POLITECNICO
MILANO 1863

Bayesian models for regression on real-estate data

Group Members:
Tommaso Bonetti
Simone Lima
Luca Panzeri
Davide Remondina
Alessandro Ruggieri
Andrea Zanin

Course: Bayesian Statistics
Academic Year: 2022/23

Contents

Introduction

In this project we have worked on real-estate data with the goal of developing a model which can predict and explain the price of houses.

Given the broadness of the scope of the project we have decided to start by developing several models coming from well-known areas of research: mixed effect models, auto-regressive processes [8], CAR models, Gaussian processes and latent Gaussian Markov random fields. We have then selected the best models considering their computational performance and the posterior inclusion probabilities of the regressors. Finally we have created an ensemble model averaging the selected models with Akaike weights [7].

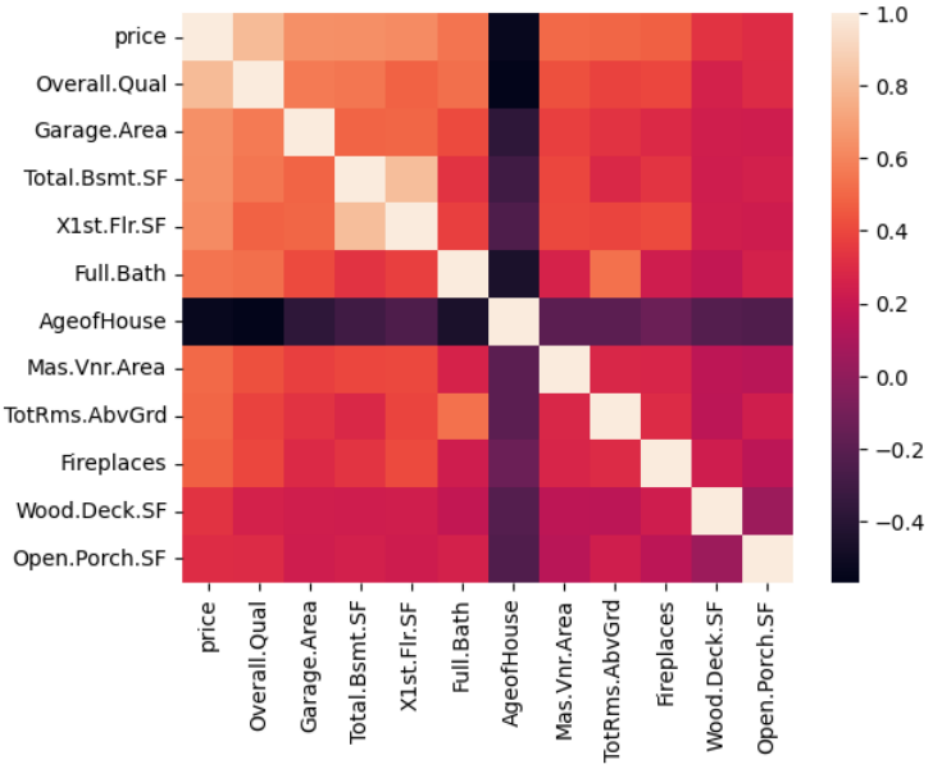
To fit and test the models we have used the Ames housing prices dataset [2], which contains data on the houses sold in Ames (Iowa) between 2006 and 2010.

Preliminary analysis and dataset preparation

Correlation Analysis

We have 2930 observations of 82 variables, so we started our work by finding the quantitative covariates most correlated to the price. We first cleaned our dataset of NaN values, and then we started our analysis, by computing correlation coefficient for every numerical variable with the price, our target variable. We decided to keep the variables which presented a correlation coefficient greater than 0.3, i.e. the variables most correlated to the price. At this point we noticed that some covariates were highly correlated to each other, for instance Garage.Area with Garage.Cars or Total.Bsmt.SF with BsmtFin.SF.1, presenting a correlation coefficient greater than 0.5, so we removed one for each pair of highly correlated variables, since they carried the same information.

Finally, we were left with 11 numerical covariates, correlated in this way to the price:



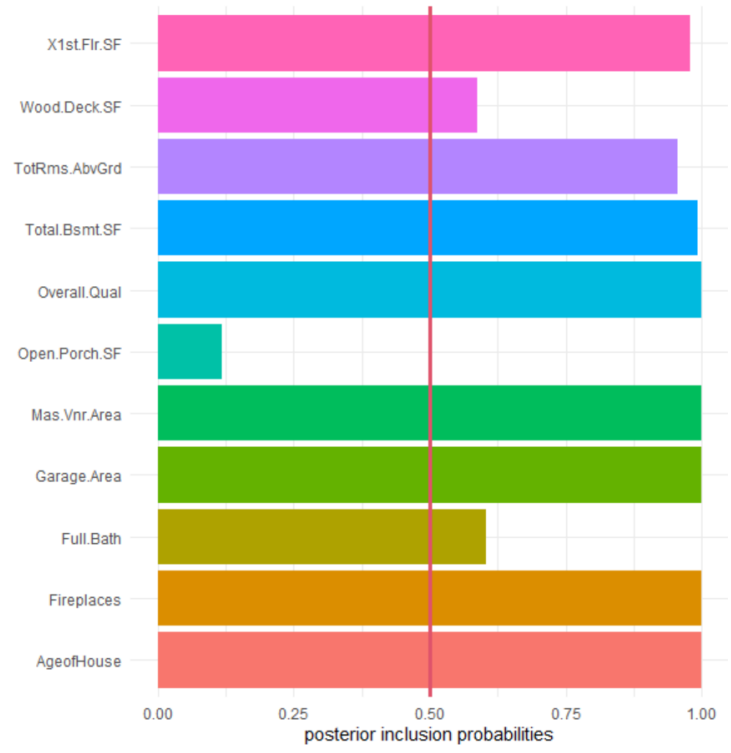
Stochastic Search Variable Selection

Before starting to formulate our models, we would like to know which are the most significant variables for price prediction among those we reduced to. So we used the Stochastic Search Variable Selection (SSVS) prior to do Bayesian Feature Selection.

The linear regression model we have considered is the following:

$$
\begin{aligned}
Y_j \,|\, \underline{\beta}, b, \sigma^2, \underline{x}_j &\overset{\text{ind}}{\sim} \mathcal{N}(b + \underline{\beta}^T \underline{x}_j, \sigma^2) \quad \forall j = 1, ..., N \\
\beta_i \,|\, \sigma_i^2 &\overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_i^2) \quad \forall i = 1, ..., p \\
\sigma_i^2 \,|\, \gamma_i &\overset{\text{ind}}{\sim} (1 - \gamma_i)\, \delta_{1860} + \gamma_i \delta_{18.6} \quad \forall i = 1, ..., p \\
\gamma_i &\overset{\text{iid}}{\sim} \text{Bernoulli}\,(0.5) \quad \forall i = 1, ..., p \\
\sigma &\sim \text{Unif}(0, 10) \\
b &\sim \mathcal{N}(0, 0.01)
\end{aligned}
$$

where $\sigma^2 = (\sigma)^2$, $N = 2305$ is the number of data points and $p = 11$ is the number of covariates in our dataset.
We fit this model using JAGS, running a chain of 110000 iterations, 10000 of which of burn-in, with thinning 10. Taking a look at the resulting traceplots, it seems to be a good mixing and convergence of the chains. With the Median Probability Model (MPM) strategy, that is picking all the variables with estimated marginal posterior inclusion probabilities larger than 0.5, we got the following results:



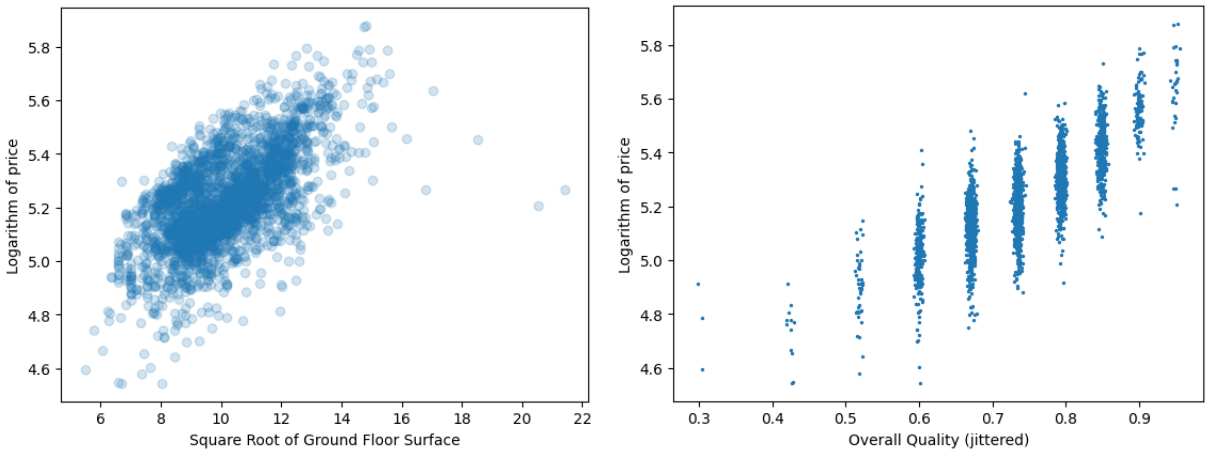
So we kept all the covariates, with the only exception of Open.Porch.SF. We obtained the same result also with the Highest Posterior Probability (HPD) strategy, and with the Hard Shrinkage (HS) strategy.
Actually, we carried on this analysis considering also the categorical variables that seemed most significant, i.e. Land.Contour, Central.Air and Condition.1, but running our future models, the effect of these categorical variables seemed really irrelevant, so we decided to go back and restart our analysis considering only numerical variables.

Preliminary Transformations

As final step in our preliminary analysis, we perform some transformations on the final covariates we obtained, through one by one variable inspection, to better adapt those to the problem we are studying.

We chose to perform the following preliminary transformations:

- Since the selling price ranged from 30.000 to 500.000 we decide to take the logarithm of the price, also to get a better distributed histogram
- We extract the square root of all covariates measuring surfaces, since the scatter plot showed us a parabolic behaviour instead of a linear one. This transformation is reasonable because the square root of a surface can have an understandable interpreation
- We observed that despite being a categorical variable, the Overall Quality follow a linear trend with respect to the price logarithm, so we decide to consider it as a numerical variable



Mixed-effect model

Model Formulation

The objective is to build a simple model for inference on the price of the house given the covariates chosen above plus the neighborhoods each house belongs to. Since there are 25 different neighborhoods, if we used categorical variables we would add too many. As a result, as wanted to focus on neighborhoods on our analisys, we opted for a linear mixed-effect model [9].

This is the model we used:

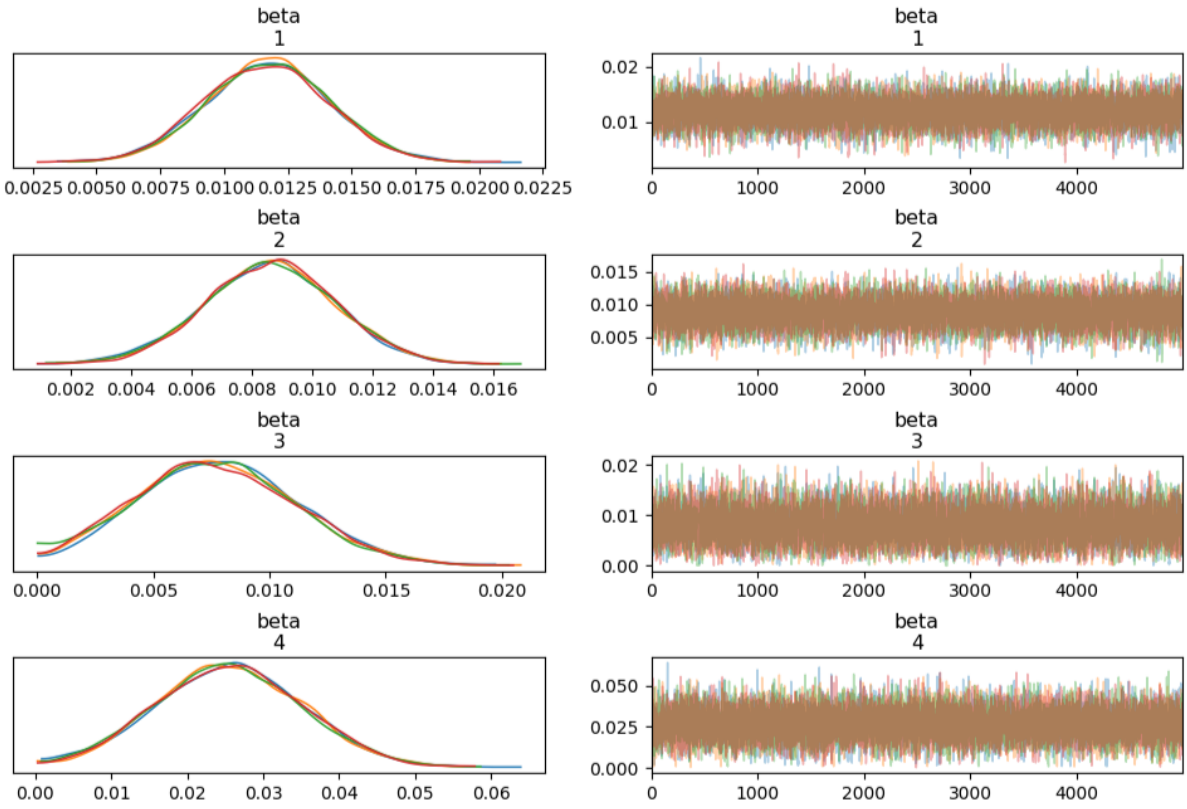
$$
\begin{aligned}
Y_{ij} \,|\, \underline{\beta}, \underline{\gamma}, \sigma &\overset{\text{ind}}{\sim} \mathcal{N}(\underline{X}_i\underline{\beta} + \underline{\gamma}_j, \sigma^2) \\
\beta_p \,|\, \tau &\overset{\text{iid}}{\sim} \text{half-}\mathcal{N}(0, \tau^2) \quad \forall p \in \{0, ..., P\} \\
\gamma_j \,|\, \xi &\overset{\text{iid}}{\sim} \mathcal{N}(0, \xi^2) \quad \forall j \in \{1, ..., M\} \\
\sigma^2 &\sim \text{InvGamma}(1, 30) \\
\tau^2 &\sim \text{InvGamma}(1, 0.05) \\
\xi^2 &\sim \text{InvGamma}(1, 0.05)
\end{aligned}
$$

where γ represent the effect of being in a specific neighborhood. A comment on the choice of the half-normal for the coefficient β has to be made. Since all the covariates we included in our model are positive features a house could have, there is no reason to believe that some β could be negative. In fact, why should a house cost less if it is bigger, has more rooms, or features a fireplace? Vice-versa, allowing negative coefficients could have resulted in noise capturing for non-significant covariates, thus we prefer to prevent it. The exception was the time passed from the last renovation, which still has a half-normal distribution but on the negative side, as the more time has passed since the last renovation, the less the house is expected to cost.

Model fitting and assessment

We fit this model using STAN, running 4 parallel chains of 6000 iteration, 1000 of which of burn-in. Treedepth, E-BFMI and Effective sample were satisfactory and no divergent transitions were found, $\hat{R} \in (1 \pm 10^{-3})$ for all parameters of the model.

	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	N_Eff/s	R_hat
lp__	2831.160000	0.057533	4.654880	2822.810000	2831.520000	2838.130000	6546.060000	6.421620	1.000470
β[1] - Overall quality	0.050307	0.000043	0.004518	0.042900	0.050271	0.057813	11244.100000	11.030300	0.999975
β[2] - Garage area	0.011731	0.000021	0.002456	0.007715	0.011747	0.015758	13511.600000	13.254700	1.000310
β[3] - Basement sq. ft.	0.008606	0.000021	0.002166	0.005075	0.008615	0.012130	10312.300000	10.116300	1.000260
β[4] - First floor sq. ft.	0.007760	0.000035	0.003463	0.002165	0.007642	0.013587	9803.130000	9.616760	1.000050
β[5] - Full bathrooms	0.025647	0.000097	0.009422	0.010076	0.025594	0.041209	9434.580000	9.255220	1.000170
β[6] - Years since last renovation	0.000238	0.000001	0.000230	0.000012	0.000167	0.000691	24333.500000	23.870900	0.999907
β[7] - Masonry veneer area	0.002025	0.000011	0.001367	0.000212	0.001802	0.004566	15601.600000	15.305000	1.000100
β[8] - Rooms above ground	0.013299	0.000030	0.003059	0.008249	0.013279	0.018328	10160.200000	9.967040	1.000140
β[9] - Fireplaces	0.025228	0.000065	0.006917	0.013760	0.025182	0.036652	11382.900000	11.166500	1.000350
β[10] - Wood deck sq. ft.	0.004011	0.000018	0.001793	0.001087	0.003969	0.007044	10193.500000	9.999760	1.000150
β[11] - Intercept	4.523840	0.000511	0.036928	4.462550	4.523800	4.584510	5221.140000	5.121890	1.000090
sigma	0.031395	0.000007	0.000945	0.029883	0.031368	0.032965	17770.189082	17.432367	0.999931
tau	2.243584	0.010459	1.171360	1.042270	1.958720	4.386380	12541.821381	12.303394	1.000058
xi	0.007191	0.000022	0.002367	0.004250	0.006768	0.011626	12050.060056	11.820981	0.999911
log_likelihood	1674.225393	0.229313	29.469497	1625.580000	1674.620000	1721.860000	16515.365306	16.201398	0.999943
log_likelihood_sq	1473.470030	0.469814	61.162233	1375.020000	1473.220000	1574.440000	16947.852538	16.625664	0.999963

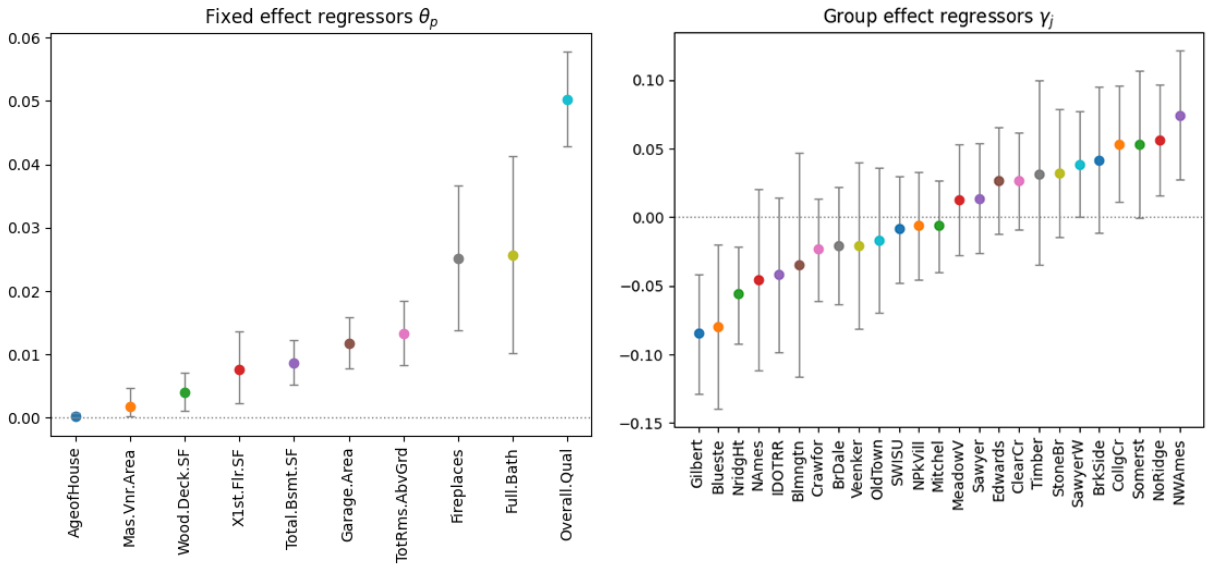


We computed WAIC [12] and we obtained a value of -3395.

$$\text{WAIC} = \; -2\sum_{i=1}^{n} \log f(y_i|\underline{y}) + 2\sum_{i=1}^{n} Var_{\theta|\underline{y}}(\log f(y_i|\theta))$$

Fixed effect parameter

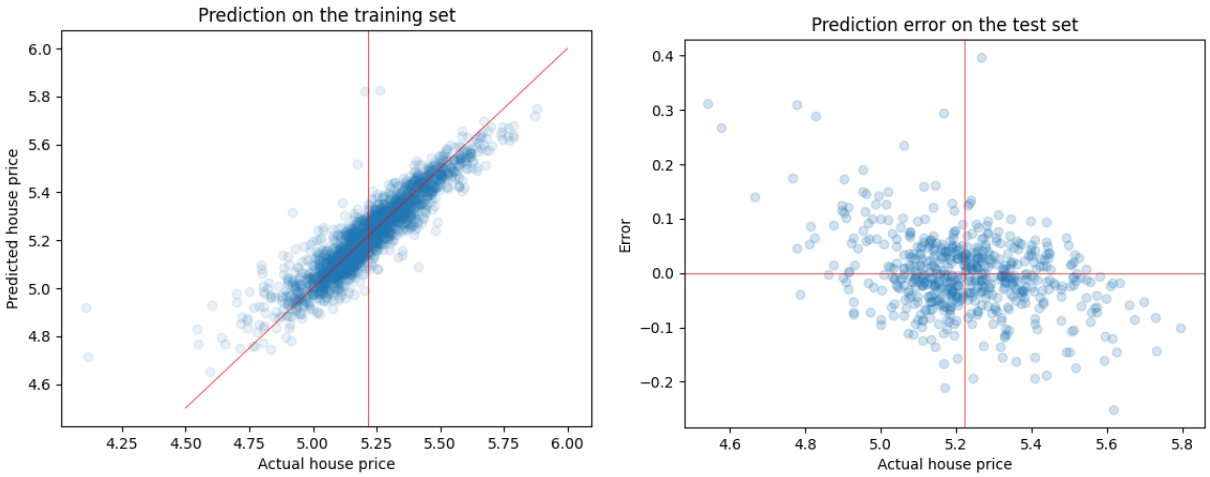
It is not easy to interpret the result due to the many transofrmations performed on the target and the covariates. In addition to this, the different scale of the variables can lead to wrong conclusions. In particular, it is hard to understand the meaning of β_4 which is the parameter of the square root of total surface of the ground floor.
It is easier to give an interpretation for β_6 or β_8: adding a room increases the price of a house by 3% on average, while every year passed from the last renovation decrases it by 1.4%.
We noticed that the overall quality of the house has a relevant impact on the final price of the house, but it is not very interesting since it is a qualitative feature and we are not sure of how this evaluation is really computed or assigned. The model seems to suggest that an additional bathroom is worth more than an additional generic room, but the variance of β_5 is too high to claim it for certain; we could perform further tests, but it is out of the scope of our analysis. Furthermore, *Garage Area* has a bigger coefficient than *First Floor Surface*: it is interesting that having a big garage has such a significant effect on the price of the house.



Neighborhood effect parameter

The majority of the neighborhoods present no relevant difference on the price of the houses. There are only six of them in which 0 does not belong to the 95% High posterior density interval. It is interesting to notice that the "Mitchel" neighborhood, which from the boxplot proves to be one of the cheapest neighborhoods, has an associated γ really close to zero. This means that the lower prices are due to the quality of the houses, and not to external factors related to the neighborhood. On the other hand, "NoRidge" has a quite high average price, but houses there seem to be average in quality. The difference in price in this case is due to the sole fact of being in that specific neighborhood.

Prediction



We perform prediction with our fitted model, using the test set we previously created. For pointwise prediction we used the mean of the predictive posterior distribution, namely

$$\mathbb{E}\left[f(y_k|\underline{y_{train}})\right] \ , \text{ where } \ f(y_k|\underline{y_{train}}) = \int_\Theta f(y_k|\theta)\pi(\theta|\underline{y_{train}})d\theta$$

We can notice that the model struggles to predict prices that are far from the mean, in particular from the *error scatterplot* we can notice that low prices are overestimated while very high price are understimated, so this model is not capturing some features that influence the price for particularly expensive or cheap houses.

CAR Model

We want to formulate a Conditional AutoRegressive (CAR) model, which deals with areal data, to take into consideration the spatial autocorrelation of our data.

CAR model with Leroux prior

As first formulation we proposed a CAR model with Leroux prior, where each house is treated as a different areal unit. So we have N different spatial random effects, one for each house in our dataset, and we have built the adjacency matrix in this way:
$A = [a_{kl}]_{k,l \in \{1,...,N\}}$ where
- $a_{kl} = 1$ if k is an house of a certain neighborhood and l is an other house of that same neighborhood or l is any house of a neighborhood adjacent with the neighborhood of k
- $a_{kl} = 0$ otherwise.

Model formulation

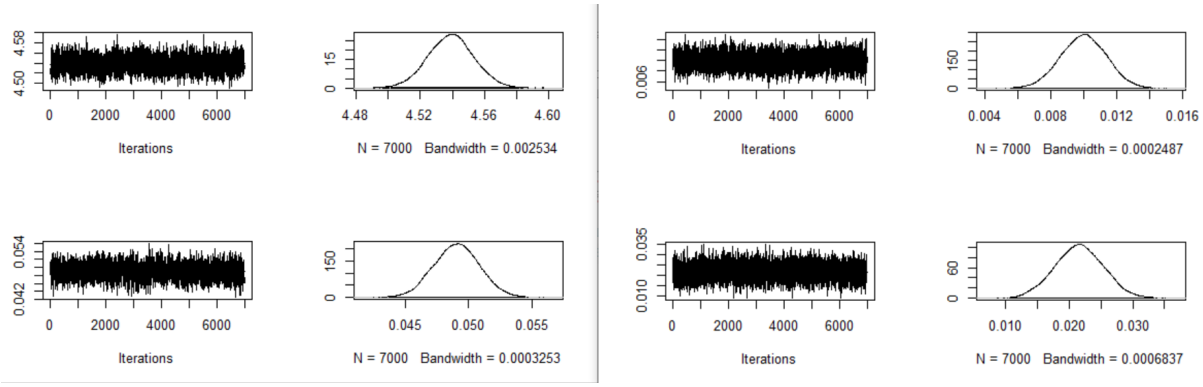
$$
\begin{aligned}
Y_j | \underline{\beta}, b, \underline{x}_j, \underline{\phi}, \nu^2 &\overset{\text{ind}}{\sim} N(b + \underline{\beta}^T \underline{x}_j + \phi_j, \nu^2) \quad \forall j = 1, ..., N \\
\beta_i &\overset{\text{iid}}{\sim} N(0, 10) \quad \forall i = 1, ..., p \\
b &\sim N(4.5, 0.01) \\
\nu^2 &\sim \text{Inv-Gamma}(0.1, 0.1) \\
\phi_k \mid \phi_{-k} &\sim \text{N}\left(\frac{\rho \sum_{l=1}^{N} a_{kl} \phi_l}{\rho \sum_{l=1}^{N} a_{kl} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{l=1}^{N} a_{kl} + 1 - \rho} \right) \quad \forall k = 1, ..., N \\
\tau^2 &\sim \text{Inv-Gamma}(0.1, 0.1) \\
\rho &\sim \text{Uniform}(0, 1)
\end{aligned}
$$

where $N = 2305$ is the number of data points and $p = 10$ is the number of covariates in our dataset.

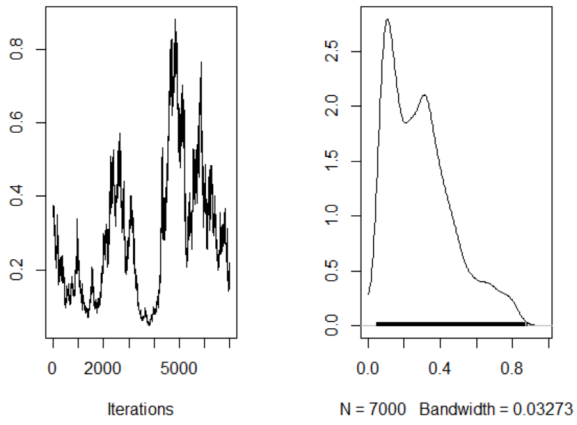
Model fitting and assessment

We wrote this model in Stan and we tried to make it run, but it was too slowly due to the huge adjacency matrix (NxN) we were working with at each iteration.
Thus we fit the model using the R package 'CARBayes', which implements spatial generalised linear mixed models for areal data, with inference in a Bayesian setting using MCMC simulations, running a chain of 15000 iterations, 1000 of which are of burn-in, with thin 2.

We obtained good mixing and covergence of the chain for betas, as we can see, for instance, in the following plots:



However, the traceplot of the spatial autocorrelation parameter rho, is not good:



the chain explore the support of rho too slowly and without a good mixing. Furthermore, the posterior mean of rho is 0.292, underlying as random effects have not modeled enough spatial autocorrelation of our data.

CAR model with multilevel Leroux prior

Since the results of the CAR model with Leroux prior was not satisfactory, we decided to try to formulate a more appropriate model, still in the field of areal data, and this was the CAR model with multilevel-Leroux prior.
In this model, each areal unit is represented by a different neighborhood, taking into consideration all the houses for every neighborhood. So, unlike the previous model where we had $N = 2305$ spatial random effect, one for each house in the dataset, this time we have 'just' $K = 25$ spatial random effects, one for each neighborhood.
We have built the new adjacency matrix in this way:
$A = [a_{kl}]_{k,l \in \{1,...,K\}}$ where
- $a_{kl} = 1$ if k and l are adjacent neighbors
- $a_{kl} = 0$ otherwise.

Model formulation

$$
\begin{aligned}
Y_j | \underline{\beta}, b, \text{Neigh}_j, \underline{x}_j, \underline{\phi}, \nu^2 &\overset{\text{ind}}{\sim} N(b + \underline{\beta}^T \underline{x}_j + \phi_{\text{Neigh}_j}, \nu^2) \quad \forall j = 1, ..., N \\
\beta_i &\overset{\text{iid}}{\sim} N(0, 10) \quad \forall i = 1, ..., p \\
b &\sim N(4.5, 0.01) \\
\nu^2 &\sim \text{Inv-Gamma}(0.1, 0.1) \\
\phi_k \mid \phi_{-k} &\sim \text{N}\left(\frac{\rho \sum_{l=1}^{K} a_{kl} \phi_l}{\rho \sum_{l=1}^{K} a_{kl} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{l=1}^{K} a_{kl} + 1 - \rho} \right) \quad \forall k = 1, ..., K \\
\tau^2 &\sim \text{Inv-Gamma}(0.1, 0.1) \\
\rho &\sim \text{Uniform}(0, 1)
\end{aligned}
$$

where $N = 2305$ is the number of data points, $p = 10$ is the number of covariates and $K = 25$ is the number of neighborhoods in our dataset.

Model fitting and assessment

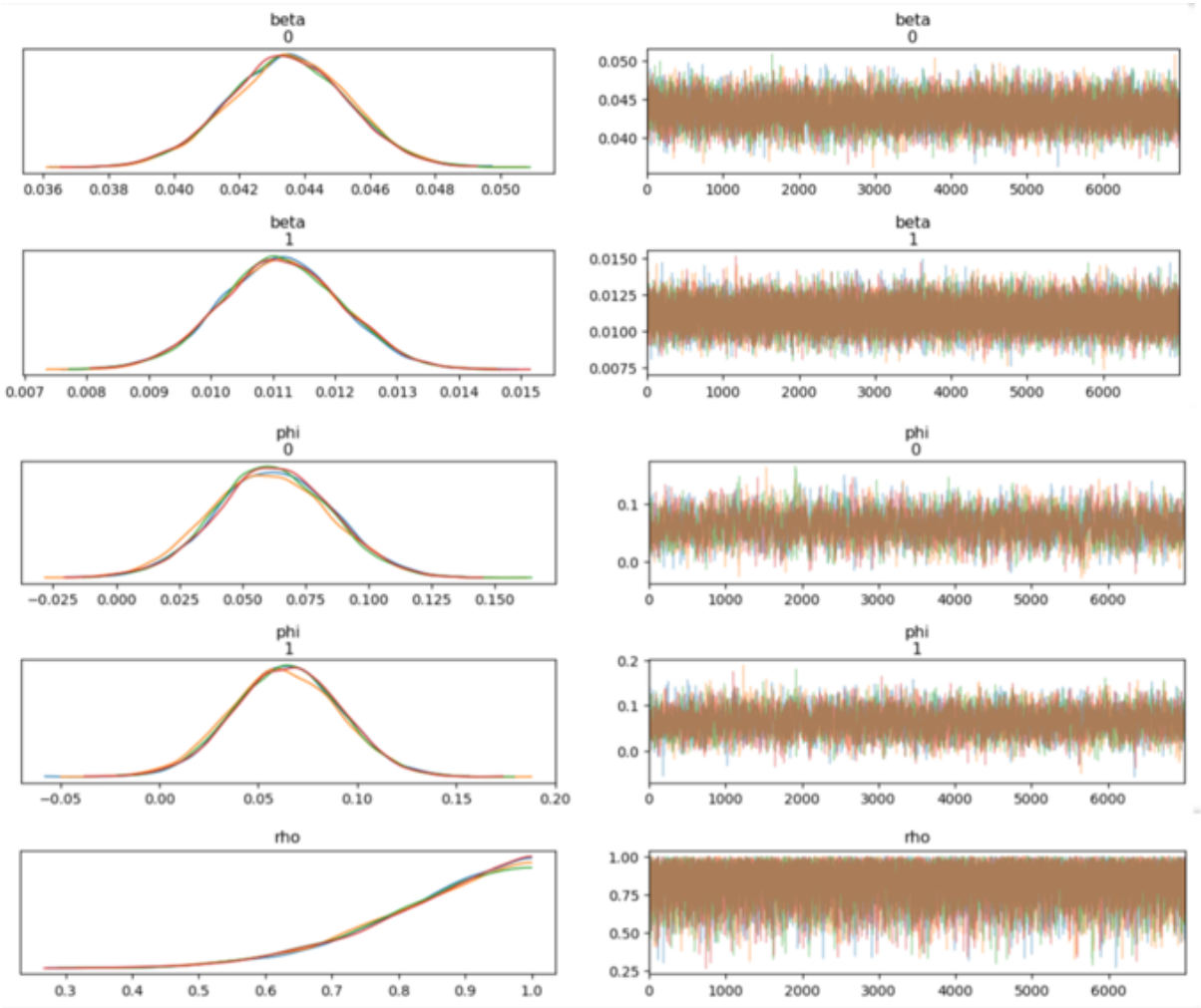
We fit this model using STAN, running 4 parallel chains of 7000 iterations, 1000 of which of burn-in. We noticed that 0.95-CI for the beta associated to Mas.Vnr.Area contains zero, thus we re-fit the model in the same way, i.e. running 4 parallel chains of 7000 iteration, 1000 of which of burn-in, but with one less covariate, Mas.Vnr.Area.

We can see the summary of the model fit in the next table:

	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	N_Eff/s	R_hat
lp__	5069.840000	0.047642	4.424360	5062.060000	5070.190000	5076.500000	8624.140000	2.445440	1.000560
beta[1]	0.043493	0.000021	0.001835	0.040463	0.043492	0.046503	7586.860000	2.151320	1.000720
beta[2]	0.011135	0.000005	0.000974	0.009540	0.011127	0.012742	34506.100000	9.784480	0.999956
beta[3]	0.009069	0.000004	0.000885	0.007614	0.009068	0.010508	39626.700000	11.236500	0.999966
beta[4]	0.008134	0.000016	0.001407	0.005827	0.008141	0.010441	7797.810000	2.211130	1.000560
beta[5]	0.015800	0.000027	0.003899	0.009391	0.015798	0.022246	20582.300000	5.836270	1.000060
beta[6]	-0.005803	0.000005	0.000489	-0.006603	-0.005802	-0.004994	10138.800000	2.874930	1.000450
beta[7]	0.014645	0.000007	0.001187	0.012690	0.014630	0.016602	29783.200000	8.445250	1.000090
beta[8]	0.027975	0.000023	0.002732	0.023479	0.027965	0.032453	14327.200000	4.062600	1.000280
beta[9]	0.003551	0.000004	0.000743	0.002323	0.003550	0.004774	36746.800000	10.419800	0.999914
beta_intercept	4.500226	0.000178	0.010127	4.483420	4.500290	4.516760	3226.419284	0.914877	1.001358
phi[1]	0.062002	0.000522	0.023365	0.023987	0.061780	0.100522	2000.960000	0.567388	1.002260
phi[2]	0.064281	0.000491	0.027127	0.019998	0.064216	0.108922	3057.090000	0.866863	1.001640
phi[3]	0.031716	0.000469	0.021562	-0.003387	0.031725	0.067412	2109.430000	0.598146	1.002370
phi[4]	0.087314	0.000474	0.018511	0.057064	0.087120	0.118193	1522.380000	0.431683	1.003440
phi[5]	0.150111	0.000515	0.021457	0.115128	0.149949	0.185736	1737.160000	0.492586	1.002730
phi[6]	0.117140	0.000506	0.018862	0.086412	0.117049	0.148772	1391.590000	0.394597	1.003450
phi[7]	0.159643	0.000503	0.019483	0.127716	0.159478	0.192063	1502.690000	0.426100	1.003260
phi[8]	0.079234	0.000484	0.018214	0.049395	0.079046	0.109694	1417.600000	0.401973	1.003500
phi[9]	0.109989	0.000486	0.018455	0.079893	0.109893	0.140679	1441.300000	0.408692	1.003050
phi[10]	0.014497	0.000471	0.018841	-0.015968	0.014309	0.045766	1599.390000	0.453520	1.003070
phi[11]	0.050402	0.000440	0.020259	0.017424	0.050243	0.084030	2121.070000	0.601447	1.002780
phi[12]	0.109459	0.000492	0.019084	0.078447	0.109352	0.141147	1501.570000	0.425783	1.003080
phi[13]	0.102812	0.000497	0.018230	0.073062	0.102593	0.133416	1344.870000	0.381347	1.003750
phi[14]	0.191115	0.000556	0.022175	0.155046	0.190854	0.228033	1593.490000	0.451845	1.003210
phi[15]	0.056686	0.000509	0.023623	0.017891	0.056573	0.095939	2151.790000	0.610156	1.002080
phi[16]	0.150791	0.000556	0.021025	0.116282	0.150652	0.185942	1430.360000	0.405589	1.003350
phi[17]	0.099629	0.000526	0.019809	0.067121	0.099469	0.132523	1419.870000	0.402615	1.003440
phi[18]	0.048340	0.000482	0.017988	0.019024	0.048282	0.078403	1393.300000	0.395080	1.003630
phi[19]	0.096590	0.000481	0.018371	0.066663	0.096447	0.127092	1460.570000	0.414157	1.003590
phi[20]	0.113864	0.000507	0.019353	0.082404	0.113723	0.145939	1457.970000	0.413419	1.003550
phi[21]	0.129563	0.000518	0.019690	0.097542	0.129358	0.162321	1445.300000	0.409825	1.003170
phi[22]	0.168512	0.000568	0.022830	0.131205	0.168307	0.206595	1618.250000	0.458867	1.002690
phi[23]	0.086235	0.000504	0.020947	0.051989	0.086200	0.121043	1728.790000	0.490212	1.002660
phi[24]	0.130283	0.000538	0.021409	0.095123	0.130029	0.165911	1583.530000	0.449023	1.003050
phi[25]	0.143282	0.000555	0.024853	0.102588	0.143270	0.184574	2003.290000	0.568048	1.002570
nu2	0.004850	0.000000	0.000140	0.004620	0.004850	0.005090	15594.928780	4.422060	1.000020
tau2	0.012880	0.000030	0.003990	0.007800	0.012170	0.020200	13086.025910	3.710650	1.000120
rho	0.848380	0.000910	0.118100	0.616430	0.874500	0.988940	16696.886730	4.734530	1.000370

Diagnostic metrics are good: the effective sample size is in the thousands, $\hat{R} < 1.01$ and there are no diverging iterations.

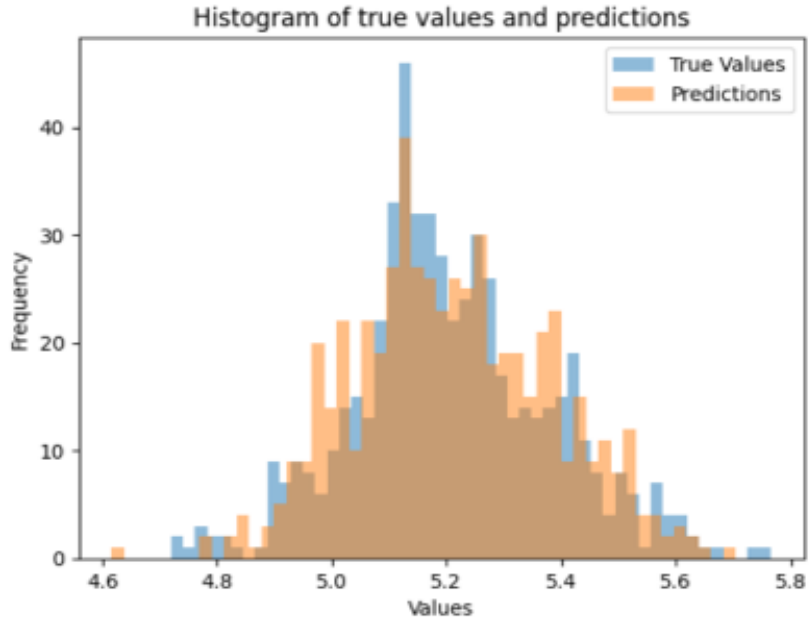
Here we can see some density and traceplot for beta, phi and rho:



- No 0.95-CI for beta contains zero, thus all the covariates are significant
- Good mixing and convergence of the chains
- Random effects have modeled a lot of spatial autocorrelation of our data, as the posterior mean for rho is 0.848

We proceeded by computing WAIC for our model, obtained a value of -5718, that we will use to compare the three models we have formulated.

Prediction



We computed the MSE that is equal to 0.0085.

Time-series model

Model formulation

We wanted to test whether time had an effect on the price of the houses; in particular we had information about the year and month of sale and we wanted to assess whether any of the two had an impact on the predicted price.

We thus formulated a Structural Time Series model [8] with 3 components: regression on the features of the house with an intercept, average price for the given month or year expressed as an auto-regressive process and a gaussian error; we chose to use uninformative priors. The following is the resulting model formulation considering the year of sale as the index of the auto-regressive process:

$$
\begin{aligned}
\text{Price}_i \mid \underline{\beta}, b, D_{\text{year}_i}, \sigma_\varepsilon^2, \underline{x}_i &\sim N\left(b + \underline{\beta}\underline{x}_i + D_{\text{year}_i}, \sigma_\varepsilon^2\right) \\
D_{t+1} \mid D_t, a &\sim N(aD_t, \sigma_\nu^2) \\
b &\sim N(0, 15) \\
a &\sim \text{Unif}(-1, 1) \\
\frac{1}{\sigma_\varepsilon^2} &\sim \text{Gamma}(3, 1) \\
\frac{1}{\sigma_\nu^2} &\sim \text{Gamma}(50, 1) \\
D_0 &= 0
\end{aligned}
\tag{1}
$$

We have fitted also an analogous model using the year and month as indexes (e.g. January 2006 corresponded to 1, while January 2007 corresponded to 13), the results and interpretation of the two models are equivalent. In the following assessment we will consider only the model based on the year.

Model fitting and assessment

We have fitted the model using STAN, in particular we have run 6 chains for 15000 iterations with 2000 iterations of warmup. Most diagnostic metrics of the Hamiltonian Monte Carlo algorithm were very good (see Table 1): effective sample size in the thousands, $\hat{R} < 1.01$ and satisfactory E-BFMI; inspection of the traceplots confirmed a very good mixing of the chains (see Figure 1).
The exception to this results was the parameter a which had $\hat{R} = 1.15$, effective sam-

ple size 50 and bad mixing; the sampler also reached maximum treedepth in 99% of the iterations.

The posterior distributions of the terms D_t are tightly centered around 0, so our conclusion has been that the bad behaviour of the parameter a is due to the model being inconsistent with the data: the effect of the year is none or negligible, thus the parameter a is estimating the ratio of terms that are very close to 0, it is then reasonable that the posterior distribution of a cannot be estimated accurately.
To confirm this assessment we have fitted a simpler linear model in which the year of sale was treated as a one-hot encoded categorical feature, in this model all the diagnostic metrics were satisfactory and the effect of the year was still negligible.

We concluded that there is no evidence that the year (or month) of sale has an effect on the price of the house when the other features are taken into account. This conclusion is supported by the previous considerations and by the fact that the posterior 90% highest probability density intervals of the parameters D_t all contain 0.

	Mean	MCSE	StdDev	5%	50%	95%	N_{eff}	$\hat{R}$
D_1	0.00839	1.32899e-04	0.00816	-0.00486	0.00831	0.02197	3775	1.00125
D_2	0.00523	1.44480e-04	0.00848	-0.00866	0.00519	0.01924	3450	1.00078
D_3	0.00686	1.40642e-04	0.00844	-0.02076	-0.00689	0.00711	3600	1.00176
D_4	0.00069	1.47918e-04	0.00992	-0.01564	0.00064	0.01713	4500	1.00116
a	0.13915	6.63900e-02	0.46502	-0.64751	0.16247	0.86370	49.05	1.15247

Table 1: Summary of the model fit in STAN

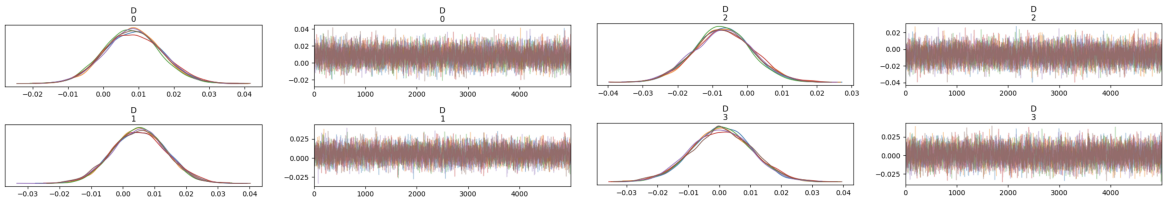


Figure 1: Traceplots for D_t

Geospatial model

We want to analyse the imapct of the georeferencial position on the price.

Teoretical Aprroach

We want to analyse the imapct of the georeferencial position on the price. Since the geostatistical models are too computationally heavy in order to be fitted using stan we used an alternative approach based on the Integrated Nested Laplace Approximation. This method consider the following model:

$$Y_i \mid \eta_i, \sigma^2 \sim \mathcal{N}\left(\eta_i, \sigma_e^2\right) \ \forall i \tag{2}$$

with $\eta_i = \alpha + \sum_{j=0}^{\eta_\beta} \beta_j Z_{i,j} + \sum_{k=0}^{\eta_f} f^{(k)}(u_{k,i})$ being the linear predictior for the mean, that comprehends a fixed effect part given by the covariates $Z_{i,j}$ and a random effect part given by the f terms.
The assumtion is that the vector $\xi = (\alpha, \beta_1, ..., \beta_{\eta_\beta}, f^{(1)}, ..., f^{(\eta_f)})$ is distributed as a Gaussian Markov Random Field with 0 mean and $Q(\theta)$ precision matrix.
Tjis asssumption allow to write the likelyhood pf the vector as:

$$\pi(\xi|\theta) \propto \pi(\theta)|Q(\theta)|^{1/2} \exp\{\frac{1}{2}\xi^T Q(\theta)\xi\} \tag{3}$$

and so the posterior distribution of hyperparameters and x is:

$$\pi(\xi, \theta|y) \propto \pi(\theta)\pi(\xi|\theta)\pi(y|\xi, \theta) \propto \pi(\theta)|Q(\theta)|^{1/2} \exp\{\frac{1}{2}\xi^T Q(\theta)\xi\}\pi(y|\xi, \theta) \tag{4}$$

alowing us to easily approximate the marginal posteriors (for more infomration Rue, Havard, Sara Martino, and Nicolas Chopin. (2009)).

Given the gaussian vector $\underline{\xi} = (\underline{\beta}, \underline{w})$ we can approximate the posterior distribution of $\theta = (\theta_1, \theta_2)$:

$$\pi(\theta|y) \propto \left.\frac{\pi(\theta,\xi,y)}{\pi(\xi|\theta,y)}\right|_{\xi=\xi^*(\theta)} \approx \left.\frac{\pi(\theta,\xi,y)}{\tilde{\pi}(\xi|\theta,y)}\right|_{\xi=\xi^*(\theta)}$$

where $\tilde{\pi}(\xi|\theta, y)$ is a gaussian approximation of $\pi(\xi|\theta, y)$.
Using this approxuimation we can compute the approximation of the posterior of the

parameters:

$$\pi(\theta_k|y) \approx \int \tilde{\pi}(\theta|y)d\theta_{-k}$$

$$\pi(\xi_j|y) \approx \int \tilde{\pi}(\xi_j|\theta, y)\tilde{\pi}(\theta|y)d\theta$$

Model formulation

$$\begin{aligned}
Y_j \mid \underline{\beta}, w, \sigma_e^2 &\sim \mathcal{N}\left(\underline{X}_j^T \underline{\beta} + w(s), \sigma_e^2\right) \quad \forall j \in 1, ..., N \\
\beta_i &\sim_{\text{iid}} \mathcal{N}(0, 1) \quad \forall i \in 1, ..., p \\
\sigma_e^2 &\sim \text{Log-Gamma}(0.1, 0.1) \\
w(s) \mid \theta &\sim \mathcal{N}\left(0, \Sigma(\theta)\right) \\
\text{Solution of } &\left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau w(s)) = \mathcal{W}(s) \\
\log(\kappa) &= log(\kappa_0) - \theta_2 \\
\log(\tau) &= log(\tau_0) - \theta_1 + \nu\theta_2 \\
\theta_1 &\sim \mathcal{N}(0, 1) \\
\theta_2 &\sim \mathcal{N}(0, 10) \\
\Sigma(s; \theta) &= \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa\|s\|)^\nu K_\nu(\kappa\|s\|)
\end{aligned}$$

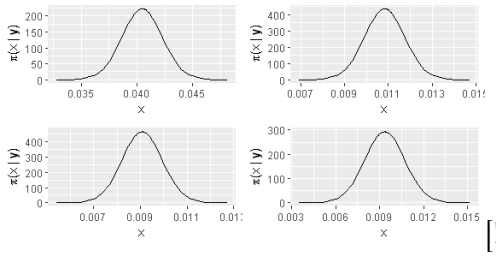
We used a popular and effective model that builds the random spatial effect $w(s)$ using an internal SPDE representation.
In practise $w(s)$ is found solving the following stochastic partial differential equation:

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau w(s)) = \mathcal{W}(s) \tag{5}$$

where $W(s)$ is a gaussian spatial white noise with 0 mean. The solution is found using the Finite Elements Method.
For what concenr the prior over the other hyperparameters we chosed some of them performing hyperparameter tuning and selecting the model with the lowest WAIC and MSE. The other hyperparameters were chosen looking at the literature. In particular, since κ_0 and τ_0 are related with σ^2 of the covariance function and the range ρ (that is the distance for which the correlation function has fallen to approximately 0.13) through the relations $\kappa_0 = \frac{\sqrt{8\nu}}{\rho}$ and $\tau_0 = \frac{1}{\sqrt{4\pi}\kappa_0\sigma}$, we focused on finding the right values.

Model fitting and assessment

We fitted the model using the INLA package available on R. From the image we can see that the model is able to predict with high accuracy the prices around the mean. It is still a bit inaccurate we trying to predict hte price of houses too cheap or too expensive. The WAIC of this model is -5526 and the MSE is 0.004.



[!b]

Figure 2: posterior distribution of the two parameters θ_1 and θ_2

	mean	sd	X0.025quant	X0.5quant	X0.975quant	mode
b0	4.6171404519	0.0206125235	4.5768360483	4.6170602636	4.657903859	4.6169051220
x1	0.0404761041	0.0017977756	0.0369509148	0.0404758178	0.044002897	0.0404752295
x2	0.0108337752	0.0009155418	0.0090383865	0.0108336692	0.012629763	0.0108334560
x3	0.0090784896	0.0008565258	0.0073987124	0.0090784347	0.010758577	0.0090783240
x4	0.0093495682	0.0013648067	0.0066726900	0.0093495814	0.012026373	0.0093496087
x5	0.0134489773	0.0036488972	0.0062944800	0.0134481899	0.020607919	0.0134466042
x7	0.0005876813	0.0007143219	-0.0008132881	0.0005876613	0.001988763	0.0005876211
x8	0.0143244761	0.0011641930	0.0120408881	0.0143245523	0.016607635	0.0143247061
x9	0.0242069644	0.0026067922	0.0190962857	0.0242062121	0.029321890	0.0242046976
x10	0.0038696164	0.0006791847	0.0025374899	0.0038696198	0.005201723	0.0038696265
x11	-0.0062336834	0.0004535148	-0.0071233714	-0.0062336156	-0.005344379	-0.0062334792

Figure 3: Summary of the model fit

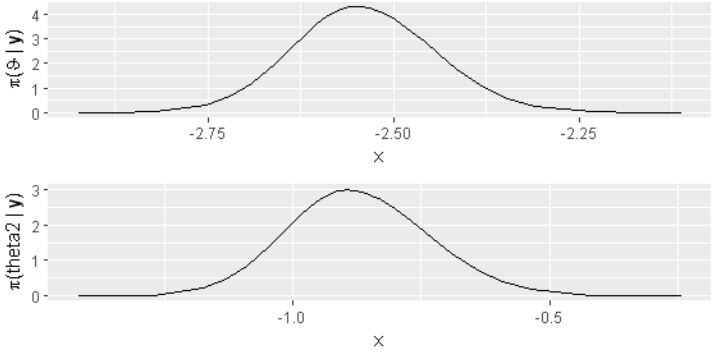


Figure 4: Posterior distribution of the betas

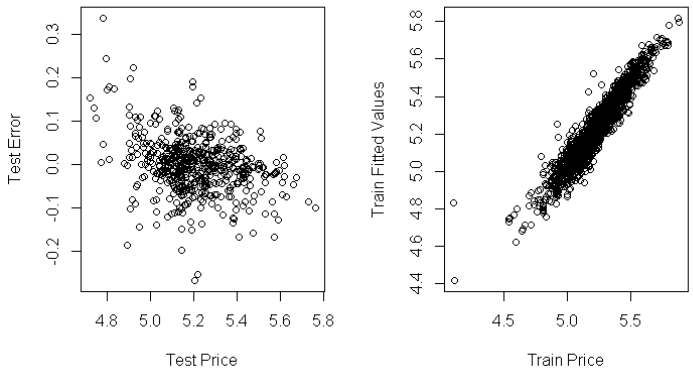


Figure 5: Results of the prediction

Model comparison

After having fitted the four models presented in the previous sections we discarded the Time Series model as there was no evidence that the date of sale had an impact on the price of the house. To compare and combine the three remaining models we followed the approach described in McElreath's book [7], which is based on the WAIC and thus allows us to combine models that have been fitted separately, this was necessary in our case because we used both STAN and INLA to fit the various models.
The following is the definition of the weights and the ensemble prediction as given by the book:

$$\mathrm{dWAIC}_i = \mathrm{WAIC}_i - \min_{j \in \{\mathrm{CAR,LME,GMRF}\}} \mathrm{WAIC}_j \qquad \forall i \in \{\mathrm{CAR}, \mathrm{LME}, \mathrm{GMRF}\} \quad (6)$$

$$w_i = \frac{\exp\left(-\frac{1}{2}\mathrm{dWAIC}_i\right)}{\sum_{j=1}^{m} \exp\left(-\frac{1}{2}\mathrm{dWAIC}_j\right)}$$

$$Y_{\text{pred, ENSEMBLE}} = w_{\mathrm{CAR}} Y_{\text{pred, CAR}} + w_{\mathrm{LME}} Y_{\text{pred, LME}} + w_{\mathrm{GMRF}} Y_{\text{pred, GMRF}}$$

The prediction that are being combined are point estimates, so in the case of the CAR model and the Linear Mixed Effects model they are the average of the posterior predictive distribution given the features of the house.

The following are the WAIC scores for the three models fitted on the training set

$$\mathrm{WAIC}_{\mathrm{CAR}} = -5717 \quad \mathrm{WAIC}_{\mathrm{LME}} = -3929 \quad \mathrm{WAIC}_{\mathrm{GMRF}} = -5526$$

Using the definition (6) the actual weights for our model are

$$w_{\mathrm{CAR}} \simeq 1 \quad w_{\mathrm{LME}} \simeq 0 \quad w_{\mathrm{GMRF}} \simeq 0$$

The interpretation is that we have enough data to conclude that the CAR model is almost surely better than either of the other two models and thus the optimal ensemble model is simply the CAR model.

Running the CAR model on the test set we get the predicted $\log_{10}$ prices, from which we compute the predicted prices; the standard deviation of the prediction is 27.000 $.

Bibliography

[1]

[2] D. De Cock. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[3] A. Guglielmi and M. Beraha. Lecture notes in Bayesian Statistics, 2022.

[4] D. Lee. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013. URL `https://www.jstatsoft.org/htaccess.php?volume=55&type=i&issue=13`.

[5] B. G. Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. E. Halloran and D. Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191, New York, NY, 2000. Springer New York. ISBN 978-1-4612-1284-3.

[6] S. Martino and A. Riebler. Integrated Nested Laplace Approximations (INLA), 2019. URL `https://arxiv.org/abs/1907.01248`.

[7] R. McElreath. *Statistical rethinking, a Bayesian course with examples in R and Stan*, pages 195–205. 2015.

[8] T. Proietti, M. Clements, and D. Hendry. *Forecasting with Structural Time Series Models*, pages 105–133. Blackwell Publishers, 1991.

[9] G. Rosner, P. Laud, and W. Johnson. *Bayesian Thinking in Biostatistics*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2021.

[10] H. Rue, F. Lindgren, Janet van Niekerk, and E. Krainski. R-INLA Project - Documentation. URL `https://www.r-inla.org/documentation`.

[11] C. N. Rue H, Martino S. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations. *Journal of the Royal Statistical*, 2009.

[12] S. Watanabe and M. Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.