(s) = X(s)\beta + \omega(s) + \epsilon(s)$$

where $Y(s) = [y_1(s), y_2(s)]$ is a vector of measurements (prices) for each dataset, in locations s. $\epsilon(s)$ represents measurement error, assumed distributed as gaussian, with a diagonal covariance matrix with parameters Ψ_1, Ψ_2. $X(s)\beta$ is the deterministic effect of each dwelling's attributes on the prices. Non-deterministic spatial effects are represented by a Gaussian Process $\omega(s)$, with spatial memory parameters ϕ_1, ϕ_2. When combining these datasets we have an additional parameter measuring the correlation among the values in the datasets, ρ. The combinations among the parameters ϕ_i and ρ_i define a cross spatial covariance matrix.

Data Misalignment Model

The statistical model assumes a gaussian distribution for $Y(s)$ with mean function $\mu(\mathbf{s};\beta)$ and covariance function $C(\mathbf{s}-\mathbf{s}';\boldsymbol{\theta}) = \sigma^2\rho(\mathbf{s}-\mathbf{s}';\boldsymbol{\phi})$ so that $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi})^T$.

So we have

$$\mathbf{Y}_s \Big| \beta, \boldsymbol{\theta} \sim N\left(\boldsymbol{\mu}_s(\beta), \sigma^2 H_s(\phi)\right)$$

where $\mu_s(\beta)_i = \mu(s_i;\beta)$ e $(H_s(\phi))_{ii'} = \rho(\mathbf{s}_i - \mathbf{s}_{i'};\phi)$.

Data Misalignment Model

Prediction for new localities in a Bayesian context result from the predictive distribution function.

$$f\left(\mathbf{Y}_{s'}|\mathbf{Y}_{s}\right) = \int f\left(\mathbf{Y}_{s'}|\mathbf{Y}_{s}, \beta, \theta\right) f(\beta, \theta|\mathbf{Y}_{s}) d\beta d\theta$$

Data Misalignment Model

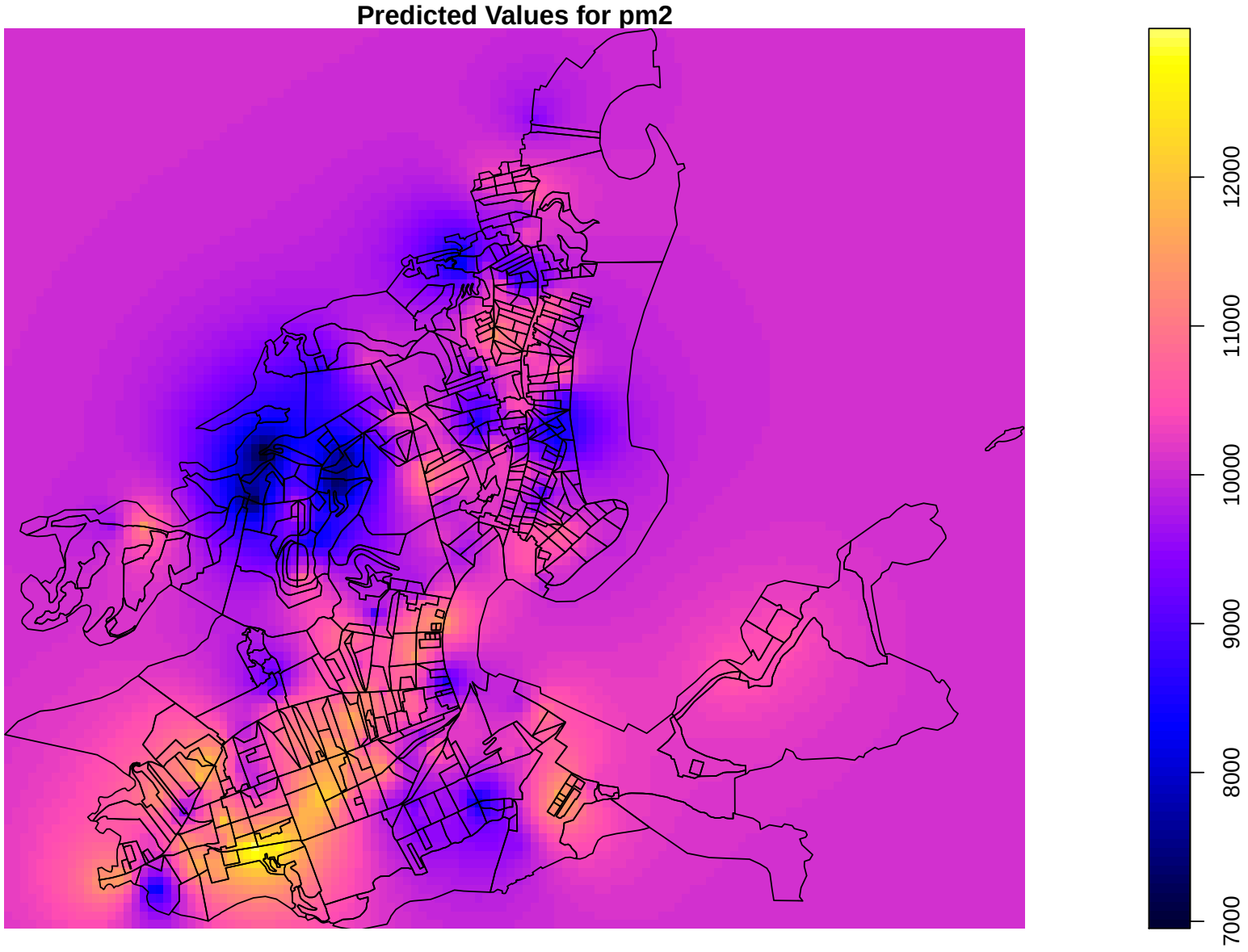
From the Gaussian Process, we can write:

$$f\left(\begin{pmatrix} \mathbf{Y}_s \\ \mathbf{Y}_{s'} \end{pmatrix} | \beta, \boldsymbol{\theta}\right) = N\left(\begin{pmatrix} \boldsymbol{\mu}_s(\beta) \\ \boldsymbol{\mu}_{s'}(\beta) \end{pmatrix}, \sigma^2 \begin{pmatrix} H_s(\phi) & H_{s,s'}(\phi) \\ H_{s,s'}^T(\phi) & H_{s'}(\phi) \end{pmatrix}\right)$$

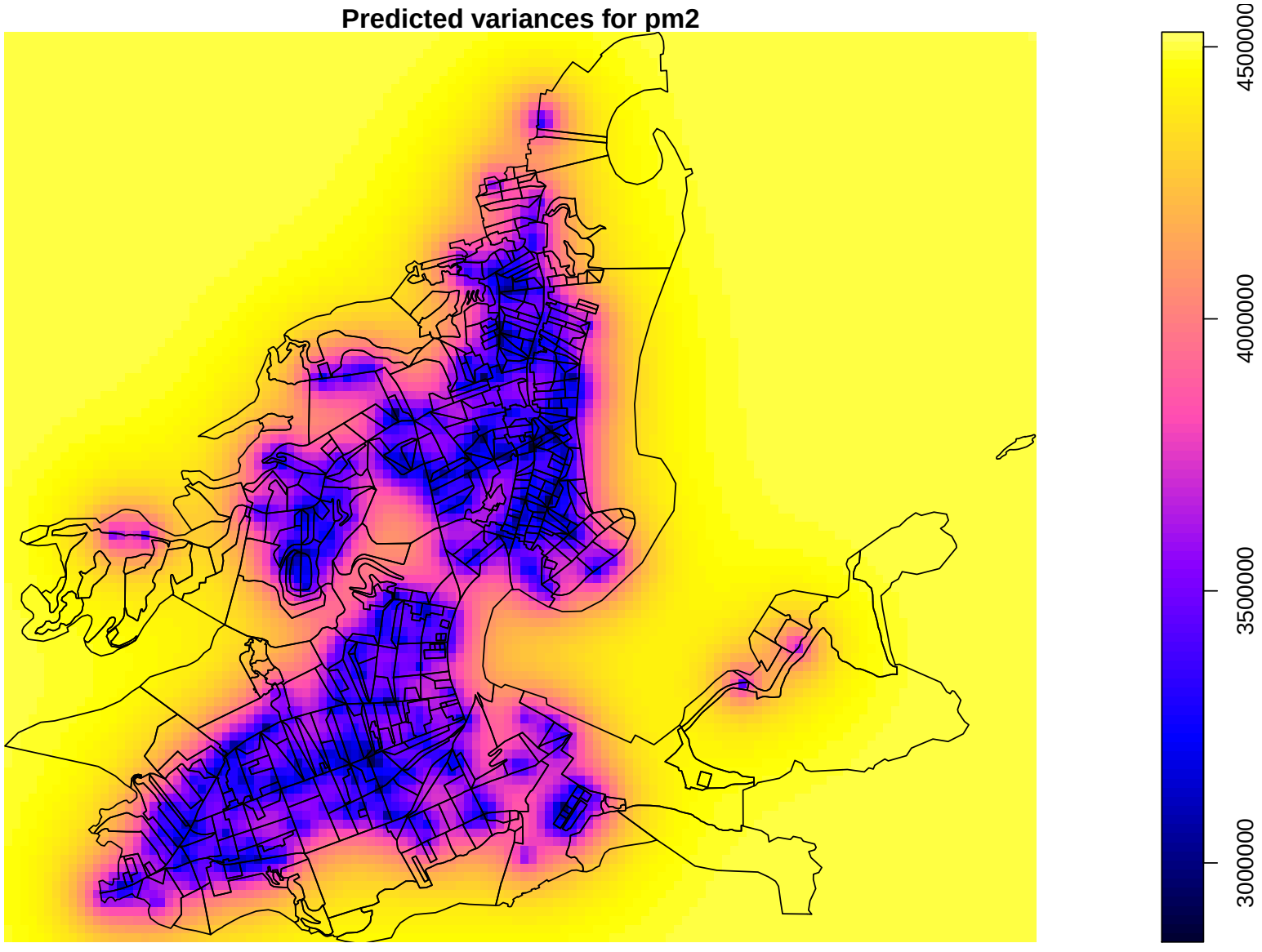
Therefore, $\mathbf{Y}_{s'} | \mathbf{Y}_s, \beta, \boldsymbol{\theta}$ is distributed as

$$N\Big(\boldsymbol{\mu}_{s'}(\beta) + H_{s,s'}^T(\phi) H_s^{-1}(\phi) \left(\mathbf{Y}_s - \mu_s(\beta)\right)$$
$$\sigma^2 \left[H_{s'}(\phi) - H_{s,s'}^T(\phi) H_s^{-1}(\phi) H_{s,s'}(\phi)\right]$$

Results



Results


Predicted variances for pm2
4500000
4000000
3500000
3000000

Results

The above methodology is applied to estimating house prices in the city of Rio de Janeiro, with data from appraisals and advertisements spanning the months between january 2014 and december 2017. The MAPE (Mean Absolute Percentage Error) evaluation metric is calculated using LOOCV (Leave one out cross validation) for both datasets. The results are compared to two other models applied individually to each dataset:

- Univariate spatial model, using the same specification and Bayesian strategy of the model presented in last section, but with latent spatial effects modeled using purely auto-regressive correlation matrices.
- Tree-based Gradient Boosting Machine model (see Picchetti (2017)), a popular machine learning algorithm, trained to predict house prices based on their intrinsic characteristics including location information.

Results (MAPE)

	Model	Appraisals	Advertisements
1	GBM	18.30	28.10
2	Spatial Univariate	19.20	31.60
3	Spatial Misaligned	16.80	25.70

Further Research

An alternative approach (see, for example Hill et alli (2017)) is to leverage the information of correlations between house prices not only across space, but across time as well.
The spatio-temporal model is

$$y_t(\boldsymbol{s}) = \boldsymbol{x}_t(\boldsymbol{s})^\top \beta_t + u_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}), \quad \epsilon_t(\boldsymbol{s}) \overset{ind}{\sim} N\left(0, \tau_t^2\right)$$

$$\beta_t = \beta_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \overset{i.i.d.}{\sim} N\left(0, \boldsymbol{\Sigma}_\eta\right)$$

$$u_t(\boldsymbol{s}) = u_{t-1}(\boldsymbol{s}) + w_t(\boldsymbol{s}), \quad w_t(\boldsymbol{s}) \overset{ind}{\sim} GP\left(\mathbf{0}, C_t\left(\cdot, \boldsymbol{\theta}_t\right)\right), \quad t = 1, 2, \ldots, N_t$$

Spatio-Temporal Model

Attractive features that make it worth exploring:

- The dynamic specification of the latent spatial effect $u_t(\boldsymbol{s})$ induces interactions between time and spatial effects on house prices not easily captured the direct specification of a proper parametric spatio-temporal correlation function.
- The induced smooth time trajectory of the intercept component in $\boldsymbol{x}_t(\boldsymbol{s})^\top \beta$ provides an attractive estimate of a time-dummy price index (as in Eurostat (2013)).
- Among the coefficients estimated in $C_t(\cdot, \boldsymbol{\theta}_t)$, the covariance matrix of the gaussian process behind the latent spatio-temporal effect, is the correlation between the values in $y_t(\boldsymbol{s})$. In our example, this is the correlation between asking prices and appraisal prices. Since it is also estimated through a dynamic specification, one can gain valuable insight on the trajectory of the differences between these prices, especially during changes in the housing market cycles.

Thank you!