# Bayesian hierarchical models for housing prices in the Helsinki-Espoo-Vantaa region

Ville Mäkinen

March 29, 2020

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

| Tiedekunta/Osasto — Fakultet/Sektion — Faculty | Laitos — Institution — Department |
|---|---|
| Matemaattis-luonnontieteellinen | Matematiikan ja tilastotieteen laitos |

Tekijä — Författare — Author
Ville Mäkinen

Työn nimi — Arbetets titel — Title

Bayesian hierarchical models for housing prices in the Helsinki-Espoo-Vantaa region

| Oppiaine — Läroämne — Subject | | |
|---|---|---|
| Matematiikka | | |
| Työn laji — Arbetets art — Level | Aika — Datum — Month and year | Sivumäärä — Sidoantal — Number of pages |
| Pro gradu -tutkielma | Maaliskuu 2020 | 87 s. |

Tiivistelmä — Referat — Abstract

Tässä tutkielmassa esitellään bayesläisten hierarkisten mallien käyttöä asuntojen hintojen mallintamiseen. Tutkielmassa käytetään aineistoa, jossa kuvataan tapahtuneita asuntokauppoja Helsingistä, Espoosta ja Vantaalta.

Tutkielmassa estimoidaan yhteensä viisi robustia regressiomallia. Malleissa käytetään Studentin $t$-jakaumaa likelihood-jakaumana, sillä aineistotarkastelut antavat viitteitä tietojen kirjausvirheistä. Neljässä mallissa on hierarkinen rakenne, joka perustuu myytyjen asuntojen kaupunginosiin. Malleista tuotetaan myös yhdistelmämalli käyttäen n.k. model stacking-menetelmää.

Mallien toimivuutta tarkastellaan posterior-jakaumasta johdettavien ennustejakaumien perusteella: Ennustejakaumista poimitaan otos, jonka perusteella muodostetaan jakaumat valituille tunnusluvuille. Tunnuslukujen jakaumia verrataan oikeasta aineistosta laskettuihin, toteutuneisiin tunnuslukuihin.

Mallien ennustekykyä vertaillaan tutkimalla ennustejakaumien kalibraatiota sekä terävyyttä. Lisäksi malleille lasketaan logaritmiset pisteet käyttäen leave-one-out ristiinvalidointia. Ristiinvalidoinnin laskennassa käytetään n.k. Pareto smoothed importance sampling-menetelmää. Ennustejakaumista tuotetaan myös piste-estimaatit käyttäen otoskeskiarvoja. Piste-estimaateille lasketaan $R^2$-suure.

Mallien tulokset ovat valtaosin uskottavia. Malleissa käytetyt selittävät muuttujat käyttäytyvät pääosin etukäteen odotetulla tavalla ja mallien ennusteet ovat järkeviä valtaosalle havainnoista. Tulokset viittaavat siihen, että hintamekanismi eroaa oleellisesti Helsingin keskustassa verrattuna muihin tutkittuihin alueisiin. Mallit kärsivät kuitenkin huonosta kalibroinnista sekä siitä, että kalliiden asuntojen hintaennusteet ovat valtaosin liian alhaisia.

Avainsanat — Nyckelord — Keywords
Asuntojen hinnat, bayesläinen mallintaminen, hierarkiset mallit, model stacking

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpulan tiedekirjasto

Muita tietoja — Övriga uppgifter — Additional information

**Abstract**

Objectives: The objective of this thesis is to illustrate the advantages of Bayesian hierarchical models in housing price modeling.

Methods: Five Bayesian regression models are estimated for the housing prices. The models use a robust Student's $t$-distribution likelihood and are estimated with Hamiltonian Monte Carlo. Four of the models are hierarchical such that the apartments' neighborhoods are used as a grouping. Model stacking is also used to produce an ensemble model. Model checks are conducted using the posterior predictive distributions. The predictive distributions are also evaluated in terms of calibration and sharpness and using the logarithmic score with leave-one-out cross validation. The logarithmic scores are calculated using Pareto smoothed importance sampling. The $R^2$-statistics from the point predictions averaged from the predictive distributions are also presented.

Results: The results from the models are broadly reasonable as, for the most part, the coefficients of the explanatory variables and the predictive distributions behave as expected. The results are also consistent with the existence of a submarket in central Helsinki where the price mechanism differs markedly from the rest of the Helsinki-Espoo-Vantaa region. However, model checks indicate that none of the models is well-calibrated. Additionally, the models tend to underpredict the prices of expensive apartments.

Keywords: housing prices, Bayesian modeling, hierarchical model, model stacking

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In 2016, Statistics Finland reported[1] that apartments formed approximately half of the wealth of households, so it seems likely that housing price predictions are of interest to households when buying or selling an apartment, for example. Similarly, price predictions are presumably useful for construction companies for revenue calculations and for banks for the purposes of assessing collateral value. It is therefore clear that housing price modeling has practical importance.

On the methodological side, housing price data provides a good way to showcase Bayesian methods, specifically Bayesian hierarchical models. The data provides many natural groupings such as the location of the apartment (the neighborhood, the city and the region) and the type of the apartment (flat, row house etc.), so hierarchical models provide a natural way to utilize this information. Additionally, if the exact locations of the sold apartments is available, Gaussian process models can be used to utilize the spatial information comprehensively.

This thesis is an attempt to illustrate the advantages of Bayesian methods for modeling housing prices because good predictions are of practical importance and because the available data provides good opportunities to do this. The specific data utilized in this thesis has been compiled from housing sales in Helsinki, Espoo and Vantaa. The housing prices are modeled through so-called hedonic pricing theory which assumes that the price of an apartment is determined by the apartment's properties. To the author's knowledge, this thesis is the first time that Bayesian methods have been applied to housing price modeling with Finnish data.

The thesis is structured as follows: Chapter 2 gives an introduction to hedonic pricing theory and its application to housing prices. Additionally, recent research on housing prices in the Finnish context is also noted. Chapter 3

[1] *https://tilastokeskus.fi/til/vtutk/2016/vtutk_2016_2018-06-05_tie_001_fi.html*

presents the methods used in this thesis. Chapter 4 describes both the data used in this thesis and the way it was preprocessed. Chapter 5 describes the general modeling process. Chapter 6 describes the structure and estimates for the individual models and chapter 7 the model comparisons and model stacking results. Chapter 8 discusses the results and concludes the work.

# Chapter 2

# Hedonic pricing theory and its application to housing prices

This chapter serves as an introduction to hedonic pricing theory and its application to housing price modeling.

## 2.1 Hedonic pricing theory

### 2.1.1 Theoretical foundations

Hedonic pricing theory is a theory of product differentiation where goods are represented as indivisible bundles of 'utility-bearing' characteristics. The foundations for the theory are presented concisely in Rosen (1974). The theory assumes that market participants, i.e. consumers and producers, are rational and seek to maximize their respective utility or profit functions. Together with additional necessary assumptions on the market structure, the formulation of goods as bundles and the rationality of market participants imply the existence of an equilibrium price function which is defined over the set of characteristics.

Rosen (1974) also sketches an outline for the empirical study of the theory. The outline includes the first step of regressing the observed prices over the characteristics. These estimates can then be used for the analysis of other parameters of interest such as the parameters that define the utility function of consumers. This thesis is essentially concerned how this first step regression could be done using Bayesian methods.

### 2.1.2 Application to housing prices

Housing serves as a natural example of a differentiated good where the utility-bearing characteristics of hedonic pricing theory correspond to the characteristics of the apartment. Sheppard (1999) provides a general overview on how hedonic pricing theory has been applied to housing prices. The article notes the special characteristics of the housing market, e.g. the effects of location, the search costs and the possibility of resale. The article also includes a general discussion on the appropriate econometric tools for estimating the hedonic prices. In this discussion it is noted that the land value plays a central part in urban economic theory and that the price of land would typically be expected to vary with location (Sheppard (1999, p. 1616)).

Early empirical articles on hedonic price estimation for housing include Dubin (1988), Dubin (1992) and Can (1992). These articles describe different approaches of how the spatial aspects of the housing data can be utilized in models.

In Dubin (1988), housing prices are modeled with a regression model with spatial autocorrelation in the error terms. The spatial autocorrelation is parameterized in the correlation matrix using a negative exponential function that utilizes distances between the observations. Dubin (1992) extends the work in Dubin (1988) by describing a method for kriging housing prices for locations that were not in the estimation data.

Can (1992) presents four models where the influence of location is modeled through so-called neighborhood effects and adjacency effects. The neighborhood effects encode the available neighborhood-specific information, e.g. census data on unemployment rates per neighborhood. The adjacency effects incorporate spatial spill-over effects, i.e. how the prices of adjacent housing might influence the price of a given apartment. The four models are formulated as combinations of alternative ways to handle the neighborhood and adjacency effects: For the neighborhood effects, the effects were modeled either directly as explanatory variables or indirectly as interaction terms. For the adjacency effects, the effects were modeled through the inclusion or exclusion of an autoregressive term defined with an a priori weight matrix based on the inter-observation distances.

Dubin (1998) compares methods for modeling spatial autocorrelation either through the use of a distance-based weight matrix, as in Can (1992), or through explicit parameterization, as in Dubin (1988) and Dubin (1992). The article suggests that it is probably better to model the spatial autocorrelation through parameterization rather than through the use of a weight matrix.

A slightly later discussion on the theoretical foundations and the appropriate econometric tools for hedonic pricing theory for housing can be found in Bowen

et al. (2001). In the discussion on the theoretical foundations, the article considers the concepts of *spatial heterogeneity*, which refers to the systematic variation of housing prices depending on location, and *spatial dependency*, which refers to the interdependence of the observations' prices that depends on the relative locations of the observations and that is not explicitly included in the model structure. The article notes that spatial patterns in housing data are likely due to both spatial heterogeneity and spatial dependence, so statistical procedures that can handle both properties should be used when constructing hedonic housing price models.

Goodman and Thibodeau (1998) describes a modeling procedure that allows for spatial heterogeneity by iteratively testing for the existence of submarkets, i.e. geographic areas with differing prices per housing characteristic. The procedure consists pairwise comparisons where a hierarchical linear model is fitted using a data set which comprises of sales located in adjacent geographic areas. If a designated group-dependent coefficient is not statistically significant, the areas are deemed to be in the same submarket and are thus combined for further comparison. Otherwise, the areas are deemed to be in separate submarkets. This process is continued until all areas are assigned to some submarkets. The article notes that the results from the procedure depend on the initial conditions, i.e. what were the areas that were compared first, and in what order the comparisons were made.

In more recent years, research on housing prices using newer estimators has started to appear. In Gelfand et al. (2003), Gaussian process models with spatially varying coefficients are introduced and the models are illustrated with an example using housing price data. In Hui et al. (2010), a Bayesian hierarchical model for housing prices is developed. The model of Hui et al. (2010) outperforms other available valuation methods whilst still being relative simple. In Yu et al. (2007), four regression models are constructed including a geographically weighted regression model (described in Fotheringham et al. (1998)) in which the regression coefficients are allowed to vary spatially. Examples of articles where machine learning algorithms are applied to housing prices include Selim (2009), where a neural net is developed, and Antipov and Pokryshevskaya (2012), where a random forest model is compared with multiple other predictive methods. Antipov and Pokryshevskaya (2012) also includes a discussion on how the data could be segmented for submodel construction.

### 2.1.3 Hedonic pricing theory for housing prices in the Finnish context

Hedonic pricing has been utilized for housing in the Finnish context in, at least, Kyllönen and Räty (2000), Kortelainen and Saarimaa (2015), Eerola and Lyytikäinen (2015), Eerola and Saarimaa (2018) and Harjunen (2018). In Kyllönen and Räty (2000), a semi-parametric model with splines is estimated for housing prices in Joensuu. Eerola and Lyytikäinen (2015) studies the impact on prices and sales times after new information channels are provided to market participants. Kortelainen and Saarimaa (2015) studies whether owner-occupied housing generates positive externalities which would capitalize into housing prices. Eerola and Saarimaa (2018) uses hedonic pricing as a tool to measure implicit rent subsidies for public housing. The doctoral thesis Harjunen (2018) consists of three articles which use hedonic pricing theory to study the price effects of school quality, heating technology choices and an announcement of a new metro line in the Helsinki metropolitan area.

The hedonic pricing theory for housing in a Finnish housing market context has also been a popular subject for postgraduate studies. Recent master's theses on the subject include, at least, Brotherus (2011), Koivuniemi (2014), Takala (2016), Luhta (2017), Jantunen (2017), Hakala (2018), Valaja (2018) and Vuorela (2019). The focus of the theses has ranged from simply estimating the price mechanism to analyzing price effects of a specific phenomenon, e.g. opening of a new tram line or urban infill. Each listed thesis has utilized OLS estimators.

## 2.2 Relation to previous work

The models estimated in this thesis are simple Bayesian hierarchical models with a focus on prediction. To this end, model stacking is also presented. In relation to the listed master's theses, this thesis contributes to this body of work by utilizing Bayesian methods instead of OLS estimators. In relation to the existing literature, this thesis can be seen as an application of well-established model structures to a novel data set. The thesis' models are closest structure-wise to the model in Hui et al. (2010) or, e.g., the models described in chapters 12 and 13 of McElreath (2016) and Gelman and Hill (2006). Additionally, an attempt is made to estimate a Gaussian process model by assigning locations to observations based on the observations' neighborhoods following an example in section 13.4 of McElreath (2016). Similar Gaussian process models are described in, e.g., Gelfand et al. (2003) or chapter 6 of Banerjee et al. (2015).

# Chapter 3

# Methods

This chapter provides an overview of the methods used in this thesis.

## 3.1 Foundations

The models in this thesis are estimated with Bayesian methods where the focus is on determining the posterior distribution $p(\theta|y)$ of parameters $\theta$ given the data $y$, likelihood distribution $p(y|\theta)$ and prior distribution $p(\theta)$. Bayes' theorem provides the connection between these terms such that

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where $p(y)$ is the marginal distribution for data $y$, i.e. $p(y) = \int p(y|\theta)p(\theta)d\theta$. The marginal distribution $p(y)$ can be intractable, so methods have been developed for determining the posterior without explicitly determining the marginal distribution $p(y)$.

In practical applications, the objective of statistical modeling is usually prediction. Bayesian methods allow formulating prediction through the posterior predictive distribution. For new data $\tilde{y}$, the posterior $p(\theta|y)$ determines the posterior predictive distribution $p(\tilde{y}|y)$ such that

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta. \tag{3.1}$$

The models in this thesis are regression models, so for data $y = (y_1, \ldots, y_n)$, each observation $y_i$ is divided into two parts such that $y_i = (v_i, x_i)$ where the term $v_i$ is the response variable and the term $x_i = (x_{1i}, \ldots, x_{ki})$ is a set of explanatory variables for observation $i$. The purpose of the models is to predict the value of the response variable $v_i$, in this case the price of an apartment,

given the values for the explanatory variables $x_i$ which in this case encode the measurable properties of the apartment. The relationship between the response variable $v_i$ and the explanatory variables $x_i$ is assumed to be linear in the expected value such that

$$\mathrm{E}(v_i|\beta, x_i) = \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

where the term $\beta = (\beta_1, \ldots, \beta_k)$ denotes the coefficients for the explanatory variables. The model parameter $\theta$ then includes the coefficients $\beta$ along with the other necessary terms related to variance and to possible group-level effects. This explicit notation is suppressed for the discussion on the methods for the rest of this chapter.

## 3.2 Markov chain Monte Carlo

Perhaps the most popular current general method for determining the posterior distribution is the Markov chain Monte Carlo (MCMC) method. In MCMC, a sample is produced by iteratively drawing values from a Markov chain that has been constructed such that its stationary distribution is the desired posterior distribution. A sufficiently large sample drawn from this Markov chain then provides a description of the posterior distribution. The material of this section has been adopted from chapter 11 of Gelman et al. (2014).

The MCMC samples have two properties that need to be considered for statistical inference, namely within-sequence correlation and the requirement for convergence. First, sampling from a Markov chain induces within-sequence correlation in the resulting sample, so inference using a MCMC sample is generally less precise than inference from an independent sample of the same length. Second, in order for the draws to describe the posterior, the Markov chain needs to have converged to its stationary distribution. The early draws from the Markov chain are influenced by the chain's starting point which is normally randomly initialized, so they are not typically representative of the stationary distribution. The early draws are therefore discarded from any analysis. To check convergence, it is typical to run the Markov chain multiple times and compare the resulting draws. When convergence has occurred, the draws from any of the chains are similar to draws from any other chain.

Convergence can be studied using so-called trace plots where the draws from separate Markov chains are plotted with respect to their iteration sequences. If the chains have converged, the draws from each chain are mixed, indicated in the trace plots by draws being located at the same specific region of values, and stationary, indicated in the trace plots by the lack of any trends with respect to

the iteration.

Convergence can also be analyzed with the $\hat{R}$-statistic and the effective sample size $\hat{n}_{\text{eff}}$-statistic which check whether the individual sequences from the Markov chains mix well with each other. The metrics do not guarantee that the Markov chains have really converged to the posterior since well-mixed chains are only a necessary condition for convergence.

The $\hat{R}$-statistic measures the potential scale reduction for the distribution of an estimand $\psi$ that could be achieved if the number of iterations of the Markov chain was increased. The $\hat{R}$-statistic is calculated using multiple Markov chain sequences that have been split into two at the middle index. Let $\psi_{ij}, i = 1, \ldots, n, j = 1, \ldots, m$ denote the $i$th draw in the $j$th half-sequence for estimand $\psi$. Using the half-sequences, the between-sequence variance $B$ and the within-sequence variance $W$ are defined as

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2, \text{ where } \bar{\psi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}, \ \bar{\psi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{\cdot j}$$

and

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{\cdot j})^2. \tag{3.2}$$

Then, the marginal posterior variance of estimand $\psi$ is estimated with

$$\hat{\text{var}}^{+}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B. \tag{3.3}$$

The estimator (3.3) for the marginal posterior variance and the within-sequence variance (3.2) can then be used to define the $\hat{R}$-statistic as

$$\hat{R} = \sqrt{\frac{\hat{\text{var}}^{+}(\psi|y)}{W}}.$$

Gelman et al. (2014, p. 287) gives a general maximum 'acceptable' threshold of at most 1.1 for the $\hat{R}$-statistic. McElreath (2016, p. 258) gives a stricter threshold of 1.01 and suggests that the estimates with $\hat{R}$-statistic greater than 1.00 should be regarded with suspicion.

The $\hat{n}_{\text{eff}}$-statistic for the effective sample size is based on considerations of the statistical efficiency of the average of draws $\bar{\psi}_{\cdot\cdot}$ as an estimate of the posterior mean $\text{E}(\psi|y)$. The construction of the statistic utilizes the asymptotic formula of the variance of the average of a correlated sequence,

$$\lim_{n\to\infty} mn\text{var}(\bar{\psi}_{..}) = \left(1 + 2\sum_{t=1}^{\infty}\rho_t\right)\text{var}(\psi|y),$$

where the term $\rho_t$ is the autocorrelation of the sequences for estimand $\psi$ at lag $t$. The effective sample size $n_{\text{eff}}$ can then be defined as

$$n_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{\infty}\rho_t}. \tag{3.4}$$

To use the definition in practical applications, the sum of the correlations in equation (3.4) has to be estimated. Therefore, define first the variogram $V_t$ for lag $t$ as

$$V_t = \frac{1}{m(n-t)}\sum_{j=1}^{m}\sum_{i=t+1}^{n}(\psi_{i,j} - \psi_{i-t,j})^2.$$

Then by inverting the formula $E(\psi_i - \psi_{i-t})^2 = 2(1-\rho_t)\text{var}(\psi)$, an estimator for correlation $\rho_t$ can be defined as

$$\hat{\rho}_t = 1 - \frac{V_t}{2\hat{\text{var}}^+} \tag{3.5}$$

where the term $\hat{\text{var}}^+$ is the estimate given by equation (3.3).

For large values of $t$, the estimates (3.5) for the correlation are too noisy. To account for the noise, the sum in equation (3.4) is replaced by a partial sum starting from lag 0 and continuing until the sum of autocorrelation estimates for two successive lags $\hat{\rho}_{2t'} + \hat{\rho}_{2t'+1}$ is negative. Thus the estimator for the effective number of samples becomes

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T}\hat{\rho}_t}, \tag{3.6}$$

where the estimated autocorrelations $\hat{\rho}_t$ are computed using formula (3.5) and $T$ is the first odd positive integer for which $\hat{\rho}_{T+1} + \hat{\rho}_{T+2}$ is negative.

In general, the effective number of samples $\hat{n}_{\text{eff}}$ should be at least 10 for all estimands of interest. Whether or not this is sufficient for any particular application depends on the application itself.

## 3.3 Hamiltonian Monte Carlo

There is a multitude of different methods for constructing Markov chains with the desired properties. The specific MCMC method used in this thesis is Hamiltonian Monte Carlo (HMC), first introduced in a statistics context in Neal (1993) and later reviewed in, e.g., Neal (2011). The material in this section has been

adopted from chapter 12 of Gelman et al. (2014).

HMC emulates a physical system driven by Hamiltonian dynamics to improve the efficiency of a Metropolis-Hastings styled sampling algorithm with better transition proposals. For each component $\theta_j$ of the model parameter $\theta = (\theta_1, \ldots, \theta_p)$, HMC adds a momentum variable $\phi_j$. The parameter pair $(\theta, \phi)$ are then updated together using a Metropolis algorithm with a specific way for generating transition proposals. The momentum variable $\phi$ is usually given a multivariate normal distribution with mean 0 and the covariance matrix $M$ set to a prespecified 'mass matrix', which can be chosen to be a diagonal matrix to simplify the algorithm implementation. Other parameters required by the algorithm are the number of leapfrog steps $L$ and a scaling factor parameter $\epsilon$. Depending on how the method has been implemented, these parameters can be tuned during the execution to improve the efficiency of the method. In addition to these parameters, HMC requires the calculation of the gradient of the log-posterior density $\nabla \log p(\theta|y)$.

A HMC iteration consists of the following steps:

1. Update $\phi$ with a draw $\phi \sim \text{Multivariate-Normal}(0, M)$.

2. Repeat the following leapfrog steps a total of $L$ times:

   (a) Use the gradient of the log-posterior density of $\theta$ to make a half-step of $\phi$:
   $$\phi \leftarrow \phi + \frac{1}{2}\epsilon \nabla \log p(\theta|y)$$

   (b) Use the momentum vector $\phi$ to update the position vector $\theta$:
   $$\theta \leftarrow \theta + \epsilon M^{-1}\phi$$

   (c) Use the gradient of the log-posterior density of $\theta$ to make a half-step of $\phi$:
   $$\phi \leftarrow \phi + \frac{1}{2}\epsilon \nabla \log p(\theta|y)$$

3. Label $\theta^{t-1}$, $\phi^{t-1}$ as the values of the parameter and momentum vectors at the start of the leapfrog process and $\theta^*$, $\phi^*$ as the values after the $L$ leapfrog steps. Calculate the accept-reject probability $r$ as
   $$r = \frac{p(\theta^*|y)p(\phi^*)}{p(\theta^{t-1}|y)p(\phi^{t-1})}.$$

4. Set
   $$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1), \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

The models in this thesis were estimated using the Stan[1] software package. Stan provides a free, ready-made, refined HMC sampler for a broad range of Bayesian models. Using Stan allows the user to avoid explicitly tuning the parameters $L$ and $\epsilon$ and determining the gradient of the log-posterior density. In Stan, the user first provides a program which specifies the distributional assumptions. Stan then translates the user-provided program to C++, compiles the C++ and runs resulting program to provide the user with a posterior sample of the model parameter using HMC. Stan also provides the user with the $\hat{R}$- and $\hat{n}_{\text{eff}}$-statistics.

## 3.4 Model comparison metrics

Model comparison in this thesis is done on the basis of predictive performance. Since the models are Bayesian, the predictions take the form of posterior predictive distributions (3.1). Naturally desirable properties of the predictive distributions include the lack of consistent over- or underprediction ('calibration') and the small dispersion of the predictive distribution ('sharpness'). The analysis of these properties can be formalized using scoring rules which evaluate a model's predictive performance based on the predictive distribution and the realized value of the response (Gneiting et al. (2007)). When a scoring rule is properly chosen, 'good' predictions, in the sense that the majority of the mass of the predictive distributions is located near the realized values, are given higher scores than 'worse' predictions, where the mass of the predictive distributions is either more dispersed or located farther away. Finally, the predictive distributions can also be transformed into point predictions which allows the use of frequentist tools for evaluating predictive performance.

### 3.4.1 Calibration and sharpness

Gneiting et al. (2007) presents how the calibration of a model's predictive distributions can be analyzed with probability integral transforms (PIT). For observation $i$, PIT $p_i$ is defined as $p_i = F(y_i)$ where the term $F$ denotes the cumulative distribution function of the predictive distribution and the term $y_i$ denotes the realized value of the response variable.

Let $G$ denote the cumulative distribution function of the true data generating process for the response variable. If the model matched the true data generating process, i.e. the equation $F = G$ would hold, the distribution of the PITs $p_i$ would be uniform. Unfortunately, the uniformity of the PITs $p_i$ is only a necessary condition as it is possible to give counterexamples where the equation

[1] *https://mc-stan.org*

$F = G$ does not hold but the PITs $p_i$ still have a uniform distribution (see e.g. Gneiting et al. (2007)). Nevertheless, the empirical PITs can be graphed as a histogram and any clear deviations from uniformity in the histogram should be taken as evidence for mismatch between the predictions and the true data generating process.

Given two models that are well-calibrated in the sense that their empirical PIT histograms are (roughly) uniform, they can be compared on the basis of how dispersed their predictive distributions are. One model could be called 'better' than the other if its predictive distributions are 'sharper', i.e. concentrated into smaller intervals, than the predictive distribution from the other model. This sharpness can be measured, for example, by tabulating the widths of the central 90 % credible intervals. These widths can be analyzed by calculating their descriptive statistics or graphing their distribution. Gneiting et al. (2007) notes that in real world applications conditional heteroscedasticity often leads to considerable variability in the widths and therefore suggests representing them using box plots.

### 3.4.2 Scoring rules

Scoring rules are summary measures that evaluate probabilistic predictions by assigning a numerical score based on the realized values and the reported predictive distributions (see e.g. Gneiting and Raftery (2007) or O'Hagan (1994, p. 56-59)). To formalize scoring rules, suppose the performance of a predictive distribution $P$ for a random variable $X$ is being evaluated. Suppose that after the predictive distribution $P$ has been fixed, realized value of the random variable $X$ is $x$. A scoring rule $S$ is then a function $S = S(P, x)$ which describes the reward for prediction $P$ given the realized value $x$. Let $S(P, Q)$ denote the expected value of reward $S(P, x)$ when $x$ is drawn from the distribution $Q$.

A scoring rule $S$ is a *proper* scoring rule if the inequality

$$S(Q, Q) \geq S(P, Q) \tag{3.7}$$

holds for all predictions $P$ and data generating processes $Q$. If the inequality (3.7) holds with equality only when the predictions match the data generating process (i.e. the equation $P = Q$ holds), the scoring rule $S$ is a *strictly proper* scoring rule. Assuming that the distribution $Q$ would now represent the total knowledge of a forecaster and $P$ the outwardly expressed prediction, under a strictly proper scoring rule the forecaster is always encouraged to state the true beliefs $Q$ outwardly, i.e. to state $P = Q$. Gneiting and Raftery (2007) notes that utility functions, which form a part of the standard formulation of inference as a statistical decision problem, give rise to proper scoring rules. Finally, scoring

rule $S = S(P, x)$ is a *local* scoring rule if, given prediction $P$, the score depends solely on the realized value $x$.

An attractive choice for a scoring rule is the logarithmic score $\mathrm{LogS}(P, x)$, defined as

$$\mathrm{LogS}(P, x) = \log p(x) \tag{3.8}$$

where $p(x)$ denotes the density for result $x$ given by distribution $P$. Bernardo and Smith (1994, p. 153) shows that all smooth (i.e. continuously differentiable), proper local scoring rules have the form $S(P, x) = A \log p(x) + B(x)$, where $A > 0$ is an arbitrary constant and $B$ is an arbitrary, subject to the existence of the expected score, function of $x$. The logarithmic score (3.8) is the obvious special case. Additionally, Gelman et al. (2014, p. 167) notes that the logarithmic score has attractive information theoretic properties in the model comparison context: With sufficiently large sample sizes, the model with the highest logarithmic score has the lowest Kullback-Leibler information and thus the highest posterior probability out of the considered models. The logarithmic score is therefore adopted for assessing the predictions in this thesis for these reasons.

### 3.4.3 Leave-one-out cross validation and PSIS-LOO

This section presents how the logarithmic score is used in this thesis in a more exact manner. Suppose data $y = (y_1, \ldots, y_n)$ is independent given parameter $\theta$. The likelihood for the model $p(y|\theta)$ then decomposes as $p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$. If the true data generating process $p_t$ was known, it would be possible to evaluate the model's performance on some new data $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_{n'})$ with the expected log pointwise predictive density, i.e. the expected logarithmic score, calculated as

$$\mathrm{elpd} = \sum_{i=1}^{n'} \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y) d\tilde{y}_i \tag{3.9}$$

where the predictive distribution $p(\tilde{y}_i|y)$ is determined by the equation (3.1). As the true data generating process $p_t$ is typically not known, it has to be approximated. Moreover, a separate data set $\tilde{y}$ is not typically available in practical settings, so the model evaluation has to be done using the data $y$ which was used for model estimation.

The true data generating process $p_t$ can be approximated through leave-one-out cross validation (LOO-CV) where the model is first estimated using data from which a single observation has been excluded and then the model performance is evaluated for the left-out observation. With the logarithmic

score, this means evaluating the expected log pointwise predictive density as

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^{n} \log p(y_i|y_{-i}) = \sum_{i=1}^{n} \log \left( \int p(y_i|\theta) p(\theta|y_{-i}) d\theta \right) \tag{3.10}$$

where the term $y_{-i}$ denotes the data from which observation $i$ has been excluded. The obvious drawback of the estimate (3.10) is that it requires estimating the model for each observation separately thereby incurring large computational costs. This makes exact LOO-CV impractical for large data sets.

Fortunately, it is possible to avoid model re-estimations with importance sampling methods that utilize the posterior samples from MCMC. Vehtari et al. (2017) presents a recent method where re-estimations are avoided by using importance sampling weights which are smoothed with a generalized Pareto distribution. To present the method, let $\theta^s, s = 1, \ldots, S$ denote sampling draws from the posterior $p(\theta|y)$. Using draw $\theta^s$ and observation $i$, define the raw importance weight $r_i^s$ as

$$r_i^s = \frac{1}{p(y_i|\theta^s)} \tag{3.11}$$

which can be used to define the importance sampling predictive distribution approximate as

$$p(y_i|y_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} r_i^s} = \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\theta^s)}}. \tag{3.12}$$

The distribution of the raw importance weights (3.11) may have a long right-tail and thereby high variance. Estimates based on the direct use of the importance weights are therefore sensitive to the largest values, so the direct use of the approximate (3.12) would suffer from instability. To overcome this issue, a generalized Pareto distribution is fitted on a fixed share of the largest importance weights and then used for smoothing the large values. For a single observation $i$, the Pareto smoothed importance sampling method of Vehtari et al. (2017) consists of the following steps:

1. A generalized Pareto distribution is fitted using 20 % of the largest importance weights $r_i^s$ defined in equation (3.11).

2. The importance weights are stabilized by replacing the $M$ largest weights with the expected values of the order statistics of the fitted Pareto distribution, e.g.

$$F^{-1}\left(\frac{z-\frac{1}{2}}{M}\right), z = 1, \ldots, M$$

where the term $M$ is the number of sampling draws used to fit the generalized Pareto distribution, i.e. $M = 0.2S$ and the term $F^{-1}$ is the inverse cumulative distribution function of the generalized Pareto distribution. The new smoothed weights are labeled as $\tilde{w}_i^s$ where $s$ indexes the sampling draw.

3. Weights $w_i^s$ for observation $i$ are generated by truncating the smoothed importance weights such that

$$w_i^s = \min(\tilde{w}_i^s, S^{3/4}\bar{w}_i)$$

where $\bar{w}_i = \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_i^s$ is the average of the smoothed weights.

Using the weights $w_i^s, s = 1, \ldots, S$ generated separately for each observation $i$, an estimate $\widehat{\text{elpd}}_{\text{psis-loo}}$ can be calculated for the expected logarithmic score (3.10) such that

$$\widehat{\text{elpd}}_{\text{psis-loo}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s}\right). \tag{3.13}$$

The $k$-parameter of the generalized Pareto distribution provides a diagnostic tool for the reliability of the weights $w_i^s$. Vehtari et al. (2017) gives the threshold for adequate performance at $\hat{k} < 0.7$. For observation $i$ with $\hat{k}_i > 0.7$, the contribution $\log(\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)/\sum_{s=1}^{S} w_i^s)$ should be replaced with the exact LOO estimate $\log p(y_i|y_{-i})$. This replacement is referred to as the PSIS-LOO+-approach.

Following Gelman et al. (2014, p. 176) and using the formula (3.13), it is also possible to define a measure of model complexity, the effective number of parameters $\text{p}_{\text{loo}}$, as

$$\text{p}_{\text{loo}} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right) - \widehat{\text{elpd}}_{\text{psis-loo}}.$$

In this thesis, the calculations of the $\widehat{\text{elpd}}_{\text{psis-loo}}$- and $\text{p}_{\text{loo}}$-estimates are done using the `loo`-package[2].

[2] *https://cran.r-project.org/web/packages/loo/*

### 3.4.4 Point predictions

A probabilistic prediction $p_i$ for observation $i$ can be transformed into a point prediction $\hat{y}_i$ by, e.g., the expected value $\hat{y}_i = \mathrm{E}_{p_i} y_i$ which can be estimated by taking the average from a sample from the predictive distribution. With the point prediction $\hat{y}_i$, the goodness of the predictions can then be measured through classical frequentist measures, e.g. the mean squared error MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ or the $R^2$-statistic,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where the term $\bar{y}$ denotes the average of the response variables $y$, i.e. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

These measures provide points of comparison with respect to models that produce only point predictions. The downside of transforming probabilistic predictions to point predictions is that the overall characterization of uncertainty, described by the predictive distribution, is lost.

## 3.5 Model stacking

The metrics described in the previous section can be used for comparing the relative predictive performance of different models. However, if the main aim of the modeling exercise is to predict well, choosing a single model with the best performance metrics may be wasteful. Rather than choosing a single model, the individual models can be combined for predictive purposes with so-called model stacking described in Yao et al. (2018). This model combination method is described next.

Given true data generating process $p_t$, a set of models $M = \{M_1, \ldots, M_K\}$, data $y = (y_1, \ldots, y_n)$, a scoring rule $S = S(p, q)$ over distributions $p$ and $q$, model stacking is formulated as an optimization problem for model weights $w = (w_1, \ldots, w_K)$ of the form

$$\max_{w} \; S\left(\sum_{k=1}^{K} w_k p(\cdot|y, M_k), p_t(\cdot|y)\right) \text{ s.t. } 0 \leq w_k \leq 1, \sum_{k=1}^{K} w_k = 1. \tag{3.14}$$

Since the true data generating process $p_t$ is again not known in practical applications, it needs to be approximated by the leave-one-out approach described in the previous section. Likewise, to finalize the operationalization of the optimization problem (3.14), the scoring rule $S$ has to be chosen. Yao et al. (2018) states a preference for the logarithmic score $S(p, q) = \log(p(q))$ since it is

equivalent to every other proper local scoring rule, as discussed in section 3.4.2. With the logarithmic score and the leave-one-out approach, the optimization problem becomes

$$\max_{w} \ \frac{1}{n}\sum_{i=1}^{n}\log\sum_{k=1}^{K} w_k p(y_i|y_{-i}, M_k) \text{ s.t. } 0 \leq w_k \leq 1, \sum_{k=1}^{K} w_k = 1. \tag{3.15}$$

The optimization problem (3.15) is solved in this thesis using the ready-made functionality provided by the `loo`-package.

The solution $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_K)$ for optimization problem (3.15) is referred to as the stacking model in this thesis. The weights $\hat{w}$ define the stacked estimate for the predictive distribution as

$$\hat{p}(\tilde{y}|y) = \sum_{k=1}^{K} \hat{w}_k p(\tilde{y}|y, M_k). \tag{3.16}$$

The predictive distribution (3.16) can be evaluated in the same way as the predictive distributions from any single model.

# Chapter 4

# Data

The data used in this thesis consists of two parts, housing sales data and the geographical data related to the neighborhoods of Helsinki, Espoo and Vantaa. This chapter describes these data sets.

## 4.1 Housing sales data

The housing sales data has been compiled from the *Asuntojen.hintatiedot.fi*-service of the Finnish Ministery of the Environment and the Housing Finance and Development Centre of Finland for the cities of Helsinki, Espoo and Vantaa. The service compiles sales records provided by a set[1] of real estate agencies or related organizations. It is unclear whether the housing sales from just these sources are representative for 'normal housing sales' in Helsinki, Espoo and Vantaa.

The service stores each record of a housing sale for a total of 12 months starting from the date when the sale was registered to the service. Due to normal processing times at the real estate agencies, the exact sale dates for the entries have occurred typically one to two months before the sale is registered to the service. The housing sales data was retrieved on 23.12.2018 via web scraping.

For a single recorded sale, the service provides information on

- the neighborhood name (a free text field),
- the type of the apartment (a free text field),
- the type of the building,

[1]List of the real estate agencies and organizations, retrieved on 8/2019: KVKL - Kiinteistönvälitysalan Keskusliitto Ry, Kiinteistömaailma Oy, OP-Kiinteistökeskus, Huoneistokeskus Oy, SKV Kiinteistönvälitys Oy, Aktia Kiinteistönvälitys Oy, RE/MAX Suomi

- the size of the apartment (measured in square meters),
- the price,
- the price per square meters,
- the year the building was built,
- the floor where the apartment is located (a free text field),
- whether the building has a elevator,
- the reported condition of the apartment

and

- the energy classification of the building.

Appendix A contains tables for the descriptive statistics for the available information after the data was preprocessed. Due to a map error, the sales in the Kalajärvi neighborhood (a total of 3 recorded sales) in Espoo were excluded from the data. It was felt that the omission of this neighborhood would not influence any estimates materially. Data preprocessing is described next.

To facilitate the construction of hierarchical models on the basis of neighborhoods, the neighborhood names needed to be uniformized by formatting. The formatting was done manually to match the recorded names to official neighborhood names. For example, if the raw recorded neighborhood was 'Alppila', it was transformed to 'Alppiharju'.

The type of the apartment is a free text field with values such as '1 h, kk, kph', '1h, kt, lasitet...' and '1h+tk+s'. In the data preprocessing, this field was uniformized by removing the white spaces, transforming everything to lower case and replacing the ','-characters with '+'-characters. Two variables were created from the uniformized field. First, a dummy describing whether the apartment had a sauna was created by checking whether the character string '+s' was present in the uniformized field. Next, an integer describing the number of rooms in the apartment was constructed by retrieving the integer next to the 'h'-character. For example, apartment with unprocessed type '1-2 h, kk, p' is recorded to have a total of two rooms.

The information regarding the floor of the apartment is recorded in a free text field with values such as '5/12' and '-1/6'. Based on this field, the floor of the apartment is read from the digits before the '/'-character. However, when the free text field has the '-1'-substring, the floor of the apartment is read from the digits after the '/'-character. This treatment was adopted on the assumption that the use of the '-1'-substring is related mostly to sales of either a town house

or a row house - these apartments can have multiple floors, so the floor variable would now describe the highest floor of the apartment.

Table 4.1 gives the descriptive statistics for the sizes, prices and prices per square meters for the recorded sales. The minimum values for the variables were regarded as indications of data entry errors in the raw data. Checking the data on an observation-by-observation basis revealed that there was a single aberrant observation which had both the suspiciously low price and the suspiciously small size. There are also other indications of possible data entry errors, e.g. there are recorded sales with apartments in buildings with a total of 9 floors but which do not have an elevator. Additional errors may have been introduced by how the data was preprocessed: There are observations where the building type is a high-rise building and the raw floor variable has the '-1'-substring, so the chosen treatment for the floor variable might have lead to further data errors.

| Variable | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Square meters | 12.50 | 45.00 | 60.50 | 67.28 | 82.00 | 320.00 |
| Price | 865 | 175000 | 240000 | 280734 | 330000 | 2500000 |
| Price per square meters | 69 | 3014 | 4174 | 4465 | 5636 | 14214 |

Table 4.1: Continuous variables, descriptive statistics (Kalajärvi excluded)

The number of rooms and the reported condition of the apartment were transformed further for model estimation. First, the number of rooms was transformed into a set of dummy variables that were used as explanatory variables as such. Next, an interaction variable was generated using the size of the apartment and a dummy variable encoding whether reported condition of the apartment is good.

The variables are referenced in the model specifications with the following terms. For observation $i$, term

- $\text{Sqm}_i$ is the size of the apartment measured in square meters,
- $\text{GoodConditionSqm}_i$ is the interaction term[2] of the size of the apartment and the reported condition of the apartment,
- $\text{Age}_i$ is the age of the apartment's building measured as the difference between the year the building was built and 2018 (i.e. $\text{Age}_i = 2018 - \text{Year}_i$ where the term $\text{Year}_i$ is the building year),
- $\text{TwoRoomsDummy}_i$ is a dummy variable describing whether the apartment had 2 rooms,

[2]The interaction term $\text{GoodConditionSqm}_i$ gets the same value as variable $\text{Sqm}_i$ if the reported condition of the apartment $i$ was good and 0 otherwise.

- ThreeRoomsDummy$_i$ is a dummy variable describing whether the apartment had 3 rooms,
- FourRoomsOrModeDummy$_i$ is a dummy variable describing whether the apartment had 4 rooms or more,
- OwnFloor$_i$ is the floor variable processed from the raw floor text field,
- SaunaDummy$_i$ is a dummy variable describing whether the apartment had a sauna

and finally

- Price$_i$ is the observed price.

## 4.2 Geographical data

It seems natural that the price of an apartment would increase as the desirability of the apartment's neighborhood increases. While the desirability of a neighborhood most likely depends on multiple factors, it seems natural to assume that these factors include the distance from the neighborhood to the center of the metropolitan area (assumed effect is 'the closer the better') and the distance to the ocean (assumed effect is 'the closer the better'). To model these types of effects along with the possible interdependence of the desirability of neighborhoods, geographical data was compiled.

The geographical data was derived from a map[3], seen in figure 4.1, that was constructed using open geographical data with QGIS[4]. The neighborhood polygons were retrieved from the open geographical services of Helsinki, Espoo and Vantaa. As can be seen from the figure, the Kalajärvi neighborhood in Espoo is missing. The centroids of the neighborhood polygons are denoted with red dots in the figure. The road network graph was retrieved from open data provided by the National Land Survey of Finland. The ocean polygon was provided by OpenStreetMap. A single point was manually added roughly next to the central railway station of Helsinki to represent the 'center of the metropolitan area'. This center point is denoted with the orange star in the figure.

Three different distances were measured using the compiled map. First, the inter-centroid distance $D_{ij}$ measures the direct distance between the centroids of neighborhoods $i$ and $j$. Second, the distance OceanDistance$_i$ measures the direct distance from centroid of neighborhood $i$ to the ocean polygon. Third,

[3] *https://github.com/villemakinen/housingprices/tree/master/map*
[4] *https://qgis.org*

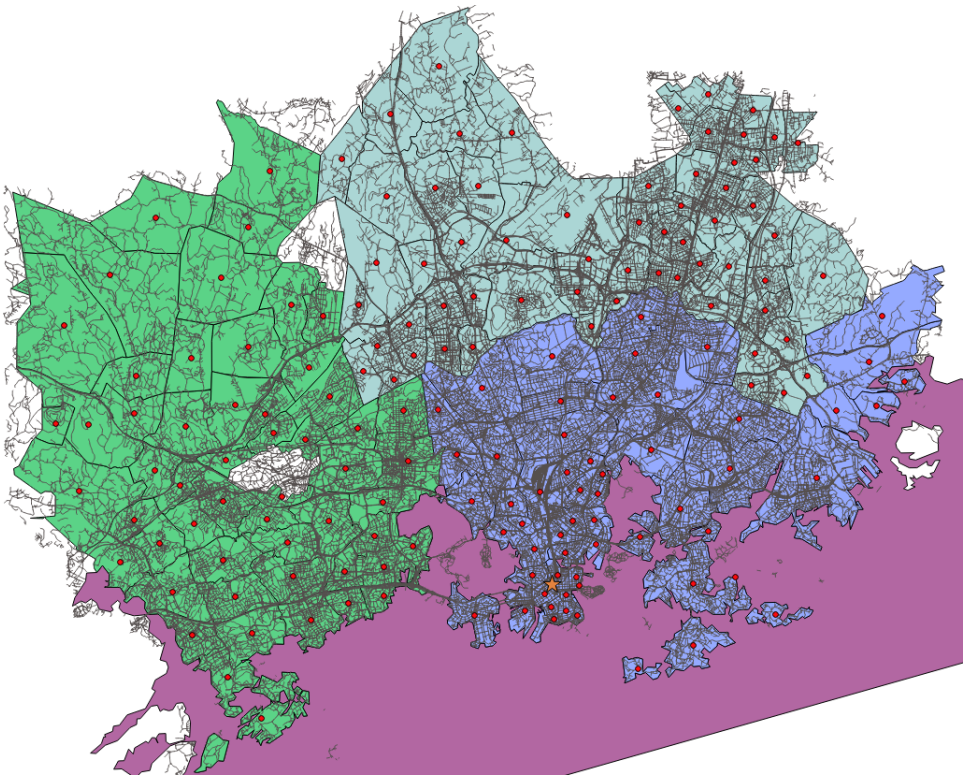

Figure 4.1: Neighborhood map

the distance RoadDistance$_i$ measures the distance along the road network from centroid of neighborhood $i$ to the point representing the 'center of the metropolitan area'. The distances OceanDistance$_i$ and RoadDistance$_i$ were recorded in meters and the distances $D_{ij}$ in kilometers. Table A.11 gives the descriptive statistics for the measured distances.

# Chapter 5

# Overall modeling approach

This chapter discusses the necessary distributional choices, the modeling work flow and model checking.

## 5.1 Distributional choices

Since there are indications of data errors (see e.g. the minimum price in table 4.1), it was felt that a robust model should be used, so Student's $t$-distribution was chosen as the likelihood. The distribution's heavier tails allow the models to better accommodate the unusual observations (Gelman et al. (2014, p. 437)). Moreover, O'Hagan (1979) shows that Student's $t$-distribution is *outlier-prone of degree* 1 meaning that the effect on the posterior of a total of $m$ outliers becomes negligible as long as there is at least a total of $2m$ observations altogether.

The use of Student's $t$-distribution as a robust likelihood is well-established in Bayesian statistics. Seminal articles where the Student's $t$-distribution has been used for Bayesian linear regression include West (1984), which discusses the construction of heavy-tailed distributions through mixtures of normal distributions and their use both as likelihoods and priors, and Geweke (1993), which describes a Gibbs sampler for a linear regression model with a Student's $t$-distribution likelihood. In the context of Gaussian process models, Student's $t$-distribution has been used in, at least, Vanhatalo et al. (2009), Jylänki et al. (2011) and Hartman and Vanhatalo (2019).

The prior distribution choices for the model parameters are essentially conventional and follow examples found in McElreath (2016) where possible. The normal distribution was used for the regression coefficients since this was seen as the simplest choice. The half-Cauchy distribution was used for the variance terms since this choice follows both McElreath (2016) and the recommendations of Gelman (2006) regarding the variance terms for the group-level effects. The

half-Cauchy distribution was used also for the covariance parameters in model 4. For the degrees of freedom parameter $\nu$ of the Student's $t$ likelihood, the Gamma distribution was used on the basis of recommendations[1] from the Stan community. Lastly, the LKJ distribution was used for the correlation matrix in model 5.

To allow for easy interpretation of the models, the explanatory variables were not scaled or centered. No 'standard' choices were therefore available for the hyperparameters for regression coefficients. Similarly, no clear hyperparameter choices were available for the rest of the parameters. The values of the hyperparameters for the prior were therefore chosen emulating the generative approach presented in Gelman et al. (2017) and Gabry et al. (2019). First, expectations were formed for the direction of influence of each explanatory variable. For example, it was expected that the price would tend to increase as the size of the apartment increases. Next, the hyperparameter values for the whole prior were chosen through multiple rounds of simulating data using a given set of hyperparameter values. When the results from the multiple data simulation rounds were deemed to be broadly reasonable, the hyperparameter values were accepted. Otherwise, the hyperparameter values were tuned and the simulation rounds continued.

A single round of simulations had three steps (for a fixed set of hyperparameter values):

1. Simulated data for the explanatory variables was drawn from a set of independent one-dimensional distributions. For example, the values for the sizes of the apartments were drawn from a Gamma distribution for which the shape and rate parameters were chosen such that the expected value matched the observed mean size from the true data.

2. The values for the model parameters were drawn from the prior with the fixed set of hyperparameter values.

3. The simulated prices were drawn from the likelihood using the explanatory variable data simulated in step 1 and the model parameters simulated in step 2. These prices were used to draw a histogram and the histogram was checked for obvious inconsistencies, e.g. a large number of negative or abnormally large prices.

A prior distribution can be called *weakly informative* when it is proper but it has been set up in such a way that the information it provides is intentionally weaker than the actual available prior knowledge (Gelman et al. (2014, p. 55)).

[1] *https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations*

Arguably the prior distribution resulting from the described procedure is weakly informative as the resulting hyperparameter values could be deemed to be acceptable even when the simulated data included some negative prices, which are not seen in real life. Alternatively, if 'intentional weakening' of the prior is defined through artificially inflated[2] dispersion, then calling the resulting prior weakly informative might be difficult as it was not possible to formulate prior beliefs in terms of explicit distributions.

## 5.2 Modeling work flow

Before starting to build the models, the housing sales data was split into separate estimation and test sets with 7-3-split using simple random sampling. The model estimation was done using the estimation set and the test set was left to represent a hypothetical future data for testing the predictive performance of the models. The same estimation and test sets were used for each of the models.

The modeling work flow consisted of the following steps:

1. Calibrate hyperparameters choices for the prior distribution as described in the previous section.
2. Generate an additional simulated data set.
3. Write the Stan program for the model.
4. Check correctness of the Stan program by fitting the model on simulated data.
5. Fit the model using the estimation set.
6. Perform model checking using the estimation set.
7. Evaluate the predictive performance on the estimation and test sets.

The correctness of the Stan program was checked by fitting the model on the additional simulated data and analyzing the resulting posterior distribution. The Stan program was deemed to be correct when the posterior distributions of the model parameters were located around the parameter values that were used for generating the data.

[2]For example, the prior distribution for coefficient $\beta$ is intentionally chosen as $\beta \sim N(\mu, K\sigma^2), K > 1$ when the distribution $\beta \sim N(\mu, \sigma^2)$ describes the actual prior belief over the coefficient.

## 5.3 Model checking

Model checking was done using draws from the predictive distributions following the general approach presented in chapter 6 of Gelman et al. (2014). Given a set of test quantities which serve as summaries for the important aspects of the data, the predictive distribution draws were used to derive distributions for the test quantities. If a test quantity calculated from the true data is located in a high-mass region in the distribution derived from the predictive distribution draws, it can be concluded that the model is believable at least in the aspect measured by this test quantity.

In more exact terms, the model checking proceeded as follows: For each posterior sample $\theta^s = (\nu^s, \sigma^s, \beta^s)$ and observation $i$, a replicated price $v_i^{\text{rep},s}$ was drawn from the likelihood for which the explanatory variables $x_i$ and the coefficients $\beta^s$ set the expected value $\mu_i^s$, i.e.

$$v_i^{\text{rep},s} \sim t_{\nu^s}(\mu_i^s, \sigma^s).$$

The individual replicated price $v_i^{\text{rep},s}$ are then used to compile a replicated data set $v^{\text{rep},s}$ such that $v^{\text{rep},s} = (v_1^{\text{rep},s}, \ldots, v_n^{\text{rep},s})$. A sample of test quantity values $T(v^{\text{rep},s})$ can then be generated for a chosen test quantity $T$ using the replicated data sets $v^{\text{rep},s}$. The resulting sample can be used for graphing a histogram, for example. Then, a 'true' value $T(v)$ for the test quantity can be calculated from the true data. Assuming that the model fits the data, the value $T(v)$ should be located within a high-mass region of the $T(v^{\text{rep},s})$-histogram.

The choice of an appropriate test quantity $T$ depends on what is being modeled. The choices for this thesis were as follows: First, using the whole data, the the mean and median prices were calculated. Then, separately for each neighborhood, the neighborhood-specific mean prices were calculated. These test quantities were chosen due to their simplicity and their ease of interpretation.

Other possible test quantities include, for example, the maximum and the minimum prices calculated for the whole data or for each neighborhood separately. The minimum and maximum prices were not used in this thesis since the heavy tails of the likelihood means that the replicated data sets will likely include artificially small or large values.

# Chapter 6

# Models

This chapter presents the structure and estimates for the individual models.

Models 2-5 include terms that vary across neighborhoods, so the following notation is adopted for the group-level effects: For observation $i$, let $j[i]$ denote the index of the neighborhood where the apartment is located. The term $\beta_{\text{Intercept},j[i]}$ then refers to the intercept for the apartment's neighborhood, for example.

The group-level effects for models 2-5 were written using non-centered parameterization (see e.g. Gelman et al. (2014, p. 394) or McElreath (2016, p. 408)) as this approach was more efficient for sampling with HMC. To illustrate the approach, let $\beta_j$ denote a group varying coefficient which is assumed to have a normal distribution with hyperparameters $\mu_0$ and $\sigma_0^2$. The distribution for $\beta_j$ was written in Stan as $\beta_j = \mu_0 + \sigma_0 z,\ z \sim N(0,1)$ instead of the simpler $\beta_j \sim N(\mu_0, \sigma_0^2)$.

The Stan programs used for estimating the models presented in this thesis, along with the supplementary R scripts, are provided in a separate GitHub repository[1].

## 6.1 Model 1 - Simple regression

Model 1 is a simple regression model. The model does not utilize any information about the neighborhoods where the apartments are located, so it serves as a benchmark for more realistic models.

### 6.1.1 Model specification

List (6.1) gives the prior distribution choices for model 1:

[1] *https://github.com/villemakinen/housingprices*

$$
\begin{aligned}
\beta_{\text{Intercept}} &\sim \text{N}(70000, 50000^2) \\
\beta_{\text{Sqm}} &\sim \text{N}(4500, 1000^2) \\
\beta_{\text{GoodConditionSqm}} &\sim \text{N}(1000, 1000^2) \\
\beta_{\text{Age}} &\sim \text{N}(-1500, 2000^2) \\
\beta_{\text{TwoRoomsDummy}} &\sim \text{N}(5000, 10000^2) \\
\beta_{\text{ThreeRoomsDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{FourRoomsOrMoreDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{OwnFloor}} &\sim \text{N}(7000, 1000^2) \\
\beta_{\text{SaunaDummy}} &\sim \text{N}(5000, 2500^2) \\
\sigma &\sim \text{Half-Cauchy}(0, 15000) \\
\nu &\sim \text{Gamma}(2, 0.1)
\end{aligned} \tag{6.1}
$$

The likelihood is $\text{Price}_i \sim t_\nu(\mu_i, \sigma)$ where the expected value $\mu_i$ is determined by the sum

$$
\begin{aligned}
\mu_i = \; & \beta_{\text{Intercept}} + \\
& \beta_{\text{Sqm}} \, \text{Sqm}_i + \\
& \beta_{\text{GoodConditionSqm}} \, \text{GoodConditionSqm}_i + \\
& \beta_{\text{Age}} \, \text{Age}_i + \\
& \beta_{\text{TwoRoomsDummy}} \, \text{TwoRoomsDummy}_i + \\
& \beta_{\text{ThreeRoomsDummy}} \, \text{ThreeRoomsDummy}_i + \\
& \beta_{\text{FourRoomsOrMoreDummy}} \, \text{FourRoomsOrModeDummy}_i + \\
& \beta_{\text{OwnFloor}} \, \text{OwnFloor}_i + \\
& \beta_{\text{SaunaDummy}} \, \text{SaunaDummy}_i.
\end{aligned}
$$

### 6.1.2 Estimates

Table 6.1 gives the descriptive statistics for the posterior distribution for the model parameters.

Surprisingly the estimates in table 6.1 indicate that the effect of the age of the apartment is reversed when going from the prior to the posterior: The majority of the mass of the posterior is located on positive numbers for the $\beta_{\text{Age}}$-coefficient, so under the posterior the price would be expected to increase as the age of the apartment increases.

| Parameter | mean | sd | 2.5% q. | 50% q. | 97.5% q. | $\hat{n}_{\text{eff}}$ | $\hat{R}$ |
|---|---|---|---|---|---|---|---|
| $\beta_{\text{Intercept}}$ | 35595.91 | 5241.74 | 25302.66 | 35580.94 | 45898.14 | 4534.08 | 1.00 |
| $\beta_{\text{Sqm}}$ | 1683.59 | 81.86 | 1528.36 | 1683.20 | 1842.71 | 4610.05 | 1.00 |
| $\beta_{\text{GoodConditionSqm}}$ | 953.66 | 46.96 | 860.53 | 953.82 | 1044.05 | 7487.12 | 1.00 |
| $\beta_{\text{Age}}$ | 808.02 | 52.53 | 703.63 | 808.58 | 909.19 | 5673.59 | 1.00 |
| $\beta_{\text{TwoRoomsDummy}}$ | 5286.38 | 3290.30 | -1127.12 | 5252.59 | 11605.64 | 5929.39 | 1.00 |
| $\beta_{\text{ThreeRoomsDummy}}$ | 7362.02 | 4455.38 | -1267.17 | 7369.99 | 15902.57 | 5141.55 | 1.00 |
| $\beta_{\text{FourRoomsOrMoreDummy}}$ | 15244.42 | 6254.10 | 3225.91 | 15231.68 | 27297.77 | 4663.38 | 1.00 |
| $\beta_{\text{OwnFloor}}$ | 9847.57 | 614.24 | 8637.13 | 9851.00 | 11025.51 | 7204.09 | 1.00 |
| $\beta_{\text{SaunaDummy}}$ | 6225.31 | 2035.01 | 2238.49 | 6222.88 | 10137.16 | 7749.65 | 1.00 |
| $\sigma$ | 65017.19 | 1465.32 | 62197.65 | 65016.24 | 67933.27 | 6012.19 | 1.00 |
| $\nu$ | 2.57 | 0.13 | 2.32 | 2.56 | 2.83 | 6148.65 | 1.00 |

Table 6.1: Parameter estimates - model 1

## 6.2 Model 2 - Varying intercepts model

Model 2 is a simple varying intercepts model where the intercept is allowed to vary across the neighborhoods. The model structure was chosen to mimic the examples in chapter 12 of McElreath (2016).

### 6.2.1 Model specification

The neighborhood-specific intercepts $\beta_{\text{Intercept},j}$ are drawn from a normal distribution representing the 'population' of neighborhoods. List (6.2) describes the structure for the group-level effects:

$$\begin{aligned} \mu_{\text{Intercept}} &\sim \text{N}(50000, 50000^2) \\ \sigma_{\text{Intercept}} &\sim \text{Half-Cauchy}(0, 11000) \\ z_j &\sim \text{N}(0, 1) \\ \beta_{\text{Intercept},j} &= \mu_{\text{Intercept}} + \sigma_{\text{Intercept}} z_j \end{aligned} \tag{6.2}$$

List (6.3) gives the rest of the prior choices for model 2:

$$
\begin{aligned}
\beta_{\text{Sqm}} &\sim \text{N}(4000, 1000^2) \\
\beta_{\text{GoodConditionSqm}} &\sim \text{N}(1000, 1000^2) \\
\beta_{\text{Age}} &\sim \text{N}(0, 2000^2) \\
\beta_{\text{TwoRoomsDummy}} &\sim \text{N}(5000, 10000^2) \\
\beta_{\text{ThreeRoomsDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{FourRoomsOrMoreDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{OwnFloor}} &\sim \text{N}(1000, 1000^2) \\
\beta_{\text{SaunaDummy}} &\sim \text{N}(5000, 2500^2) \\
\sigma &\sim \text{Half-Cauchy}(0, 15000) \\
\nu &\sim \text{Gamma}(2, 0.1)
\end{aligned} \tag{6.3}
$$

The likelihood is $\text{Price}_i \sim t_\nu(\mu_i, \sigma)$ where the expected value $\mu_i$ is determined by the sum

$$
\begin{aligned}
\mu_i = \; & \beta_{\text{Intercept},j[i]} + \\
& \beta_{\text{Sqm}} \ \text{Sqm}_i + \\
& \beta_{\text{GoodConditionSqm}} \ \text{GoodConditionSqm}_i + \\
& \beta_{\text{Age}} \ \text{Age}_i + \\
& \beta_{\text{TwoRoomsDummy}} \ \text{TwoRoomsDummy}_i + \\
& \beta_{\text{ThreeRoomsDummy}} \ \text{ThreeRoomsDummy}_i + \\
& \beta_{\text{FourRoomsOrMoreDummy}} \ \text{FourRoomsOrModeDummy}_i + \\
& \beta_{\text{OwnFloor}} \ \text{OwnFloor}_i + \\
& \beta_{\text{SaunaDummy}} \ \text{SaunaDummy}_i.
\end{aligned}
$$

### 6.2.2 Estimates

Table 6.2 gives the descriptive statistics for the posterior distribution of the model parameters. Figure B.1 in appendix B describes the posterior distributions for the neighborhood-specific intercepts $\beta_{\text{Intercept},j}$ for model 2.

Comparing these estimates with those of model 1, it can be seen that the effect of age on the price now works in the expected direction, i.e. the price of the apartment decreases as the apartment ages. The average effect of the floor variable has also more than doubled compared to the prior distribution choices. However, the $\hat{R}$-statistic for the group-level mean $\mu_{\text{Intercept}}$ is greater than the

| Parameter | mean | sd | 2.5% q. | 50% q. | 97.5% q. | $\hat{n}_{\text{eff}}$ | $\hat{R}$ |
|---|---|---|---|---|---|---|---|
| $\beta_{\text{Sqm}}$ | 2027.24 | 73.00 | 1890.33 | 2026.20 | 2171.99 | 4104.02 | 1.00 |
| $\beta_{\text{GoodConditionSqm}}$ | 555.33 | 27.11 | 502.58 | 555.26 | 608.21 | 8066.64 | 1.00 |
| $\beta_{\text{Age}}$ | -1132.39 | 50.77 | -1234.57 | -1132.57 | -1032.10 | 6546.75 | 1.00 |
| $\beta_{\text{TwoRoomsDummy}}$ | 10095.72 | 2249.52 | 5725.86 | 10084.09 | 14419.52 | 4688.93 | 1.00 |
| $\beta_{\text{ThreeRoomsDummy}}$ | 12121.05 | 3281.41 | 5726.30 | 12143.39 | 18447.92 | 3815.76 | 1.00 |
| $\beta_{\text{FourRoomsOrMoreDummy}}$ | 21667.82 | 4924.33 | 11810.41 | 21797.96 | 31205.34 | 3659.20 | 1.00 |
| $\beta_{\text{OwnFloor}}$ | 2433.37 | 412.00 | 1630.01 | 2434.16 | 3236.64 | 8511.35 | 1.00 |
| $\beta_{\text{SaunaDummy}}$ | 10332.65 | 1590.03 | 7161.58 | 10329.42 | 13437.81 | 9168.64 | 1.00 |
| $\sigma$ | 33195.75 | 836.64 | 31582.90 | 33187.38 | 34884.66 | 6114.17 | 1.00 |
| $\nu$ | 1.99 | 0.09 | 1.82 | 1.99 | 2.17 | 6279.33 | 1.00 |
| $\mu_{\text{Intercept}}$ | 104144.63 | 9031.64 | 86836.82 | 104090.54 | 122017.77 | 343.05 | 1.02 |
| $\sigma_{\text{Intercept}}$ | 90700.10 | 6161.38 | 79783.23 | 90370.51 | 103848.28 | 477.55 | 1.01 |

Table 6.2: Parameter estimates - model 2

1.01 threshold given in McElreath (2016), so more samples should have been drawn from the posterior.

## 6.3 Model 3 - Varying intercepts model with distance measures

Model 3 is the obvious extension of model 2 where the group-level intercepts are influenced by the road and ocean distances discussed in section 4.2. Similar extensions are presented in Gelman and Hill (2006, p. 241).

The attractive property of model 3 is that it shows how group-level information can be utilized for predictions at the observation-level. Models of this type could be used to study whether investment on neighborhood infrastructure is reflected in apartment prices, for example. Moreover, the models of this type should be better at extrapolating price predictions in neighborhoods for which the necessary distance data exists but no observations have been recorded in the estimation set.

### 6.3.1 Model specification

For model 3, the intercepts are now determined through a group-level linear regression model where the road and ocean distances are used as explanatory variables. The intercept for neighborhood $j = 1, \ldots, 172$[2] is determined by the group-level model described by list (6.4):

[2]The indexes run up to 172 since they now include also the neighborhoods where no housing sales were recorded.

$$
\begin{aligned}
\alpha_{\text{Intercept}} &\sim \text{N}(150000, 50000^2) \\
\alpha_{\text{OceanDistance}} &\sim \text{N}(-5, 3^2) \\
\alpha_{\text{RoadDistance}} &\sim \text{N}(-5, 3^2) \\
\sigma_{\text{Intercept}} &\sim \text{Half-Cauchy}(0, 20000) \\
z_j &\sim \text{N}(0, 1) \\
\beta_{\text{Intercept},j} &= \alpha_{\text{Intercept}} + \\
&\quad \alpha_{\text{OceanDistance}}\ \text{OceanDistance}_j + \\
&\quad \alpha_{\text{RoadDistance}}\ \text{RoadDistance}_j + \\
&\quad \sigma_{\text{Intercept}} z_j
\end{aligned} \tag{6.4}
$$

List (6.5) gives the rest of the prior distribution choices for model 3:

$$
\begin{aligned}
\beta_{\text{Sqm}} &\sim \text{N}(5000, 1000^2) \\
\beta_{\text{GoodConditionSqm}} &\sim \text{N}(2000, 1000^2) \\
\beta_{\text{Age}} &\sim \text{N}(-1000, 1000^2) \\
\beta_{\text{TwoRoomsDummy}} &\sim \text{N}(5000, 10000^2) \\
\beta_{\text{ThreeRoomsDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{FourRoomsOrMoreDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{OwnFloor}} &\sim \text{N}(1000, 1000^2) \\
\beta_{\text{SaunaDummy}} &\sim \text{N}(5000, 2500^2) \\
\sigma &\sim \text{Half-Cauchy}(0, 15000) \\
\nu &\sim \text{Gamma}(2, 0.1)
\end{aligned} \tag{6.5}
$$

The likelihood is $\text{Price}_i \sim t_\nu(\mu_i, \sigma)$ where the expected value $\mu_i$ is determined by the sum

$$
\begin{aligned}
\mu_i \quad = \quad & \beta_{\text{Intercept},j[i]} + \\
& \beta_{\text{Sqm}}\ \text{Sqm}_i + \\
& \beta_{\text{GoodConditionSqm}}\ \text{GoodConditionSqm}_i + \\
& \beta_{\text{Age}}\ \text{Age}_i + \\
& \beta_{\text{TwoRoomsDummy}}\ \text{TwoRoomsDummy}_i + \\
& \beta_{\text{ThreeRoomsDummy}}\ \text{ThreeRoomsDummy}_i + \\
& \beta_{\text{FourRoomsOrMoreDummy}}\ \text{FourRoomsOrModeDummy}_i + \\
& \beta_{\text{OwnFloor}}\ \text{OwnFloor}_i + \\
& \beta_{\text{SaunaDummy}}\ \text{SaunaDummy}_i.
\end{aligned}
$$

### 6.3.2 Estimates

Table 6.3 gives the descriptive statistics for the posterior distribution of the model parameters. Figure B.2 in appendix B describes the posterior distributions for the neighborhood-specific intercepts for model 3.

| Parameter | mean | sd | 2.5% q. | 50% q. | 97.5% q. | $\hat{n}_{\text{eff}}$ | $\hat{R}$ |
|---|---|---|---|---|---|---|---|
| $\beta_{\text{Sqm}}$ | 2061.53 | 70.83 | 1925.64 | 2060.37 | 2202.35 | 6038.30 | 1.00 |
| $\beta_{\text{GoodConditionSqm}}$ | 556.21 | 26.53 | 503.98 | 555.70 | 609.36 | 13749.18 | 1.00 |
| $\beta_{\text{Age}}$ | -1137.56 | 49.04 | -1236.56 | -1137.38 | -1042.66 | 10172.75 | 1.00 |
| $\beta_{\text{TwoRoomsDummy}}$ | 9727.64 | 2269.99 | 5349.39 | 9736.52 | 14260.48 | 6646.34 | 1.00 |
| $\beta_{\text{ThreeRoomsDummy}}$ | 11009.79 | 3197.04 | 4797.43 | 11012.36 | 17255.47 | 5351.71 | 1.00 |
| $\beta_{\text{FourRoomsOrMoreDummy}}$ | 20098.59 | 4875.50 | 10435.47 | 20140.87 | 29571.48 | 5380.90 | 1.00 |
| $\beta_{\text{OwnFloor}}$ | 2379.42 | 426.56 | 1525.02 | 2378.46 | 3222.84 | 15302.33 | 1.00 |
| $\beta_{\text{SaunaDummy}}$ | 10478.75 | 1582.91 | 7419.20 | 10484.28 | 13545.04 | 15726.36 | 1.00 |
| $\sigma$ | 33222.39 | 823.16 | 31663.19 | 33209.31 | 34870.29 | 9665.80 | 1.00 |
| $\nu$ | 1.99 | 0.09 | 1.83 | 1.99 | 2.17 | 10537.20 | 1.00 |
| $\alpha_{\text{Intercept}}$ | 245452.50 | 10406.41 | 224501.57 | 245546.45 | 265798.18 | 1102.08 | 1.00 |
| $\alpha_{\text{OceanDistance}}$ | -1.00 | 1.56 | -4.06 | -0.97 | 1.95 | 1482.89 | 1.00 |
| $\alpha_{\text{RoadDistance}}$ | -10.78 | 0.97 | -12.66 | -10.77 | -8.86 | 1417.34 | 1.00 |
| $\sigma_{\text{Intercept}}$ | 45989.01 | 3438.07 | 39892.54 | 45803.56 | 53119.29 | 1647.43 | 1.00 |

Table 6.3: Parameter estimates - model 3

Comparing table 6.3 to table 6.2, it can be seen that the non-group-level coefficient estimates are essentially the same for both models. Nevertheless, model 3 gives clear indications that there exists an relationship between a neighborhood's desirability, as measured by the intercepts, and the road distance from the neighborhood centroids to the center of the metropolitan area.

## 6.4 Model 4 - Gaussian process model

Model 4 is a Gaussian process model where the neighborhood intercepts are assumed to be simultaneously drawn from a specific group-level multivariate normal distribution. The covariance structure of the multivariate normal distribution depends on the distances between the neighborhood centroids - the smaller the distance is, the higher the covariance between the intercept terms of these neighborhoods is. This in turn means that the intercepts of nearby neighborhoods tend to be similar ceteris paribus. The model structure was chosen to mimic the example in section 13.4 of McElreath (2016).

### 6.4.1 Model specification

For model 4, the construction of the group-level model starts with the construction of its covariance matrix $K$. The elements $K_{ij}$ of the covariance matrix $K$ are determined by the inter-centroid distances as

$$K_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + 0.01\ \delta_{ij} \tag{6.6}$$

where the term $D_{ij}$ is the inter-centroid distance of neighborhoods $i$ and $j$ and the term $\delta_{ij}$ is the Kronecker delta. This covariance structure is a special case of the isotropic Gaussian covariance function (Banerjee et al. (2015, p. 28)). Parameters $\eta^2$ and $\rho^2$ have the prior distributions

$$\rho^2 \sim \text{Half-Cauchy}(0, 1)$$

and

$$\eta^2 \sim \text{Half-Cauchy}(0, 300).$$

To speed up sampling, the model definition utilizes the Cholesky decomposition of the covariance matrix $K$. Let $L$ denote the Cholesky decomposition of the covariance matrix $K$, i.e. matrix $K$ decomposes as $K = LL'$. The unscaled intercepts $\alpha$ are then given by the following multivariate normal distribution, written with non-centered parameterization,

$$\begin{aligned} z &\sim \text{Multivariate-Normal}([0, \ldots, 0]', I_{128}) \\ \alpha &= [70, \ldots, 70]' + Lz \end{aligned}$$

where the term $I_{128}$ denotes a $128 \times 128$ identity matrix and the vectors $[0, \ldots, 0]'$ and $[70, \ldots, 70]'$ have a total of 128 elements. The constant vector $[70, \ldots, 70]'$

serves as the expected value vector of the multivariate normal distribution.

To complete the group-level model, the neighborhood-specific intercepts $\beta_{\text{Intercept}}$ are determined by scaling the $\alpha$-parameter such that

$$\beta_{\text{Intercept}} = 1000\alpha.$$

This scaling was adopted to ease the tuning of the hyperparameter values for the $\rho^2$ and $\eta^2$ parameters.

List (6.7) gives the rest of the prior distribution choices for model 4:

$$\begin{aligned}
\beta_{\text{Sqm}} &\sim \text{N}(6000, 3000^2) \\
\beta_{\text{GoodConditionSqm}} &\sim \text{N}(1000, 1500^2) \\
\beta_{\text{Age}} &\sim \text{N}(-2000, 2500^2) \\
\beta_{\text{TwoRoomsDummy}} &\sim \text{N}(5000, 10000^2) \\
\beta_{\text{ThreeRoomsDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{FourRoomsOrMoreDummy}} &\sim \text{N}(7500, 10000^2) \\
\beta_{\text{OwnFloor}} &\sim \text{N}(1000, 1000^2) \\
\beta_{\text{SaunaDummy}} &\sim \text{N}(5000, 2500^2) \\
\sigma &\sim \text{Half-Cauchy}(0, 15000) \\
\nu &\sim \text{Gamma}(2, 0.1)
\end{aligned} \tag{6.7}$$

The likelihood is $\text{Price}_i \sim t_\nu(\mu_i, \sigma)$ where the expected value $\mu_i$ is determined by the sum

$$\begin{aligned}
\mu_i = \; & \beta_{\text{Intercept},j[i]} + \\
& \beta_{\text{Sqm}}\ \text{Sqm}_i + \\
& \beta_{\text{GoodConditionSqm}}\ \text{GoodConditionSqm}_i + \\
& \beta_{\text{Age}}\ \text{Age}_i + \\
& \beta_{\text{TwoRoomsDummy}}\ \text{TwoRoomsDummy}_i + \\
& \beta_{\text{ThreeRoomsDummy}}\ \text{ThreeRoomsDummy}_i + \\
& \beta_{\text{FourRoomsOrMoreDummy}}\ \text{FourRoomsOrModeDummy}_i + \\
& \beta_{\text{OwnFloor}}\ \text{OwnFloor}_i + \\
& \beta_{\text{SaunaDummy}}\ \text{SaunaDummy}_i.
\end{aligned}$$

### 6.4.2 Estimates

Table 6.4 gives the descriptive statistics for the posterior distributions of the model parameters. Figure B.3 in appendix B describes the posterior distributions for the neighborhood-specific intercepts for model 4.

| Parameter | mean | sd | 2.5% q. | 50% q. | 97.5% q. | $\hat{n}_{\text{eff}}$ | $\hat{R}$ |
|---|---|---|---|---|---|---|---|
| $\beta_{\text{Sqm}}$ | 1882.29 | 72.93 | 1742.47 | 1881.13 | 2028.74 | 3035.68 | 1.00 |
| $\beta_{\text{GoodConditionSqm}}$ | 5.70 | 0.29 | 5.12 | 5.70 | 6.27 | 12564.26 | 1.00 |
| $\beta_{\text{Age}}$ | -1185.34 | 47.66 | -1280.07 | -1184.41 | -1092.50 | 6472.40 | 1.00 |
| $\beta_{\text{TwoRoomsDummy}}$ | 14494.17 | 2255.40 | 10053.33 | 14468.57 | 19017.86 | 6895.47 | 1.00 |
| $\beta_{\text{ThreeRoomsDummy}}$ | 17529.81 | 3250.47 | 11104.17 | 17515.61 | 24031.83 | 5182.00 | 1.00 |
| $\beta_{\text{FourRoomsOrMoreDummy}}$ | 23245.79 | 4839.45 | 13878.61 | 23198.25 | 32871.56 | 4653.42 | 1.00 |
| $\beta_{\text{OwnFloor}}$ | 2484.09 | 423.91 | 1669.79 | 2488.23 | 3310.76 | 9157.43 | 1.00 |
| $\beta_{\text{SaunaDummy}}$ | 11214.31 | 1572.23 | 8108.09 | 11223.76 | 14271.47 | 14528.64 | 1.00 |
| $\sigma$ | 33548.49 | 839.82 | 31901.58 | 33547.23 | 35227.89 | 8316.70 | 1.00 |
| $\nu$ | 1.97 | 0.09 | 1.81 | 1.97 | 2.15 | 9098.78 | 1.00 |
| $\eta^2$ | 4478.22 | 681.87 | 3315.68 | 4419.83 | 5956.42 | 1412.54 | 1.00 |
| $\rho^2$ | 0.27 | 0.03 | 0.21 | 0.26 | 0.34 | 1389.58 | 1.00 |

Table 6.4: Parameter estimates - model 4

Comparing the estimates in table 6.4 to estimates of the previous models, it can be seen that the effects of the interaction term $\beta_{\text{GoodConditionSqm}}$ is suspiciously decreased 100-fold. The estimates of the effects for the number of rooms have also noticeably shifted upwards compared to the previous models.

## 6.5 Model 5 - Varying intercepts and slopes model

Model 5 is a varying intercepts and slopes model where the intercepts along with the coefficients of the size variables Sqm and GoodConditionSqm are allowed to vary across the neighborhoods. The model is similar in spirit to models presented in Gelfand et al. (2003) with the difference that in the presented model, the coefficient vary over a discrete set of neighborhoods, whereas in the article, the coefficients are modeled with Gaussian process models which allow them to vary over a spatial surface. The model structure was chosen to mimic examples in sections 13.1-13.3 of McElreath (2016).

### 6.5.1 Model specification

For model 5, the coefficients $\beta_{\text{Intercept}}$, $\beta_{\text{Sqm}}$ and $\beta_{\text{GoodConditionSqm}}$ are determined through a group-level three dimensional multivariate normal distribution. The structure of this distribution is described next.

First, let $\mu^* = [\mu_{\text{Intercept}}, \mu_{\text{Sqm}}, \mu_{\text{GoodConditionSqm}}]'$ denote the expected value vector of the group-level model. The priors for the expected values $\mu_{\text{Intercept}}$, $\mu_{\text{Sqm}}$ and $\mu_{\text{GoodConditionSqm}}$ were chosen component-wise as

$$\begin{aligned} \mu_{\text{Intercept}} &\sim \text{N}(50000, 50000^2), \\ \mu_{\text{Sqm}} &\sim \text{N}(4000, 1000^2), \end{aligned}$$

and

$$\mu_{\text{GoodConditionSqm}} \sim \text{N}(1000, 1000^2).$$

The covariance matrix $\Sigma$ of the group-level model is parameterized using the covariance decomposition

$$\Sigma = \text{Diag}(\sigma^*)\ R\ \text{Diag}(\sigma^*)$$

where the term $\sigma^* = [\sigma_{\text{Intercept}}, \sigma_{\text{Sqm}}, \sigma_{\text{GoodConditionSqm}}]'$ is a vector for the standard deviations and the term $R$ is a correlation matrix. The priors for the standard deviation terms $\sigma_{\text{Intercept}}$, $\sigma_{\text{Sqm}}$ and $\sigma_{\text{GoodConditionSqm}}$ were chosen as

$$\begin{aligned} \sigma_{\text{Intercept}} &\sim \text{Half-Cauchy}(0, 7000), \\ \sigma_{\text{Sqm}} &\sim \text{Half-Cauchy}(0, 1500), \end{aligned}$$

and

$$\sigma_{\text{GoodConditionSqm}} \sim \text{Half-Cauchy}(0, 350).$$

The prior for the correlation matrix $R$ was chosen as

$$R \sim \text{LKJcorr}(2).$$

The coefficient vector $\beta^* = [\beta_{\text{Intercept}}, \beta_{\text{Sqm}}, \beta_{\text{GoodConditionSqm}}]'$ is then written with non-centered parameterization as

$$\beta^* = \mu^* + \text{Diag}(\sigma^*)Lz$$

where the term $L$ denotes the Cholesky decomposition of the correlation matrix $R$ (i.e. $R = LL'$) and $z$ is an offset term drawn from the standard three dimensional normal distribution, i.e.

$$z \sim \text{Multivariate-Normal}([0,0,0]', I_3).$$

List (6.8) gives the prior distribution choices for the rest of the parameters for model 5:

$$\begin{aligned}
\beta_{\text{Age}} &\sim \text{N}(-2000, 1500^2) \\
\beta_{\text{TwoRoomsDummy}} &\sim \text{N}(5000, 5000^2) \\
\beta_{\text{ThreeRoomsDummy}} &\sim \text{N}(7500, 5000^2) \\
\beta_{\text{FourRoomsOrMoreDummy}} &\sim \text{N}(7500, 5000^2) \\
\beta_{\text{OwnFloor}} &\sim \text{N}(1000, 1000^2) \\
\beta_{\text{SaunaDummy}} &\sim \text{N}(5000, 2500^2) \\
\sigma &\sim \text{Half-Cauchy}(0, 15000) \\
\nu &\sim \text{Gamma}(2, 0.1)
\end{aligned} \tag{6.8}$$

The likelihood is $\text{Price}_i \sim t_\nu(\mu_i, \sigma)$ where the expected value $\mu_i$ is determined by the sum

$$\begin{aligned}
\mu_i = {} & \beta_{\text{Intercept},j[i]} + \\
& \beta_{\text{Sqm},j[i]}\ \text{Sqm}_i + \\
& \beta_{\text{SqmGoodCondition},j[i]}\ \text{GoodConditionSqm}_i + \\
& \beta_{\text{Age}}\ \text{Age}_i + \\
& \beta_{\text{TwoRoomsDummy}}\ \text{TwoRoomsDummy}_i + \\
& \beta_{\text{ThreeRoomsDummy}}\ \text{ThreeRoomsDummy}_i + \\
& \beta_{\text{FourRoomsOrMoreDummy}}\ \text{FourRoomsOrModeDummy}_i + \\
& \beta_{\text{OwnFloor}}\ \text{OwnFloor}_i + \\
& \beta_{\text{SaunaDummy}}\ \text{SaunaDummy}_i.
\end{aligned}$$

### 6.5.2 Estimates

Table 6.5 gives the descriptive statistics for the posterior distributions of the neighborhood-invariant model parameters. Figures B.4, B.5 and B.6 in appendix B describe the posterior distributions for the neighborhood-specific intercepts and coefficients.

From table 6.5 it can be seen that the effects of having two rooms or three rooms have greatly diminished in model 5 compared with the previous models.

| Parameter | mean | sd | 2.5% q. | 50% q. | 97.5% q. | $\hat{n}_{\text{eff}}$ | $\hat{R}$ |
|---|---|---|---|---|---|---|---|
| $\beta_{\text{Age}}$ | -917.45 | 44.14 | -1005.98 | -917.33 | -832.79 | 10010.80 | 1.00 |
| $\beta_{\text{TwoRoomsDummy}}$ | -669.12 | 1630.51 | -3820.42 | -674.10 | 2478.98 | 10010.99 | 1.00 |
| $\beta_{\text{ThreeRoomsDummy}}$ | 3783.15 | 2277.29 | -847.33 | 3816.87 | 8225.26 | 10511.94 | 1.00 |
| $\beta_{\text{FourRoomsOrMoreDummy}}$ | 20887.59 | 3271.09 | 14479.89 | 20867.76 | 27407.43 | 11760.38 | 1.00 |
| $\beta_{\text{SaunaDummy}}$ | 14498.95 | 1505.08 | 11575.91 | 14509.64 | 17420.81 | 14590.59 | 1.00 |
| $\beta_{\text{OwnFloor}}$ | 1380.66 | 358.92 | 672.40 | 1381.02 | 2080.88 | 16706.39 | 1.00 |
| $\sigma$ | 27683.58 | 713.55 | 26312.06 | 27676.23 | 29088.38 | 7188.24 | 1.00 |
| $\nu$ | 2.27 | 0.11 | 2.05 | 2.26 | 2.50 | 11485.67 | 1.00 |
| $R_{1,2}$ | 0.34 | 0.11 | 0.11 | 0.34 | 0.54 | 1221.70 | 1.00 |
| $R_{1,3}$ | -0.15 | 0.17 | -0.48 | -0.15 | 0.20 | 4551.53 | 1.00 |
| $R_{2,3}$ | 0.23 | 0.15 | -0.06 | 0.23 | 0.51 | 7092.46 | 1.00 |
| $\sigma_{\text{Intercept}}$ | 32720.56 | 3544.95 | 26228.32 | 32578.52 | 40168.12 | 3557.89 | 1.00 |
| $\sigma_{\text{Sqm}}$ | 1323.92 | 91.21 | 1157.94 | 1318.70 | 1515.82 | 3645.68 | 1.00 |
| $\sigma_{\text{GoodConditionSqm}}$ | 229.97 | 38.69 | 156.17 | 229.01 | 308.10 | 3247.46 | 1.00 |
| $\mu_{\text{Intercept}}$ | 110668.17 | 4586.20 | 101638.97 | 110607.81 | 119864.77 | 4628.50 | 1.00 |
| $\mu_{\text{Sqm}}$ | 2112.79 | 128.01 | 1866.58 | 2113.57 | 2358.70 | 2782.98 | 1.00 |
| $\mu_{\text{GoodConditionSqm}}$ | 564.67 | 34.41 | 496.73 | 564.73 | 631.31 | 7966.70 | 1.00 |

Table 6.5: Parameter estimates - model 5

The effect of having four or more rooms has remained in the same order of magnitude as in the previous models. It would seem plausible that the 'four or more rooms?'-dummy could now encode the effects of the building type rather than the number of rooms as row houses or town houses typically have more than three rooms. The estimates also indicate that some correlation seems to exist between the intercepts and effects of the apartment size.

# Chapter 7

# Model comparison and model stacking

This chapter presents the model checks and the comparative predictive performance of the individual models. The model stacking results are also presented. Appendix C includes the model comparison figures for the individual models and appendix D the corresponding figures for the stacking model.

## 7.1 Model comparison

### 7.1.1 Model checks

Using the whole estimation set, it can be seen that the replicated mean prices are generally too small and the replicated median prices too high with models 4 and 5 standing out. For model 4, both the replicated mean and median prices are smaller than the observed mean and median prices. For model 5, the replicated mean and median prices are generally better than those of the other models even though model 5 still underpredicts the mean price and overpredicts the median price. See figures C.1 and C.2 for the distributions.

For Helsinki, the replicated neighborhood mean prices tend to be unsurprisingly wrong for model 1 where the neighborhood information is not utilized. When the neighborhood information is utilized, as in models 2 and 3, the replicated mean prices are roughly correct for majority of the neighborhoods. The Gaussian process approach of model 4 produces worse replicated mean prices than those of models 2 and 3. There are also neighborhoods, e.g. Kaartinkaupunki, Eira and Kruunuhaka, where the replicated mean prices are wrong even when the intercepts are allowed to vary across the neighborhoods. With the varying intercepts and slopes approach of model 5, the replicated mean

prices for these neighborhoods tend to be correct. There are also neighborhoods, e.g. Katajanokka and Länsisatama, where model 5 is also insufficient. See figure C.3 for the distributions.

For Espoo and Vantaa, the replicated mean prices for neighborhoods behave similarly to the majority of Helsinki's neighborhoods. When the neighborhood information is not utilized, the replicated mean prices tend to be wrong. Once the neighborhood information is included into the intercept, as in models 2 and 3, the replicated mean prices are largely correct. In Vantaa there does not seem to exist neighborhoods where the replicated mean prices would benefit significantly from model 5's varying slopes and intercepts. This also holds largely for Espoo with the arguable exception of the Kaupunginkallio neighborhood where the varying slopes of model 5 are needed. Again, the Gaussian process approach of model 4 produces worse replicated mean prices as models 2, 3 and 5 for both Espoo and Vantaa. See figures C.4 and C.5 for the distributions.

### 7.1.2 Predictive performance

Predictive performance is analyzed in terms of the calibration and sharpness of the predictive distributions and in terms of the PSIS-LOO scores and $R^2$-statistics of the point predictions.

None of the models can be said to be well-calibrated as the empirical PIT histograms differ markedly from the uniform distribution for all models. For model 4, the discrepancy is especially clear as there is much mass at high PIT values indicating that the model tends to underpredict. In terms of sharpness of the predictive distributions, the models fall into three categories: First, model 1 produces the most widely-dispersed predictive distributions, models 2, 3, and 4 produce similarly sharp distributions and model 5 the sharpest distributions compared to the other models. The average width of the central 90 % credible interval is approximately 330 000 for model 1, 200 000 for models 2-4 and 150 000 for model 5. See figure C.6 for the PIT histograms and figure C.7 for the histograms of the central 90 % credible interval widths.

Table 7.1 gives the model-specific estimates for the expected log pointwise predictive density values $\widehat{\text{elpd}}_{\text{psis-loo}}$ and the effective number of parameters $\text{p}_{\text{loo}}$. For model 5, there were 9 observations with $\hat{k}$-parameter estimates greater than 0.7. To account for these observations, the values presented in table 7.1 for model 5 were calculated using the PSIS-LOO+-approach of Vehtari et al. (2017) where exact LOO calculations are conducted for the problematic observations.

Larger values of the $\widehat{\text{elpd}}_{\text{psis-loo}}$-statistic indicate better predictive performance, so model 5 has the best value in table 7.1. The complexity of model 5, as measured by the $\text{p}_{\text{loo}}$-statistic, is also markedly larger than the other models.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| $\widehat{\text{elpd}}_{\text{psis-loo}}$ | -48249.87 | -46294.43 | -46285.67 | -46334.23 | -45451.76 |
| $\text{p}_{\text{loo}}$ | 11.23 | 162.92 | 146.88 | 134.30 | 333.11 |

Table 7.1: Estimates for the expected log pointwise predictive density values and effective number of parameters

The table also suggests that the distance information is useful for determining the neighborhood specific intercepts. Comparing models 2 and 3, it can be seen that when distance information is utilized, the predictive performance improves and, surprisingly, the effective of number of parameters decreases. A similar decrease in the effective number of parameters can be seen between models 2 and 4, although the predictive performance of model 4 is slightly worse than that of model 2.

For observations in the estimation set, the point predictions averaged from the predictive distributions tend to be too small. For example, none of the point predictions of models 1-4 exceeds 900 000 euros whereas the highest observed price is approximately 2 500 000 euros. The point predictions from model 5 perform better but visual inspection still suggests underprediction also for model 5. The classical $R^2$-statistic for the estimation set point predictions is approximately 0.42 for model 1, 0.71 for models 2 and 3, 0.63 for model 4 and 0.85 for model 5. See figure C.8 for the scatter plots for the estimation set.

Point predictions in the test set behave in a similar fashion as in the estimation set. Models 1-4 again never produce point predictions that exceed 900 000 euros while the highest observed price in the test set is exactly 2 500 000 euros. Model 5 also underpredicts but the underpredictions are smaller than in the other models. The classical $R^2$-measure for the test set point predictions is approximately 0.42 for model 1, 0.70 for models 2 and 3, 0.61 for model 4 and 0.87 for model 5. See figure C.9 for the test set scatter plots.

In summary, the model checks indicate problems with the current model specifications with model 5 standing out as the least problematic model. Despite the problems, the models produce fairly reasonable predictions for a large proportion of the observations. It is clear that the neighborhood information is fundamentally important for the price predictions.

## 7.2 Model stacking

The stacking model built with models 1-5 and the associated results are presented next. Table 7.2 gives the weights $\hat{w}$ solved from problem (3.15). The weights were calculated after applying the PSIS-LOO+-adjustment to model 5.

Surprisingly model 4 contributes a non-negligible weight to the stacking model despite the model's problems discussed in the previous section.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| $\hat{w}$ | 0.000 | 0.000 | 0.008 | 0.072 | 0.920 |

Table 7.2: Stacking weights

The same model checks were applied to the stacking model as to the individual models. For the replicated mean prices, the stacking model distribution is slightly sharper than that of model 5 and a bit further away from the observed mean price. For the replicated median prices, the stacking model distribution is better than that of any other single model, although the distribution's mass is still not centered on the observed median price. See figures D.1 and D.2 for the distributions.

For the neighborhood-specific mean prices, the distributions of the stacking model are roughly equal to those from model 5, as would be expected from the stacking weights. The influence of the weights of models 3 and 4 shifts the mean price distributions slightly downwards. See figures D.3, D.4 and D.5 for the distributions.

As with the component models, the stacking model is not well-calibrated in the sense that its empirical PIT histogram is dissimilar to a uniform distribution. In terms of sharpness, the stacking model produces somewhat wider predictive distributions than model 5. The average width of the central 90 % credible intervals for the predictive distributions from the stacking model is approximately 168 000. See figure D.6 for the PIT histogram and figure D.7 for the sharpness histogram.

The point predictions averaged from the stacking model's predictive distributions are roughly equal to those from model 5's predictive distributions. The classical $R^2$-statistic for the point predictions is approximately 0.85 for the estimation set and 0.86 for the test set. See figures D.8 and D.9 for the estimation and test set scatter plots.

It was expected ex ante that the stacking model would clearly outperform any single model in terms of prediction but, surprisingly, the stacking model does not outperform model 5. The model check results and the predictive performance of the stacking model in the estimation and test sets are very similar to those of model 5.

# Chapter 8

# Discussion

This chapter discusses the results and possible extensions of the presented models.

## 8.1 Results

### 8.1.1 Overall results

Overall, the results of the models seem reasonable. The coefficients behave in a plausible manner for each explanatory variable. Given the general impressions of the housing markets in the Helsinki-Espoo-Vantaa region, the neighborhood-specific estimates are seemingly acceptable as, for example, neighborhoods 'with expensive reputations' in the center of Helsinki get larger estimates for the intercepts. The estimates also represent the expected uncertainty coherently: When there is only a limited number of observations for a given neighborhood, the posterior distributions for the neighborhood-specific estimands tend to be more widely dispersed representing the larger degree of uncertainty related to the estimand.

Model checks indicate that the models fail to capture some aspects of the underlying phenomenon. For example, there were neighborhoods in Helsinki where the observed mean prices were not located in high mass regions of the replicated mean price distributions of any of the models. Additionally, none of the models was well-calibrated in terms of uniform probability inverse transforms. Finally, the point predictions, especially for models 1-4, tend to be too low for expensive apartments.

### 8.1.2 Poor performance of model 4

Checks on the posterior of model 4 suggest that model is misspecified. The estimate for the coefficient of interaction term is suspiciously small and the empirical PIT histogram of the model indicates that the model's calibration is especially bad. In hindsight, the covariance structure (6.6) could be poorly chosen. Banerjee et al. (2015, p. 28) suggests including an additional variance term for the neighborhood-specific intercept, i.e. choosing $K_{ij}$ such that

$$K_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \tau^2 \ \delta_{ij}$$

where the new $\tau^2$-parameter allows the model more leeway for setting the neighborhood-specific intercepts. Other covariance functions listed in Banerjee et al. (2015, p. 28) could also be tried. Additionally, the choice of constant value of 70 over all neighborhoods for the expected value does not reflect the prior knowledge of the relative desirability of the neighborhoods. All in all, more work is needed to successfully utilize Gaussian process models.

### 8.1.3 Lack of predictive improvements from model stacking

The results indicate that model stacking does not lead to clear improvements in predictive performance. Barring the possibility of programming errors, the lack of improvement from the stacking model is speculated to be due to the limited set of models. For housing price modeling, it would seem natural to adopt the M-open view (see e.g. Bernardo and Smith (1994, p. 385) or Yao et al. (2018)) where it is accepted that the true data generating process will not be among the specified models and that including it is not feasible due to the complexity of the phenomenon. For example, possible modeling choices include which explanatory variables should be included, whether to model for possible nonlinearities in the relationship between the response and some explanatory variables, and how to model for the spatial effects. A comprehensive model set, covering all reasonable combinations of the modeling choices, would seem necessary to be able to approximate the true data generating process. The models presented in this thesis are clearly just the first step towards the comprehensive model set.

### 8.1.4 Results in terms of existing literature

Following the definition of spatial heterogeneity in Bowen et al. (2001) as systematic differences in prices in terms of location, suppose that submarkets are defined as geographical areas where the behavior of some of the explanatory variables differ markedly compared to some other well-defined geographical

area. Then estimates for model 5, seen e.g. in figure B.5, can be interpreted as evidence that there exists a separate submarket around central southern Helsinki where the relationship between the price and the size of the apartment is markedly different from other regions.

## 8.2 Further model development

As discussed in section 8.1.3, there are many possibilities for model development in terms of the choice of explanatory variables, the model structures and how the spatial information is used.

### 8.2.1 Variables

There are explanatory variables that can be constructed from the current data but which have not been utilized in the models. For example, dummy variables could be constructed to describe the building type (table A.5, also discussed in section 6.5), the existence of an elevator (table A.7) or the energy classifications (table A.9). The age of the building could also be used to generate variables which describe the possible need for expensive near-future renovations. For example, suppose that the lifetime of a typical plumbing system is approximately 50 years. Then, an interesting 'time to plumbing renovations' variable for apartment $i$ could be constructed as

$$\text{TimeToPlumbingRenov}_i = 50 - (\text{Age}_i \bmod 50)$$

so that the models would be able to represent possible cyclical effects on price due to necessary renovations. Such variables could be constructed also for other subsystems of a building, e.g. the facade and heating system.

Other variables of interest include at least information on whether the housing company owns the lot, the size of the lot and whether the apartment has a garage. Similarly, additional group-level variables could be also compiled to measure, for example, the average educational level of the neighborhood's inhabitants. These variables were not unfortunately available in the current data.

### 8.2.2 Model structure

The model structure could be further explored with respect to the utilized groupings and how the continuous variables are handled.

In addition to the neighborhood grouping, other reasonable groupings include the city where the observations are located and the building type. The city grouping could now reflect, for example, whether the differences in local

tax rates are actually included in the housing prices. Likewise, it might be reasonable to model the effects of size based on the building type, so the building type could be utilized as a grouping. Lastly, if further data sets were gathered at different points of time, a temporal grouping could be included to model possible time-variant trends in the housing market.

The models in this thesis assume that the relationships between the price and the continuous explanatory variables are linear. This assumption could be relaxed by using, for example, splines to allow the models to find possible non-linear relationships. For instance, it would seem natural that the price increase would be steeper ceteris paribus for a size increase from 21 $m^2$ to 22 $m^2$ compared to an increase from 121 $m^2$ to 122 $m^2$, so the size variable seems a natural candidate for applying a spline transformation. Other possibilities include the addition of quadratic terms or the use of logarithmic transformations.

Lastly, additional variables could be included to the varying slopes as in model 5. For example, the effect of age could be allowed to vary across the neighborhoods as the housing stock within a neighborhood is usually built roughly in the same time period.

### 8.2.3 Spatial aspects

The available spatial information might be insufficient for capturing the underlying spatial effects since the official neighborhoods can include disparate regions within a given neighborhood. For example, in Helsinki's Vuosaari neighborhood there exists the Aurinkolahti subregion which seemingly differs noticeably from the rest of Vuosaari: Observed prices in Aurinkolahti tend to larger compared to the rest of the Vuosaari, so the models tend to underpredict the prices for apartments in Aurinkolahti. Obviously, information on the exact locations of the apartments would allow the proper use of point-reference models presented, e.g., in Banerjee et al. (2015) or Gelfand et al. (2003) which should help with these types of problems. Moreover, exact locations would be necessary to study the existence of spatial dependency discussed in Bowen et al. (2001).

## 8.3 Conclusions

Despite the problems with the models, it seems reasonable to conclude that the results serve as further evidence of the importance of location for housing price modeling. Despite fairly stripped-down data, the predictions from the models are broadly reasonable. The results also indicate that the price mechanism differs by location in the Helsinki-Espoo-Vantaa region.

Bayesian hierarchical models are clearly a powerful tool for the modeling of

housing prices. With these types of models, it is fairly straightforward to utilize geographical information either as a grouping or explicitly through Gaussian process models. Likewise, it is simple to estimate robust models by choosing a robust likelihood distribution. When the focus is simply on predicting well, model stacking can be used as a straightforward method for aggregating predictions from multiple models.

The models also have apparent practical uses as they indicate which neighborhoods have the highest premiums for apartments in good condition. Consider an investment strategy of buying apartments in poor condition, renovating them and then selling them. This investment strategy could then utilize the models in the choice of which apartments to buy: For example, given two otherwise identical apartments, one being located in Koivukylä and the other in Martinlaakso, model 5 suggests[1] that the former should probably produce better returns.

[1] See figure B.6.

# Appendix A

# Descriptive statistics for data

Tables in this appendix dealing with the housing sales have been compiled without the recorded sales from Kalajärvi.

Table A.1 describes how the data is distributed over the different neighborhoods after the neighborhood names have been formatted. Derived from the free text field describing the type of the apartment, table A.2 describes the distribution of the number of rooms and table A.3 the distribution of the dummy variable describing whether the apartment has a sauna. Table A.4 describes the distribution of the floor variable. Table A.5 describes the distribution of the building types. Table A.6 describes the year when the buildings were built. Table A.7 describes the distribution of the dummy variable describing whether the building has an elevator. Table A.8 describes the distribution of the reported conditions of the apartments. Table A.9 describes the distribution of the energy classifications of the apartments. For energy classfications, the year information has been dropped in the table, e.g. the raw energy classifcation '$G_{2013}$' has been recorded as class G when compiling the table. Table A.10 gives the descriptive statistics for the continuous price and size variables. Table A.11 gives the descriptive statistics for the distance measures.

| Name | Frequency | Name | Frequency | Name | Frequency |
|---|---|---|---|---|---|
| Alppiharju | 82 | Koskela | 13 | Pasila | 57 |
| Askisto | 2 | Kruununhaka | 59 | Perusmäki | 10 |
| Asola | 88 | Kulosaari | 16 | Piispankylä | 7 |
| Eira | 9 | Kumpula | 11 | Pitäjänmäki | 47 |
| Espoon keskus | 80 | Kuninkaala | 2 | Pohjois-Tapiola | 15 |
| Espoonlahti | 58 | Kuninkaanmäki | 3 | Pukinmäki | 34 |
| Etu-Töölö | 116 | Kuurinniitty | 4 | Punavuori | 87 |
| Haaga | 207 | Laajalahti | 11 | Rajakylä | 17 |
| Hakunila | 26 | Laajasalo | 64 | Rekola | 9 |
| Hämeenkylä | 37 | Laakso | 2 | Ruskeasanta | 14 |
| Hämevaara | 2 | Laaksolahti | 49 | Ruskeasuo | 21 |
| Haukilahti | 12 | Lahnus | 1 | Saunalahti | 14 |
| Havukoski | 40 | Länsimäki | 12 | Sepänkylä | 15 |
| Henttaa | 21 | Länsisatama | 88 | Simonkylä | 30 |
| Hermanni | 32 | Latokaski | 5 | Sörnäinen | 55 |
| Herttoniemi | 149 | Lauttasaari | 163 | Soukka | 31 |
| Hiekkaharju | 60 | Leppäkorpi | 6 | Suurmetsä | 81 |
| Ilola | 12 | Leppävaara | 90 | Suutarila | 46 |
| Itä-Hakkila | 9 | Lintuvaara | 32 | Taka-Töölö | 85 |
| Järvenperä | 12 | Lippajärvi | 21 | Tammisalo | 10 |
| Jokiniemi | 21 | Malmi | 117 | Tammisto | 16 |
| Kaarela | 136 | Mankkaa | 8 | Tapaninkylä | 87 |
| Kaartinkaupunki | 8 | Martinlaakso | 94 | Tapiola | 93 |
| Kaitaa | 18 | Matari | 10 | Tikkurila | 73 |
| Kaivoksela | 28 | Matinkylä | 88 | Toukola | 20 |
| Kallio | 194 | Meilahti | 46 | Tuomarinkylä | 41 |
| Kamppi | 84 | Mellunkylä | 181 | Ullanlinna | 72 |
| Käpylä | 42 | Metsola | 10 | Vaarala | 13 |
| Karakallio | 20 | Mikkola | 11 | Vallila | 65 |
| Karhusuo | 6 | Munkkiniemi | 156 | Vanhakaupunki | 8 |
| Karvasmäki | 2 | Muurala | 5 | Vantaanlaakso | 4 |
| Katajanokka | 63 | Myyrmäki | 91 | Vanttila | 5 |
| Kauklahti | 26 | Niipperi | 6 | Vapaala | 13 |
| Kaupunginkallio | 3 | Niittykumpu | 30 | Varisto | 3 |
| Keimola | 3 | Nikinmäki | 5 | Vartiokylä | 143 |
| Kilo | 20 | Nöykkiö | 15 | Viertola | 25 |
| Kivistö | 91 | Nupuri | 1 | Vierumäki | 1 |
| Kluuvi | 4 | Olari | 83 | Viherlaakso | 23 |
| Koivuhaka | 5 | Otaniemi | 1 | Viikki | 43 |
| Koivukylä | 24 | Oulunkylä | 120 | Vuosaari | 159 |
| Kolmperä | 1 | Päiväkumpu | 6 | Westend | 2 |
| Konala | 34 | Pakila | 39 | Ylästö | 20 |
| Korso | 31 | Pakkala | 27 | | |

Table A.1: Frequencies for the neighborhoods

| Number of rooms | Frequency |
|---|---|
| 1 | 885 |
| 2 | 1880 |
| 3 | 1379 |
| 4 | 795 |
| 5 | 309 |
| 6 | 68 |
| 7 | 15 |
| 8 | 3 |

Table A.2: Frequencies for the number of rooms

| Sauna existence | Frequency |
|---|---|
| does not have a sauna | 3738 |
| has a sauna | 1596 |

Table A.3: Frequencies for the existence of a sauna in the apartment

| Floor | Frequency |
|---|---|
| 1 | 1447 |
| 2 | 1342 |
| 3 | 1018 |
| 4 | 650 |
| 5 | 417 |
| 6 | 254 |
| 7 | 128 |
| 8 | 53 |
| 9 | 11 |
| 10 | 3 |
| 11 | 2 |
| 12 | 4 |
| 13 | 2 |
| 15 | 1 |
| 16 | 1 |
| 17 | 1 |

Table A.4: Own floor, frequencies

| Building type | Frequency |
|---|---|
| high-rise building apartment ('kt') | 4334 |
| town house apartment ('ok') | 187 |
| row house apartment ('rt') | 813 |

Table A.5: Frequencies for the building types

| Year | Frequency | Year | Frequency | Year | Frequency |
|---|---|---|---|---|---|
| 1850 | 1 | 1933 | 6 | 1978 | 56 |
| 1874 | 1 | 1934 | 8 | 1979 | 58 |
| 1883 | 1 | 1935 | 6 | 1980 | 72 |
| 1885 | 1 | 1936 | 36 | 1981 | 74 |
| 1886 | 2 | 1937 | 41 | 1982 | 61 |
| 1888 | 1 | 1938 | 67 | 1983 | 61 |
| 1889 | 1 | 1939 | 50 | 1984 | 78 |
| 1890 | 2 | 1940 | 29 | 1985 | 102 |
| 1891 | 4 | 1941 | 15 | 1986 | 66 |
| 1892 | 2 | 1942 | 1 | 1987 | 65 |
| 1895 | 1 | 1943 | 1 | 1988 | 57 |
| 1896 | 3 | 1944 | 4 | 1989 | 79 |
| 1897 | 1 | 1945 | 12 | 1990 | 66 |
| 1898 | 1 | 1946 | 16 | 1991 | 29 |
| 1899 | 1 | 1947 | 6 | 1992 | 20 |
| 1900 | 3 | 1948 | 12 | 1993 | 21 |
| 1902 | 10 | 1949 | 5 | 1994 | 20 |
| 1903 | 6 | 1950 | 25 | 1995 | 25 |
| 1904 | 11 | 1951 | 21 | 1996 | 19 |
| 1905 | 4 | 1952 | 37 | 1997 | 14 |
| 1906 | 23 | 1953 | 19 | 1998 | 17 |
| 1907 | 12 | 1954 | 59 | 1999 | 19 |
| 1908 | 15 | 1955 | 40 | 2000 | 46 |
| 1909 | 7 | 1956 | 62 | 2001 | 50 |
| 1910 | 6 | 1957 | 78 | 2002 | 25 |
| 1911 | 16 | 1958 | 54 | 2003 | 40 |
| 1912 | 30 | 1959 | 69 | 2004 | 52 |
| 1913 | 19 | 1960 | 63 | 2005 | 69 |
| 1914 | 9 | 1961 | 107 | 2006 | 42 |
| 1915 | 5 | 1962 | 119 | 2007 | 66 |
| 1918 | 1 | 1963 | 101 | 2008 | 50 |
| 1919 | 1 | 1964 | 98 | 2009 | 27 |
| 1920 | 6 | 1965 | 95 | 2010 | 20 |
| 1921 | 1 | 1966 | 55 | 2011 | 38 |
| 1922 | 4 | 1967 | 61 | 2012 | 53 |
| 1923 | 9 | 1968 | 59 | 2013 | 59 |
| 1924 | 17 | 1969 | 53 | 2014 | 82 |
| 1925 | 22 | 1970 | 76 | 2015 | 78 |
| 1926 | 35 | 1971 | 81 | 2016 | 37 |
| 1927 | 46 | 1972 | 109 | 2017 | 105 |
| 1928 | 86 | 1973 | 99 | 2018 | 226 |
| 1929 | 39 | 1974 | 114 | 2019 | 335 |
| 1930 | 12 | 1975 | 81 | 2020 | 5 |
| 1931 | 7 | 1976 | 51 | | |
| 1932 | 12 | 1977 | 52 | | |

Table A.6: Frequencies for the years when the building was built

| Elevator existence | Frequency |
|---|---|
| no elevator exists | 2640 |
| elevator exists | 2694 |

Table A.7: Frequencies for the existence of an elevator in the building

| Condition | Frequency |
|---|---|
| not recorded | 42 |
| bad | 149 |
| good | 3477 |
| adequate | 1666 |

Table A.8: Reported conditions, frequencies

| Class | Frequencies |
|---|---|
| not recorded | 1307 |
| A | 63 |
| B | 166 |
| C | 698 |
| D | 864 |
| E | 1048 |
| F | 867 |
| G | 321 |

Table A.9: Energy classifications, classes, frequencies

| Variable | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Square meters | 12.50 | 45.00 | 60.50 | 67.28 | 82.00 | 320.00 |
| Price | 865 | 175000 | 240000 | 280734 | 330000 | 2500000 |
| Price per square meters | 69 | 3014 | 4174 | 4465 | 5636 | 14214 |

Table A.10: Continuous variables, descriptive statistics

| Distance measure | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Direct distance to ocean | 2.00 | 867.81 | 4868.30 | 5496.62 | 8833.72 | 18704.17 |
| Road distance to center of Helsinki | 652.11 | 10133.17 | 15478.40 | 14890.31 | 20205.15 | 29525.70 |
| Inter-centroid distance ($D_{ii}$ excluded) | 0.41 | 9.00 | 14.01 | 14.56 | 19.42 | 38.03 |

Table A.11: Distances, descriptive statistics

# Appendix B

# Posterior distributions for the neighborhood-specific terms for models 2, 3, 4 and 5

This appendix includes the box plots figures of the posterior distributions for the neighborhood-specific terms for models 2-5.

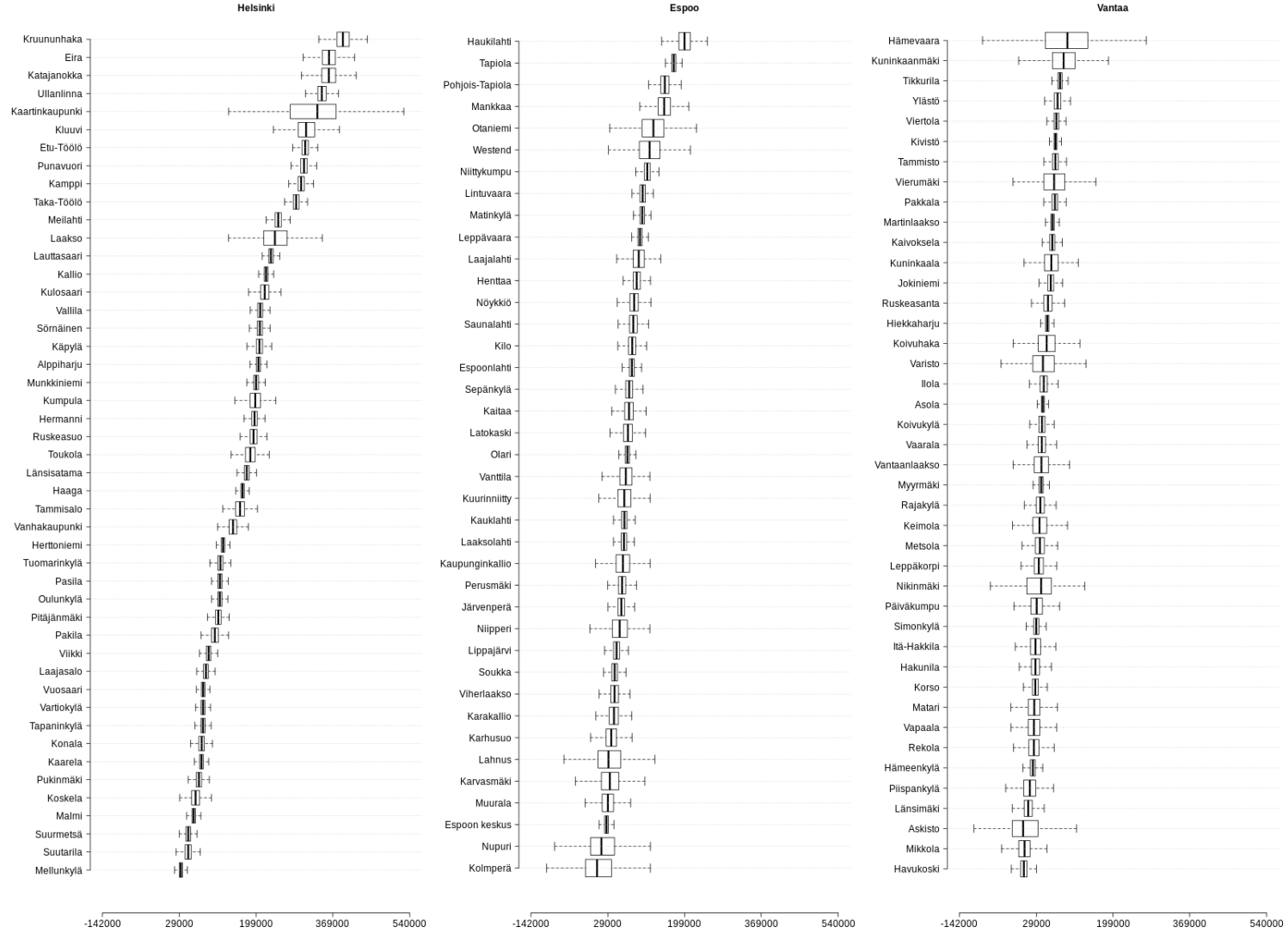


Figure B.1: Posterior box plots for the intercepts, model 2

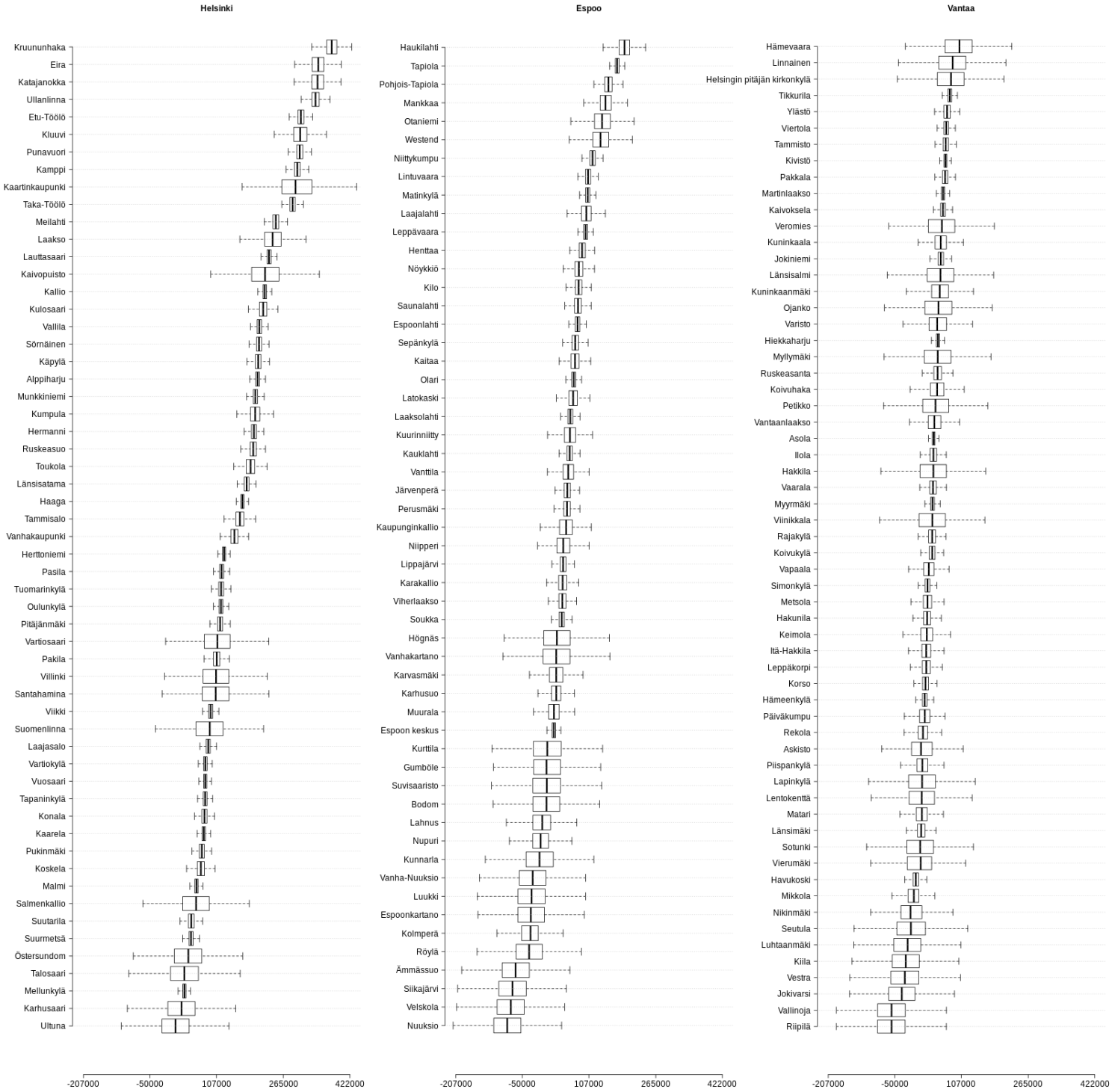


Figure B.2: Posterior box plots for the intercepts, model 3

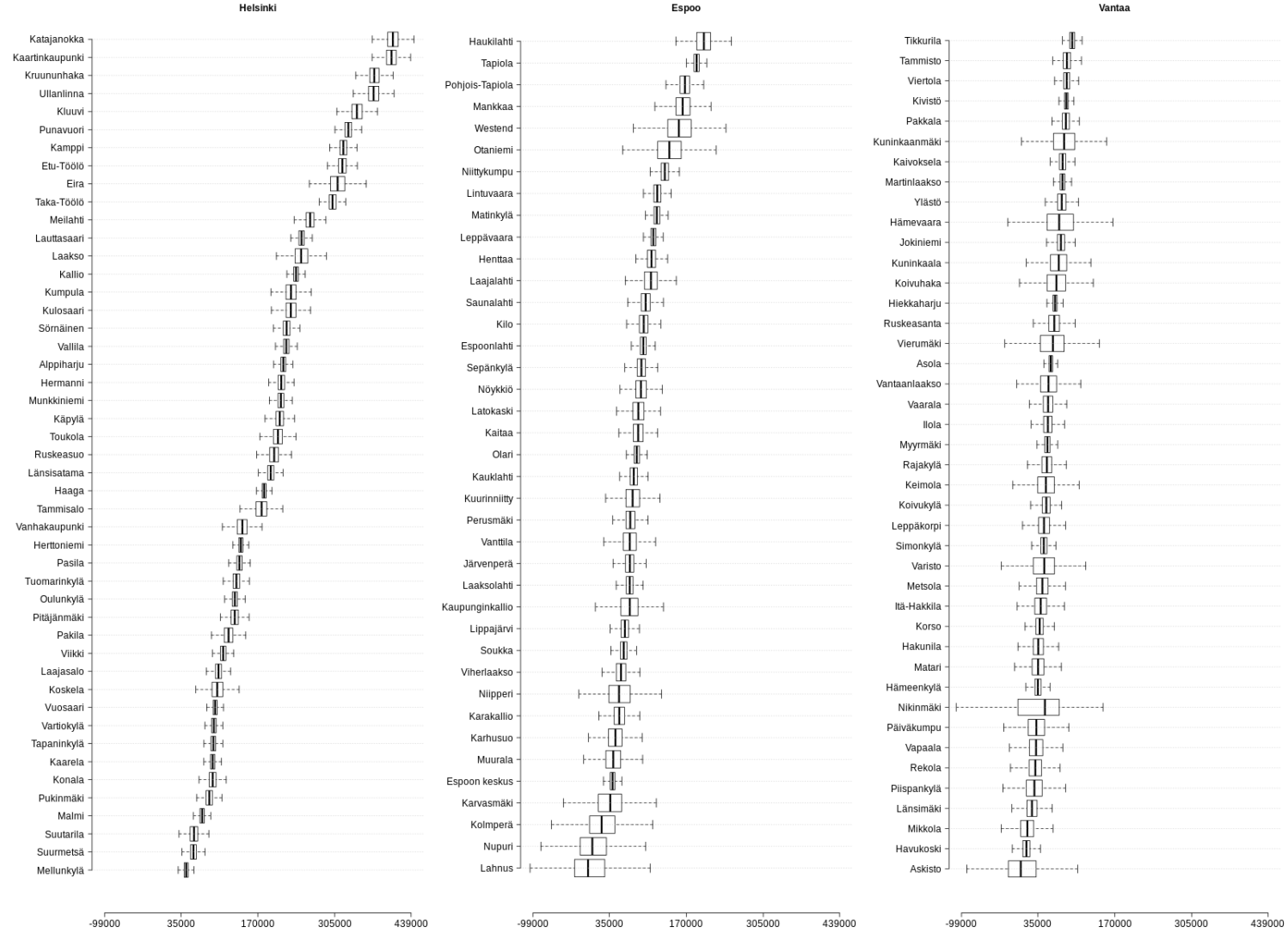


Figure B.3: Posterior box plots for the intercepts, model 4

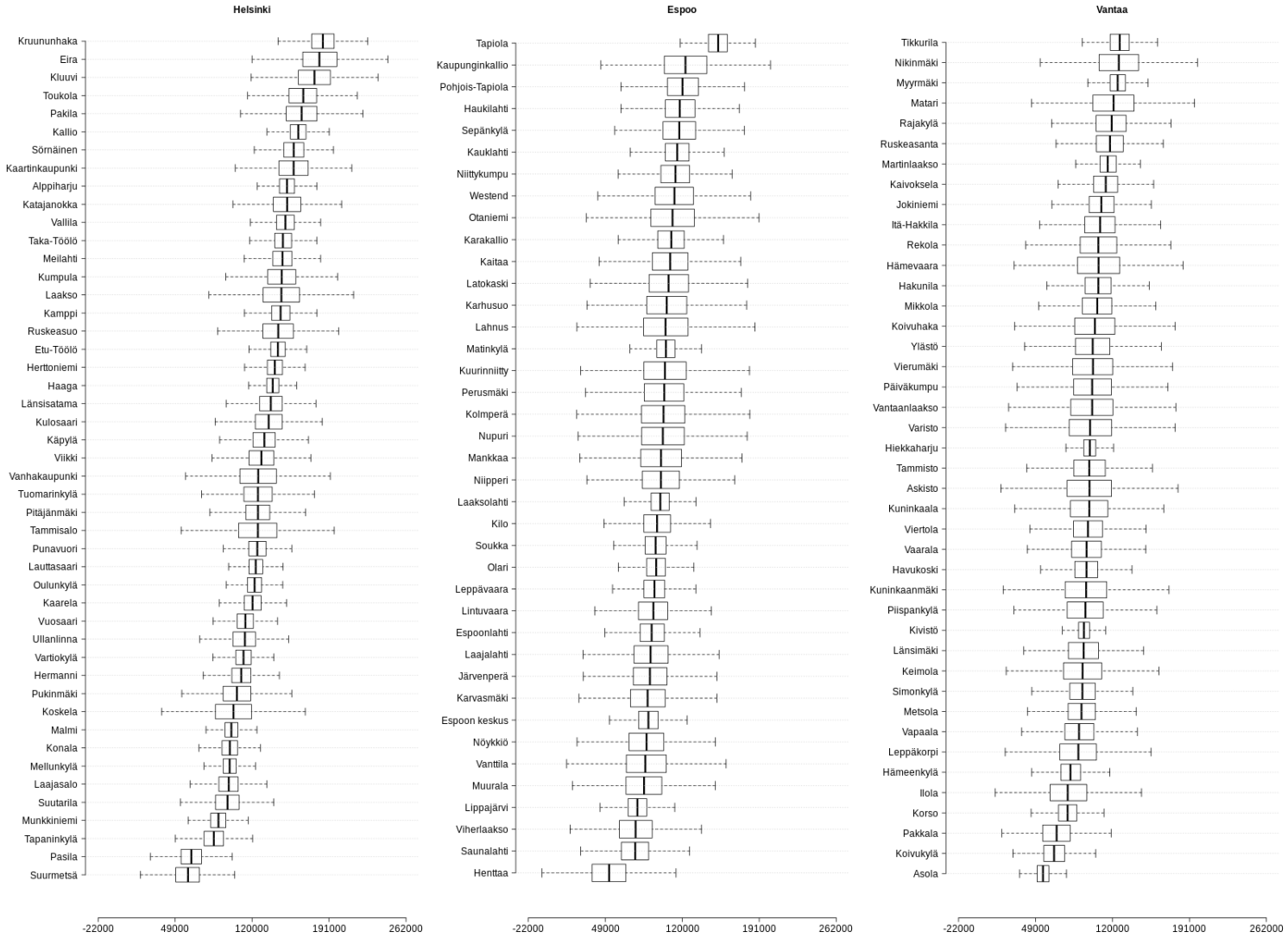


Figure B.4: Posterior box plots for the intercepts, model 5

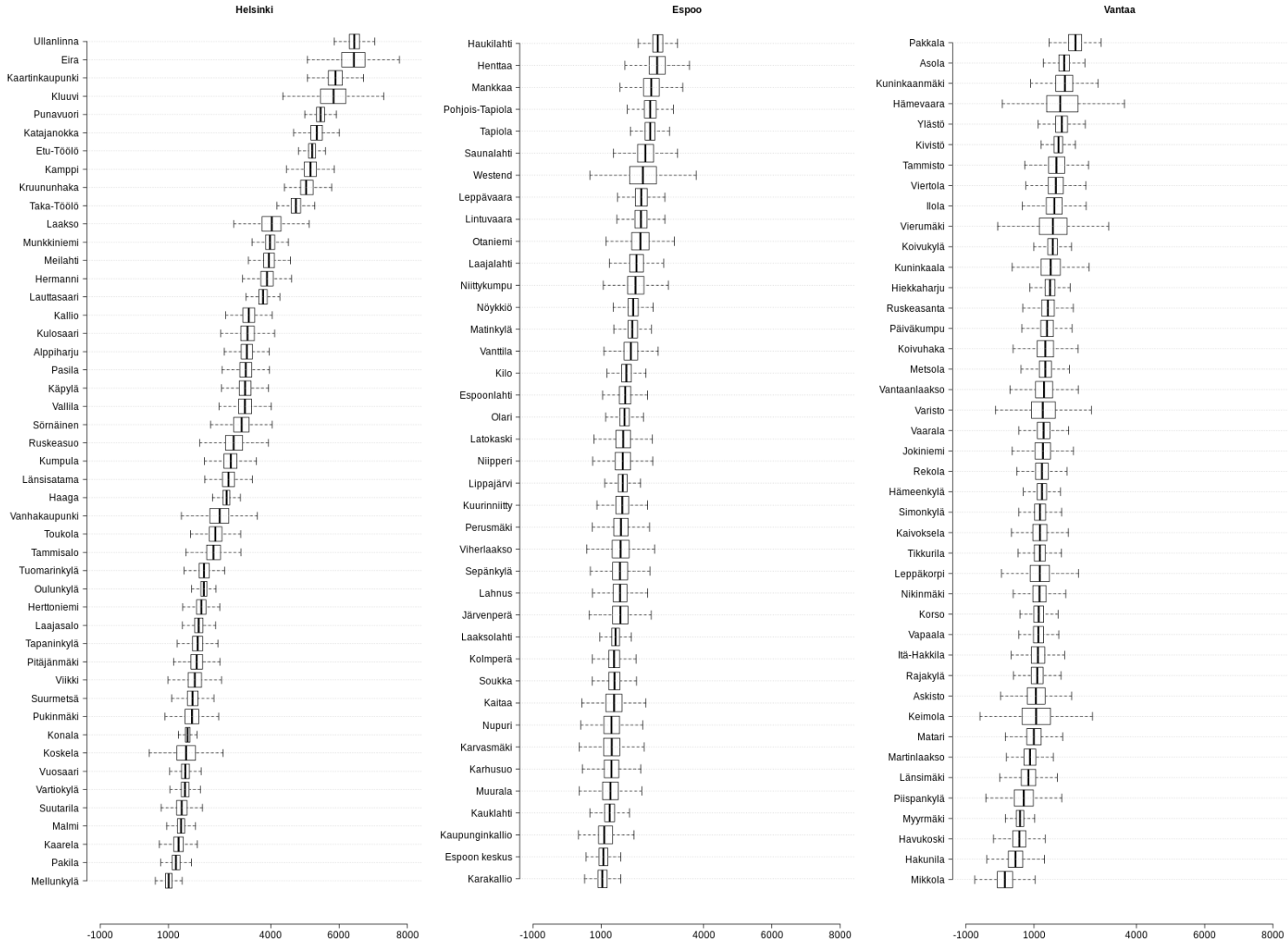


Figure B.5: Posterior box plots for the size coefficient, model 5

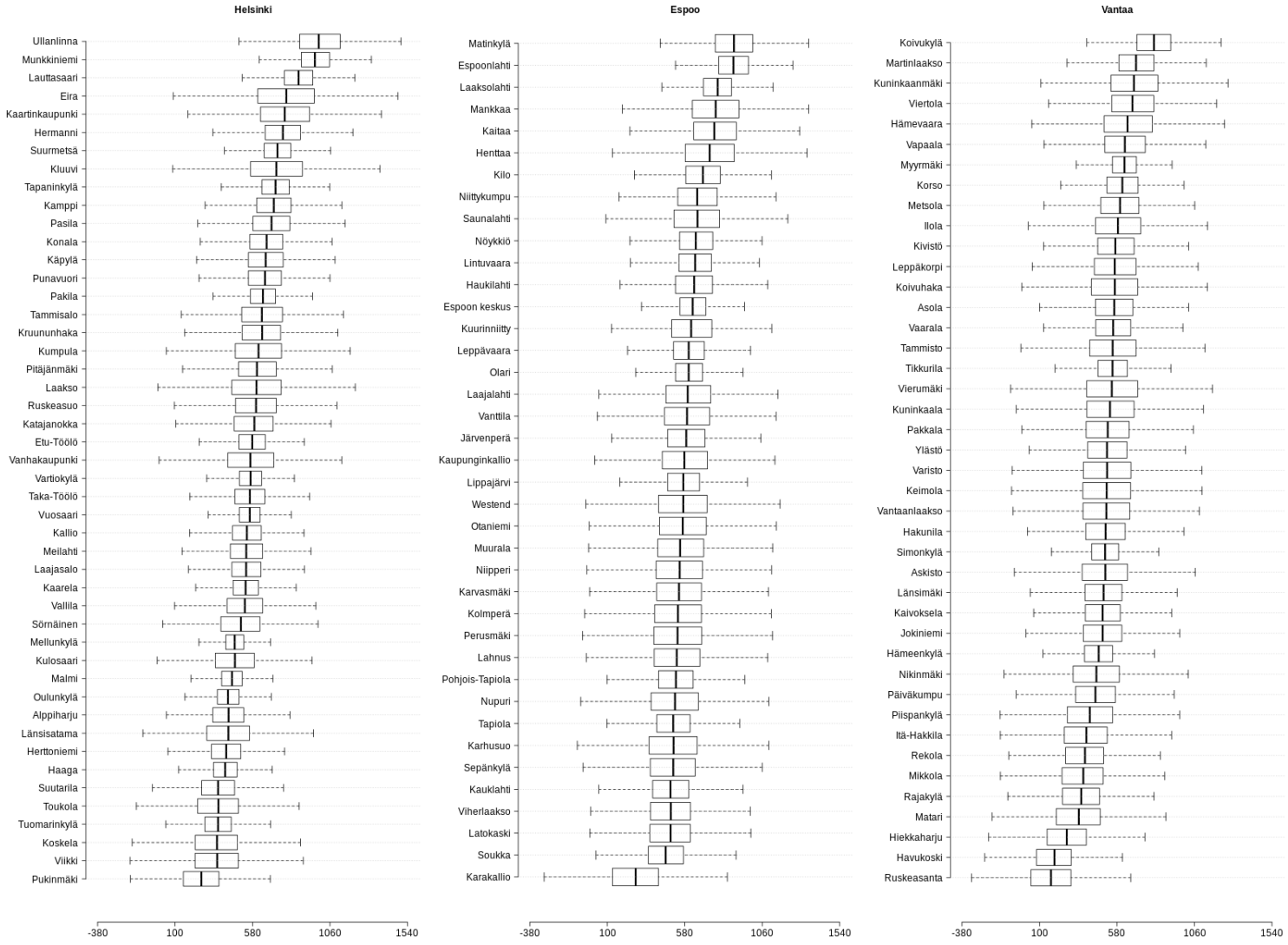


Figure B.6: Posterior box plots for the interaction term, model 5

# Appendix C

# Model comparison figures

Figures C.1 and C.2 include the distributions for the replicated mean and median prices under each model from the estimation set. The observed mean and median prices are indicated by the red vertical broken lines.

Figures C.3, C.4 and C.5 include the distributions for the replicated mean prices per neighborhood from the estimation set for each model. The figures include only those neighborhoods with at least one observation. The observed mean prices are denoted with the red markers.

Figure C.6 gives the histograms for the probability inverse transforms for each model. Figure C.7 describes the sharpness of the predictive distributions from each model.

Figure C.8 gives the scatter plots of each models point predictions, averaged from the predictive distributions, and the realized price in the estimation set. The red broken line in the figures is the $y = x$-line: The points should lie on this line for an ideal predictor. Figure C.9 includes the scatter plot of the point predictions and the realized prices of each model for the test set.

Four observations (identifiers 2748, 3084, 3089 and 2753 in figure C.9) were chosen to illustrate the underlying predictive distributions. Figure C.10 describes the predictive distributions of each model for these observations. The realized prices are denoted with the vertical red broken line.

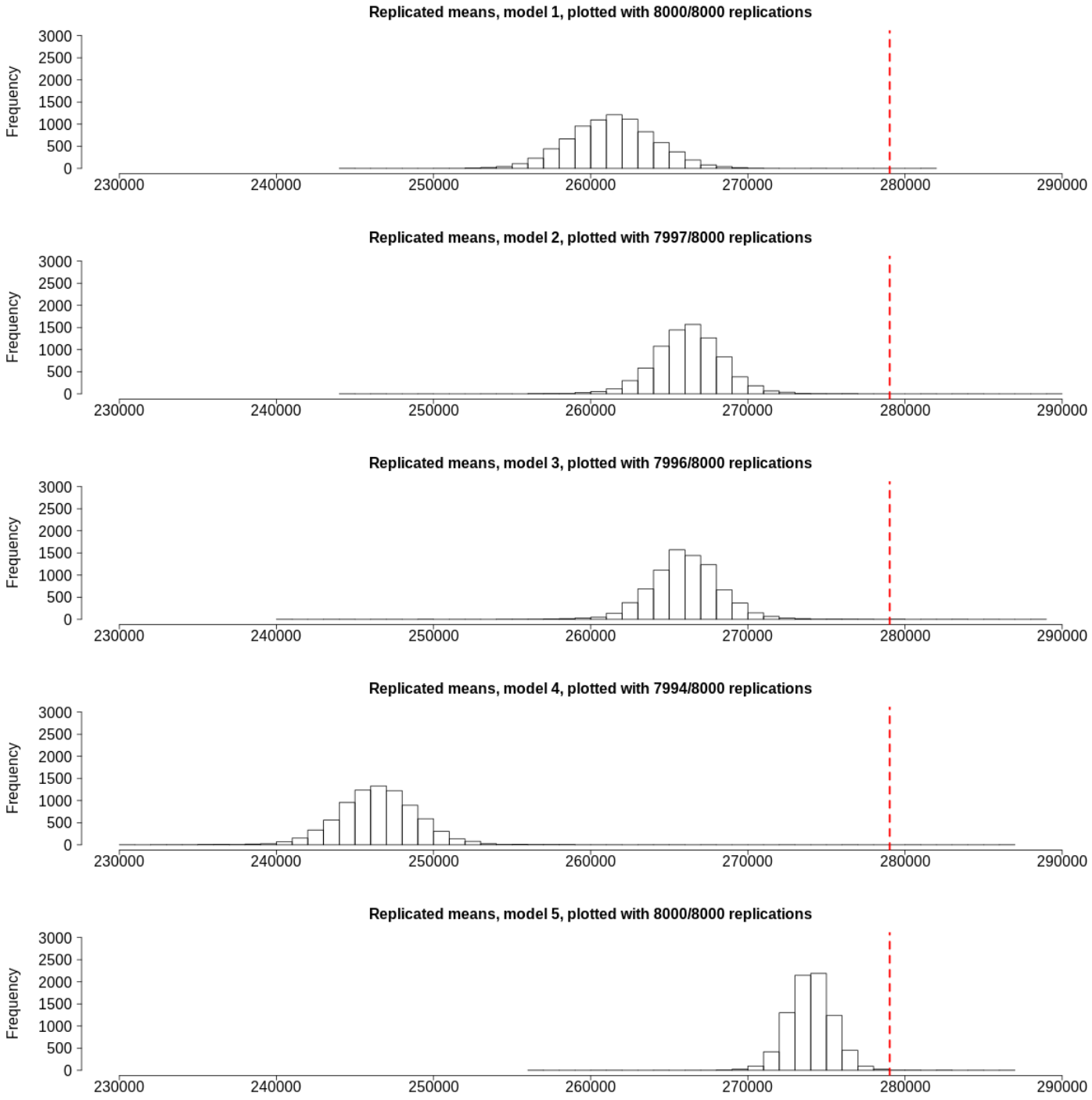


Figure C.1: Replicated mean histograms, estimation set

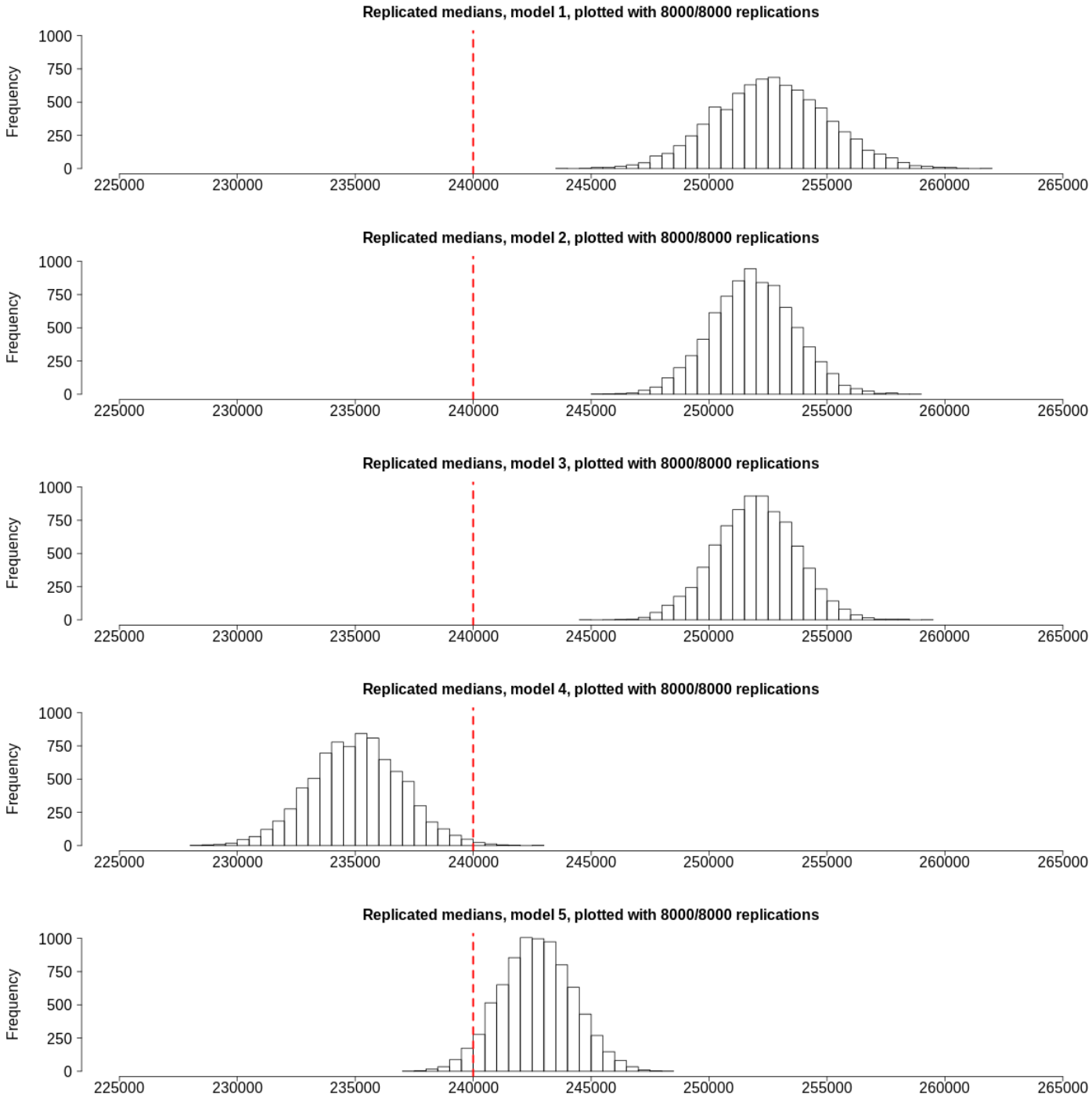


Figure C.2: Replicated median histograms, estimation set

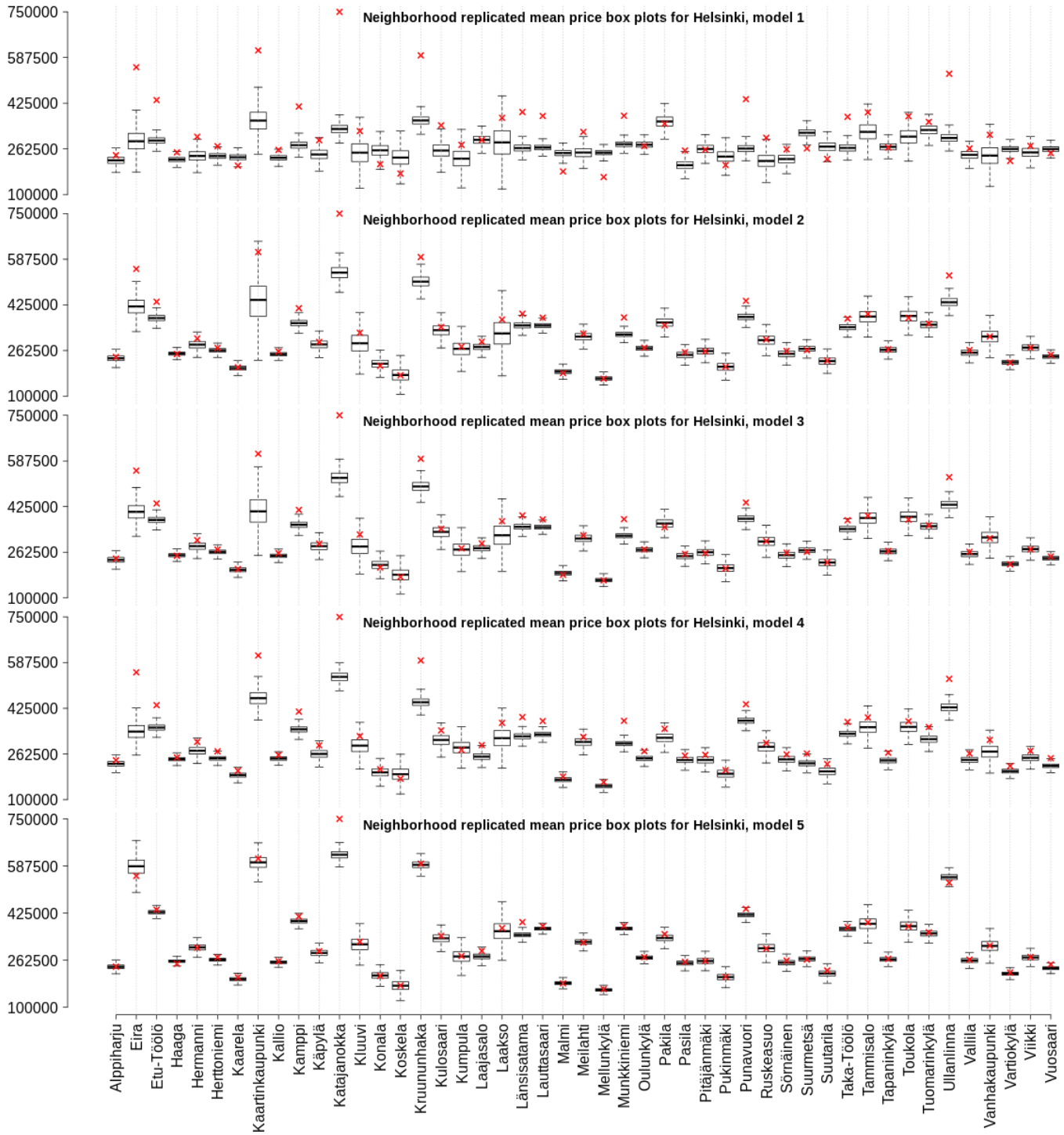


Figure C.3: Distributions of replicated mean prices per neighborhood, Helsinki, estimation set

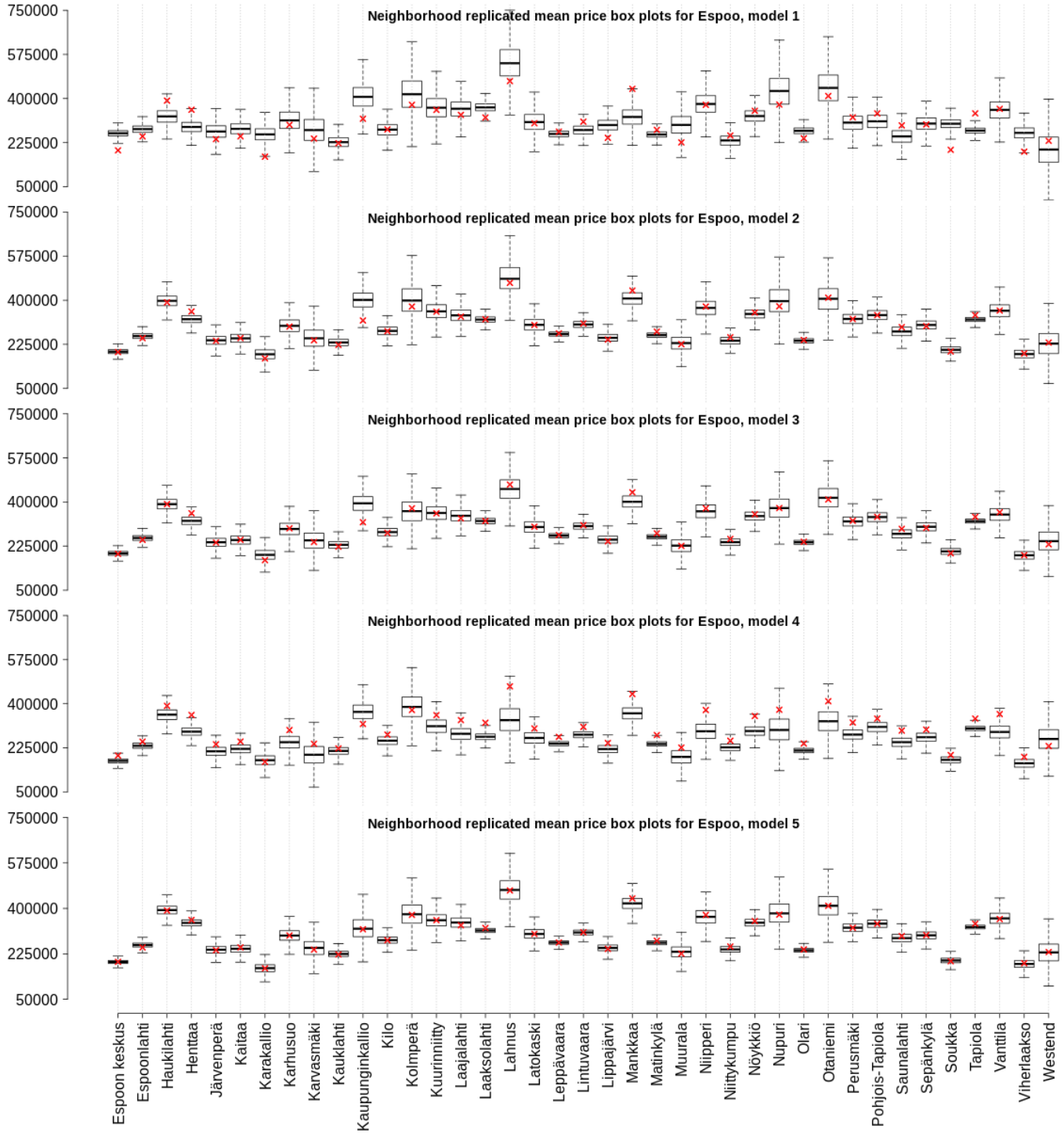


Figure C.4: Distributions of replicated mean prices per neighborhood, Espoo, estimation set

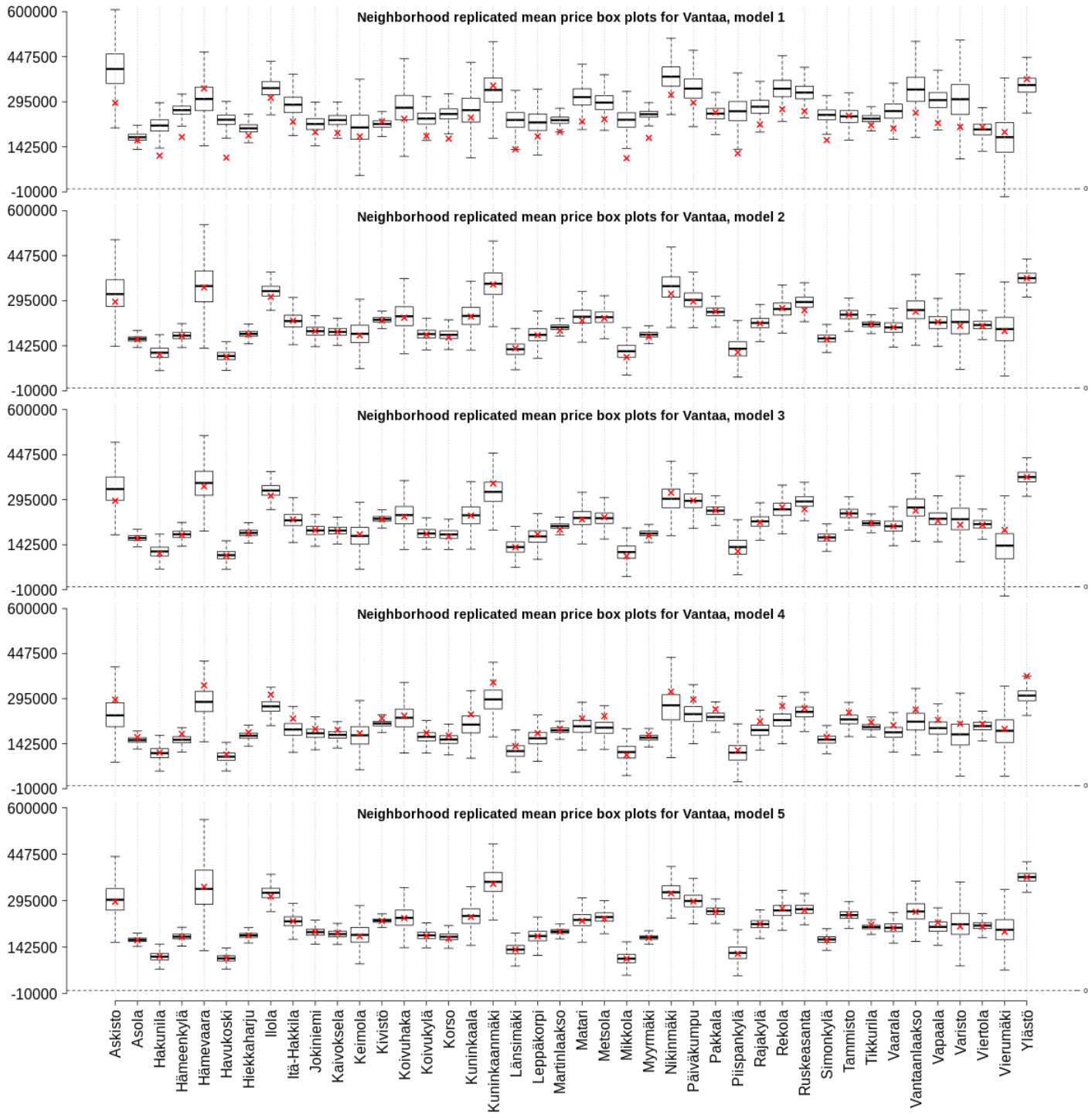


Figure C.5: Distributions of replicated mean prices per neighborhood, Vantaa, estimation set

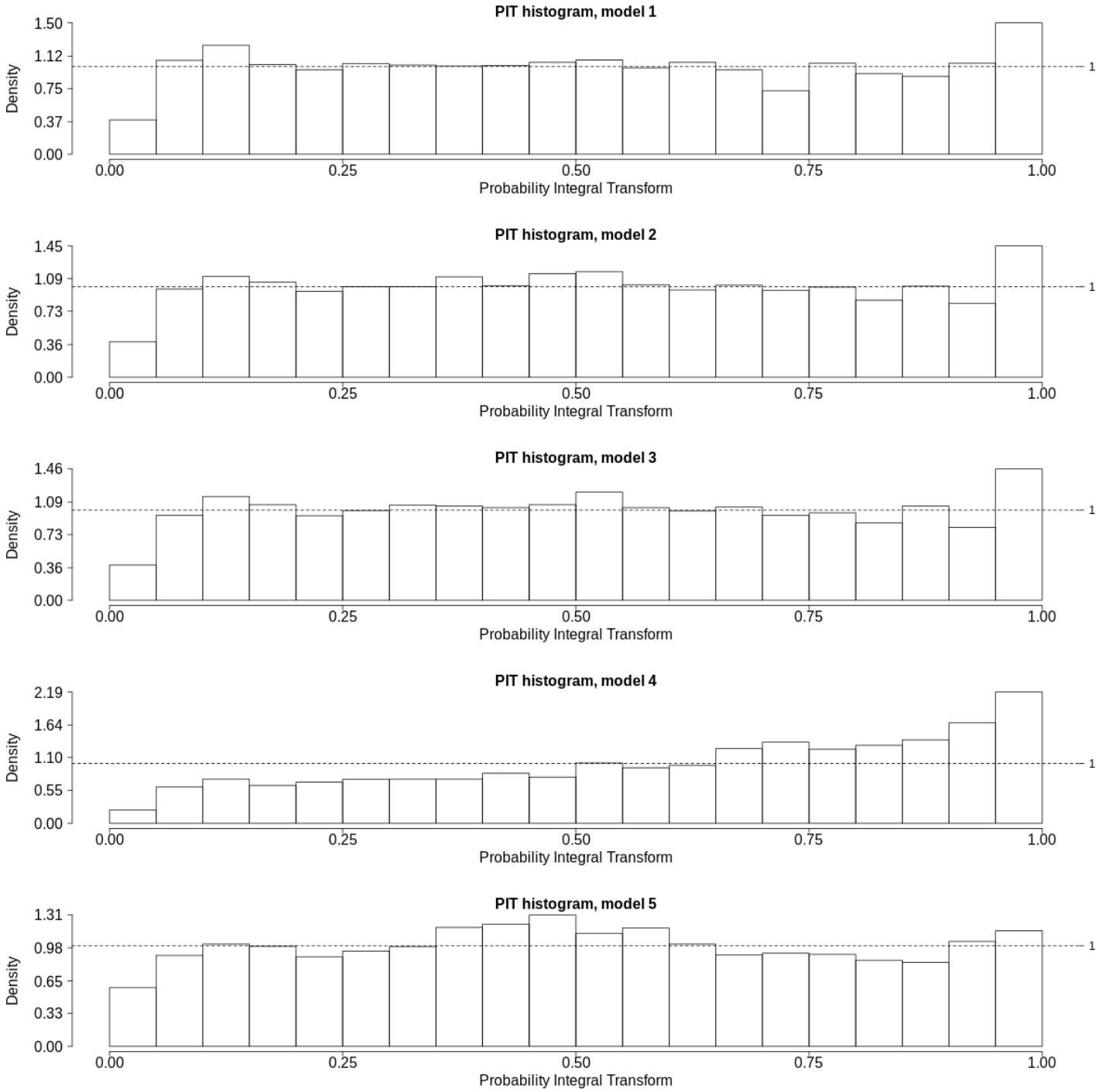


Figure C.6: PIT histograms, estimation set

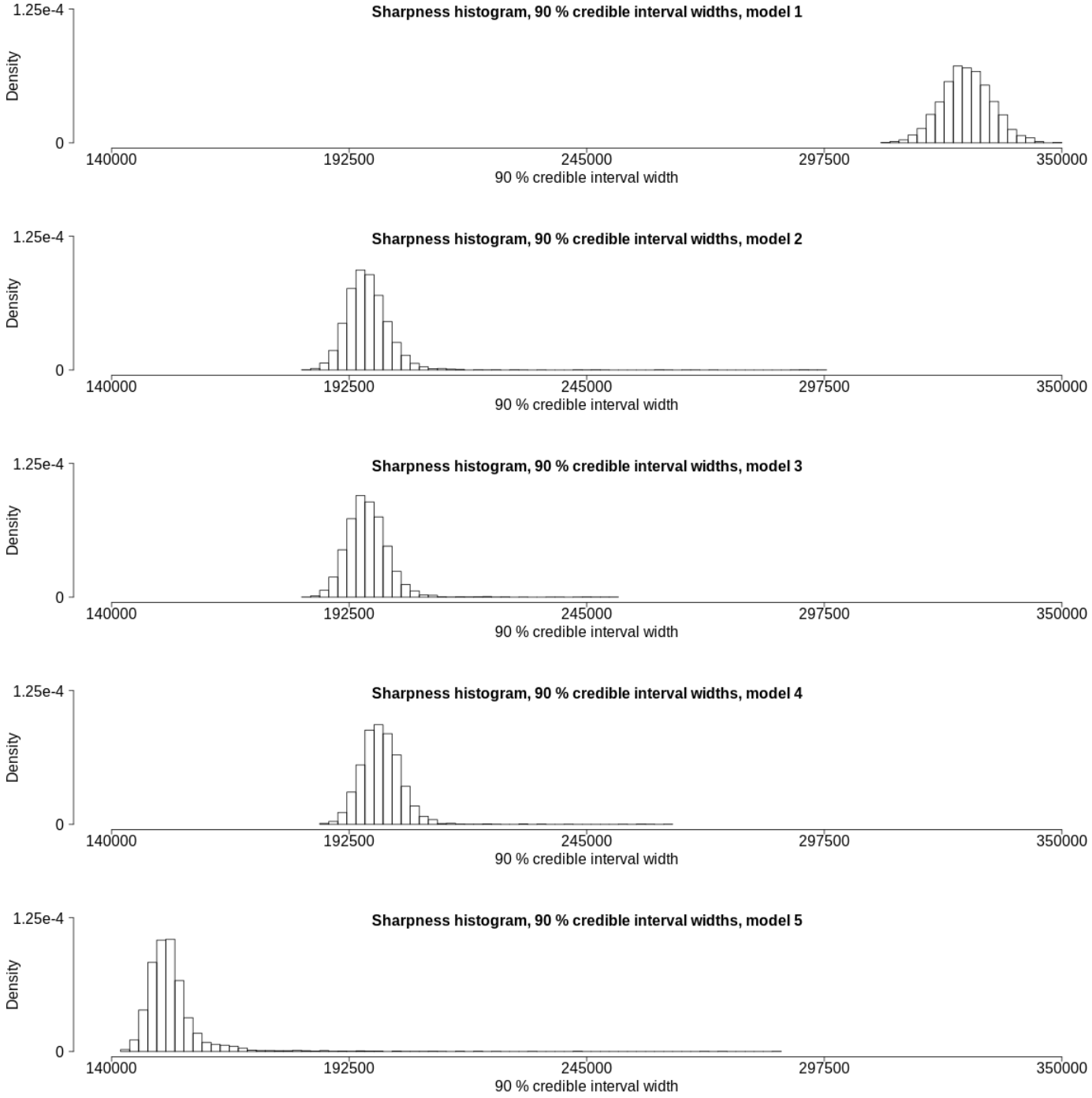


Figure C.7: Sharpness histograms, estimation set

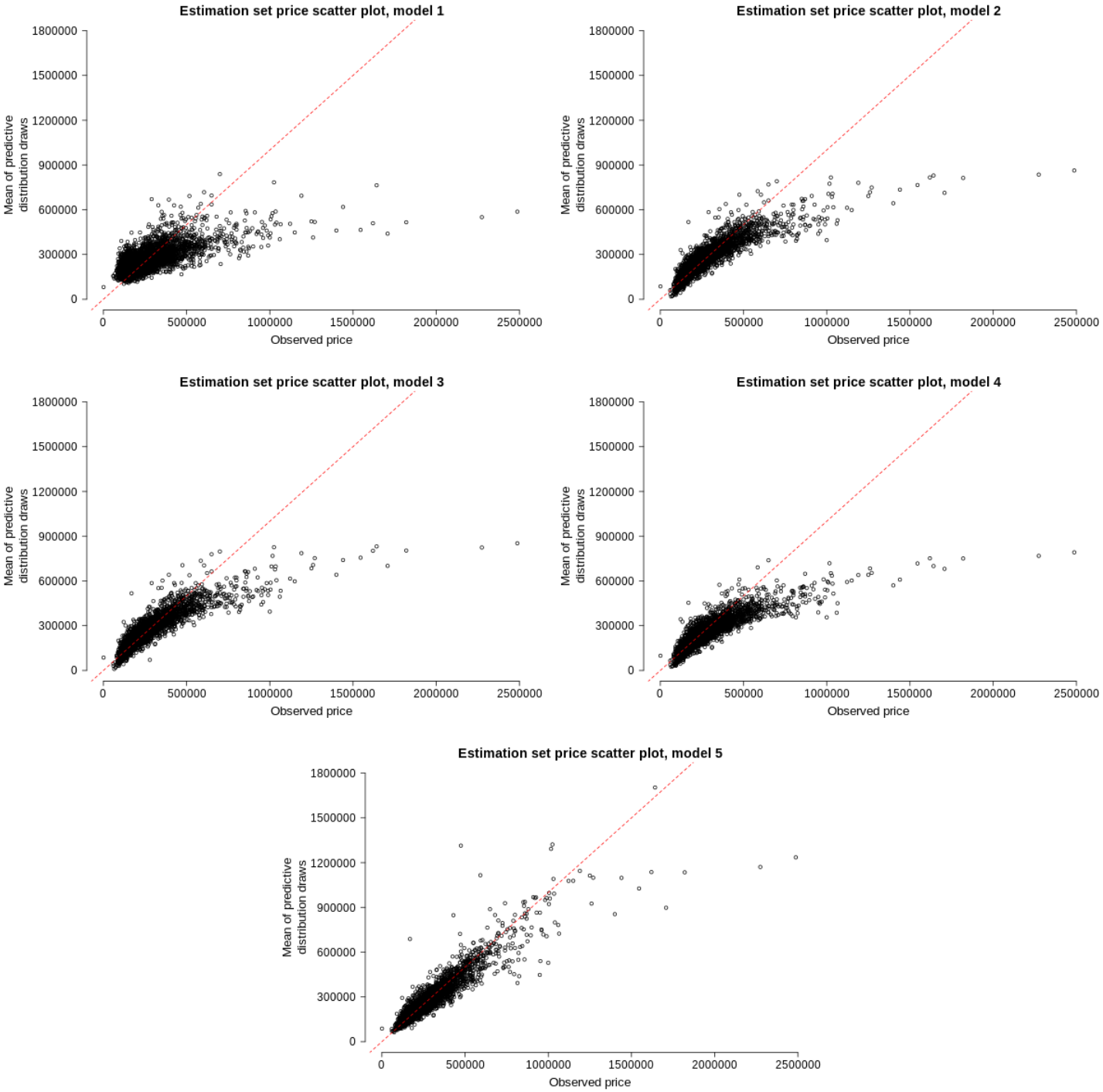


Figure C.8: Estimation set scatter plots

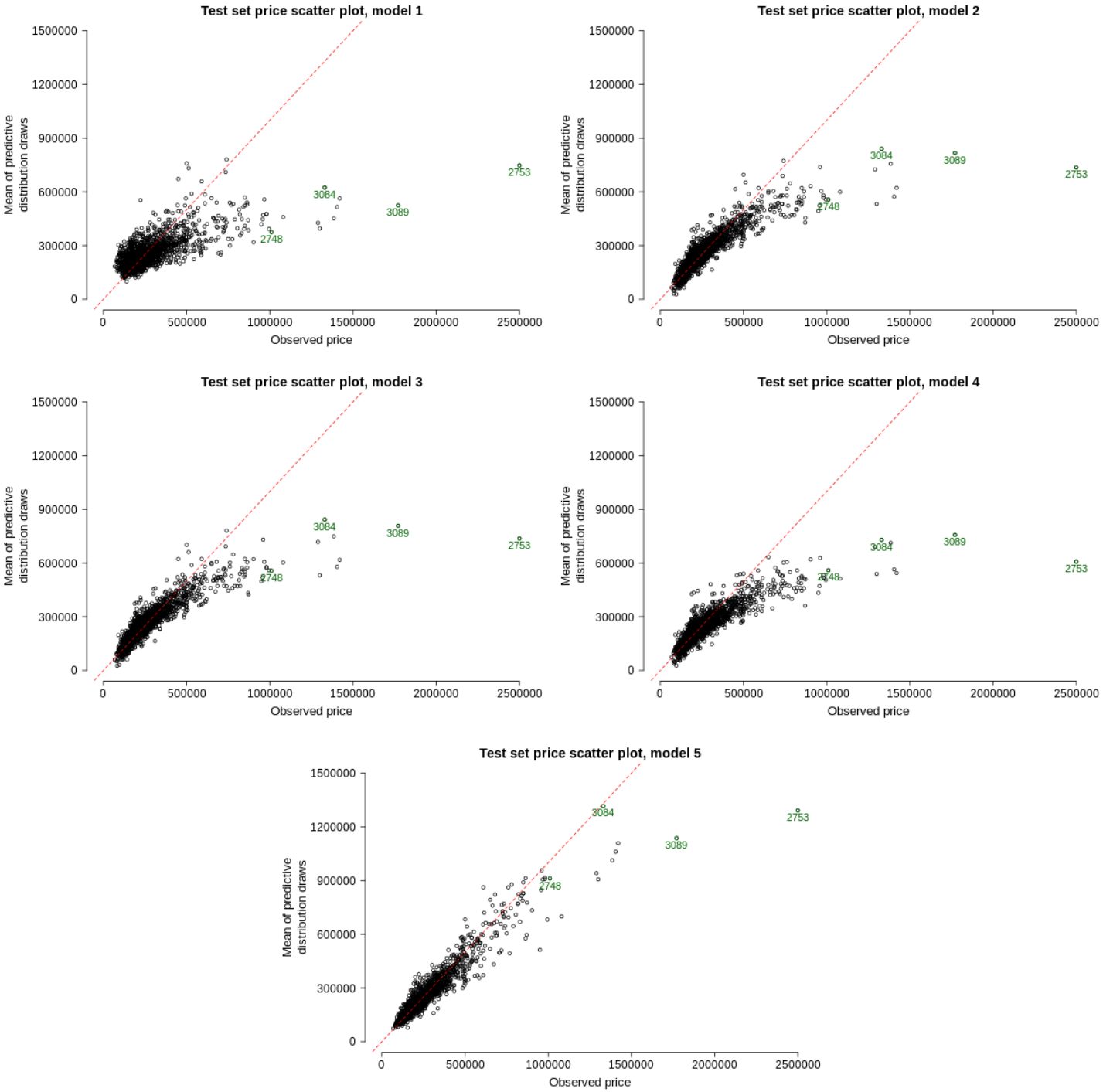


Figure C.9: Test set scatter plots

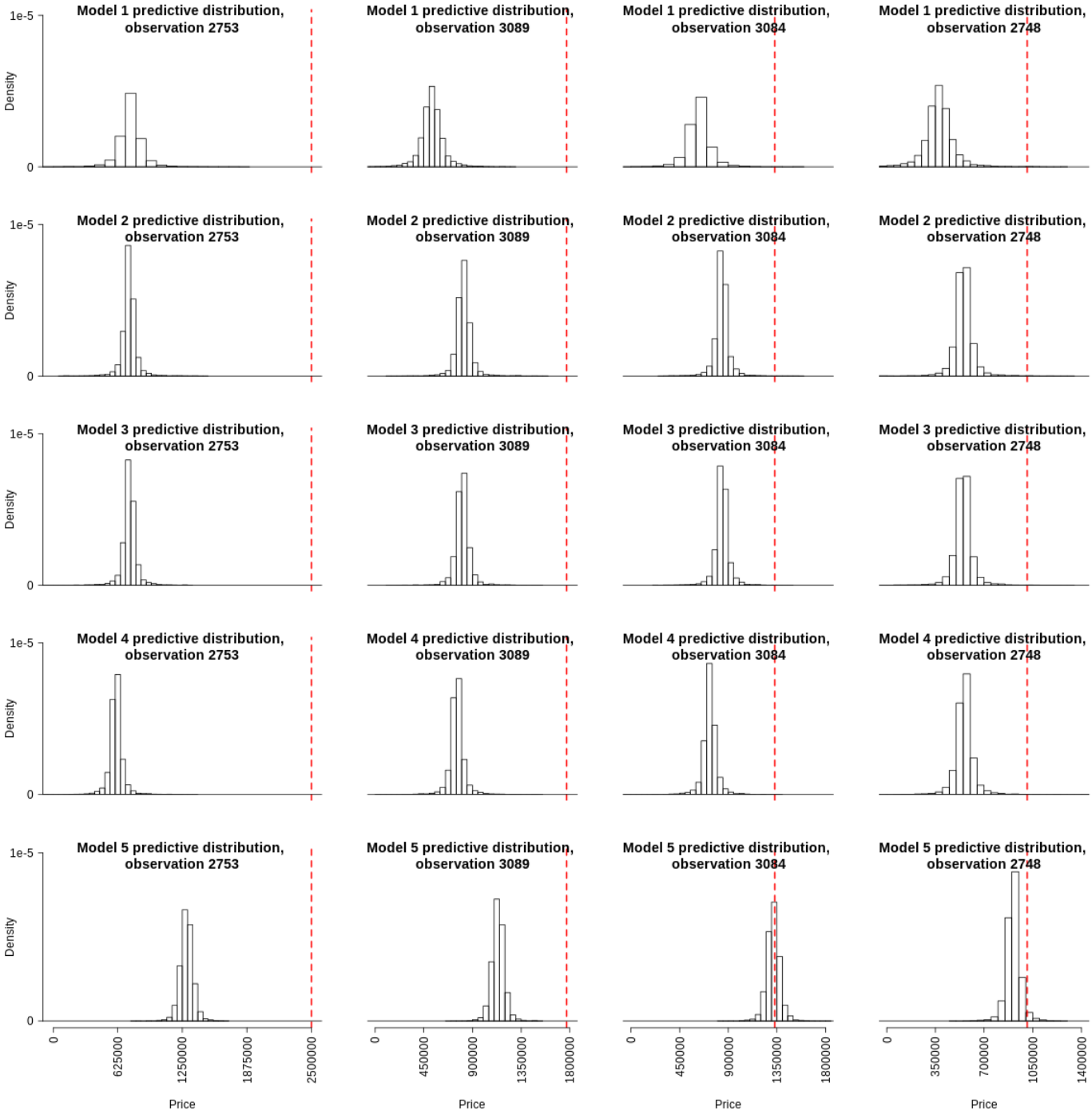


Figure C.10: Predictive distributions for chosen observations

# Appendix D

# Figures for the stacking model

Figures D.1 and D.2 give the distributions of the replicated mean and median prices of the estimation set from the stacking model. Figures D.3, D.4 and D.5 describe the replicated mean prices per neighborhood for the estimation set using the stacking model. Figure D.6 gives the PIT histogram for the stacking model. Figure D.7 gives the sharpness histogram for the stacking model. Figure D.8 is the scatter plot of the observed prices of the estimation set and the point predictions, averaged from the predictive distributions of the stacking model. Figure D.9 gives the scatter plot of the test set. The red broken line in the figures is the $y = x$-line: The points should lie on this line for an ideal predictor. Figure D.10 gives the predictive distributions for observations 2748, 3084, 3089 and 2753.

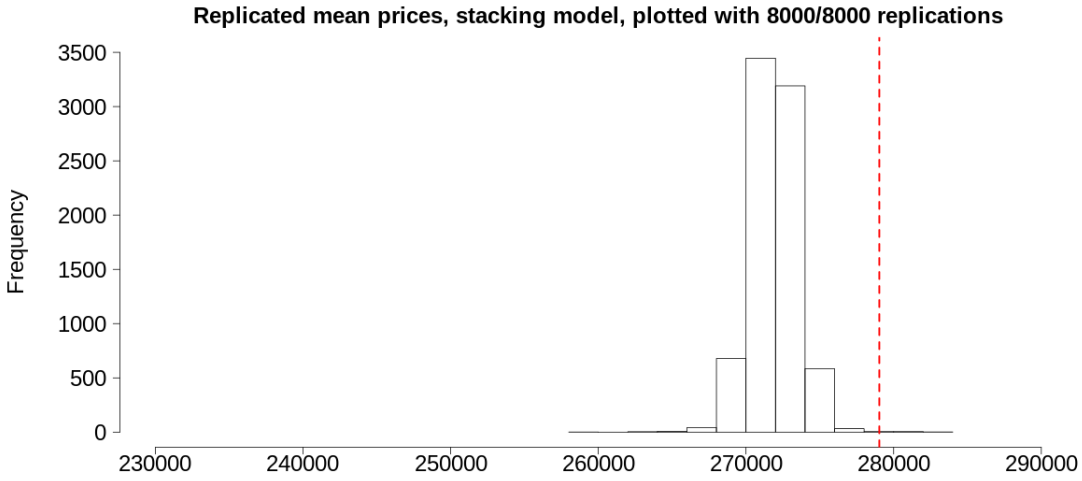


Figure D.1: Replicated mean prices histogram, estimation set, stacking model

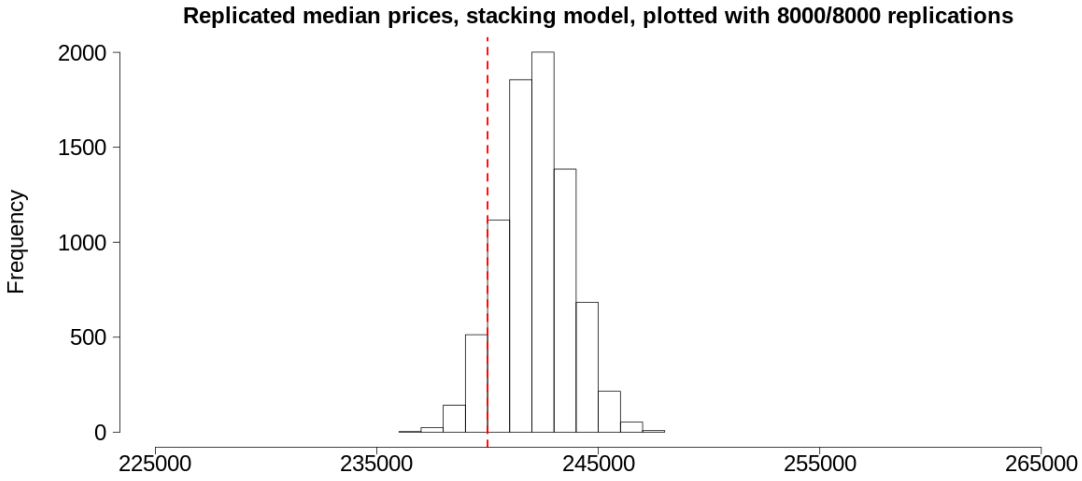


Figure D.2: Replicated median prices histogram, estimation set, stacking model

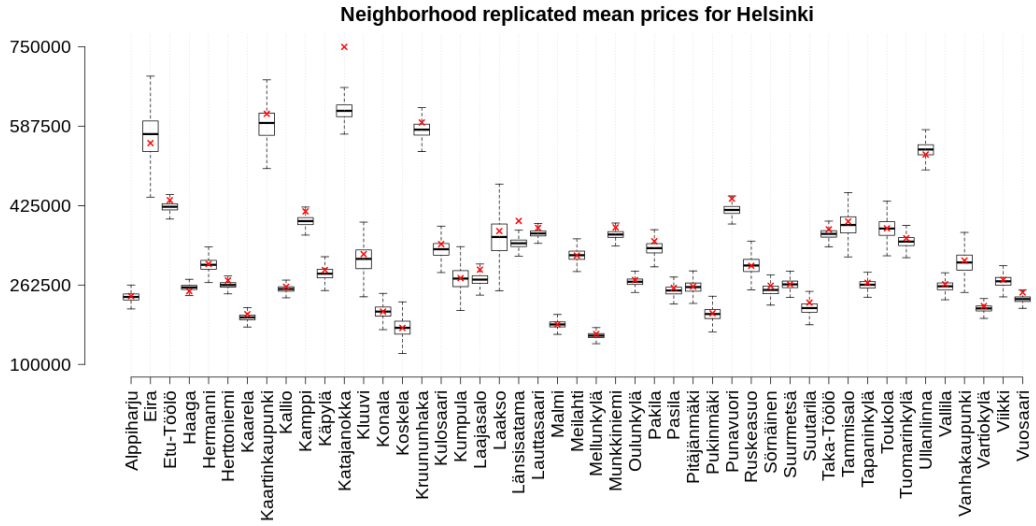


Figure D.3: Distributions of replicated mean prices per neighborhood for Helsinki, estimation set, stacking model

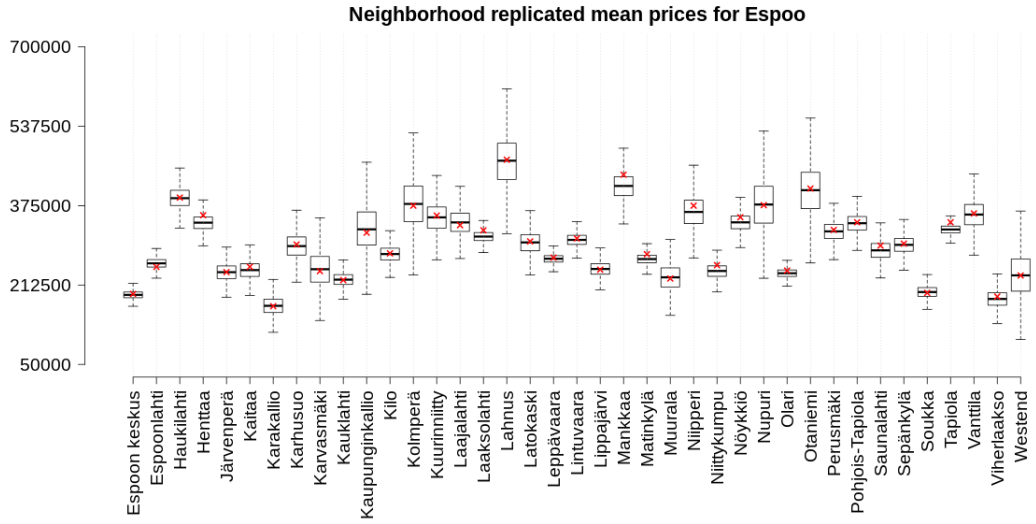


Figure D.4: Distributions of replicated mean prices per neighborhood for Espoo, estimation set, stacking model

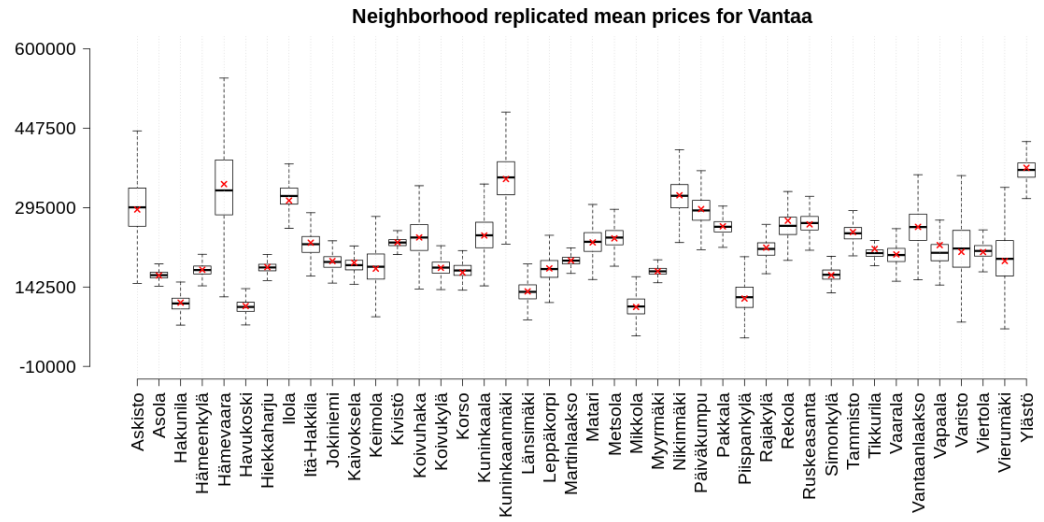


Figure D.5: Distributions of replicated mean prices per neighborhood for Vantaa, estimation set, stacking model

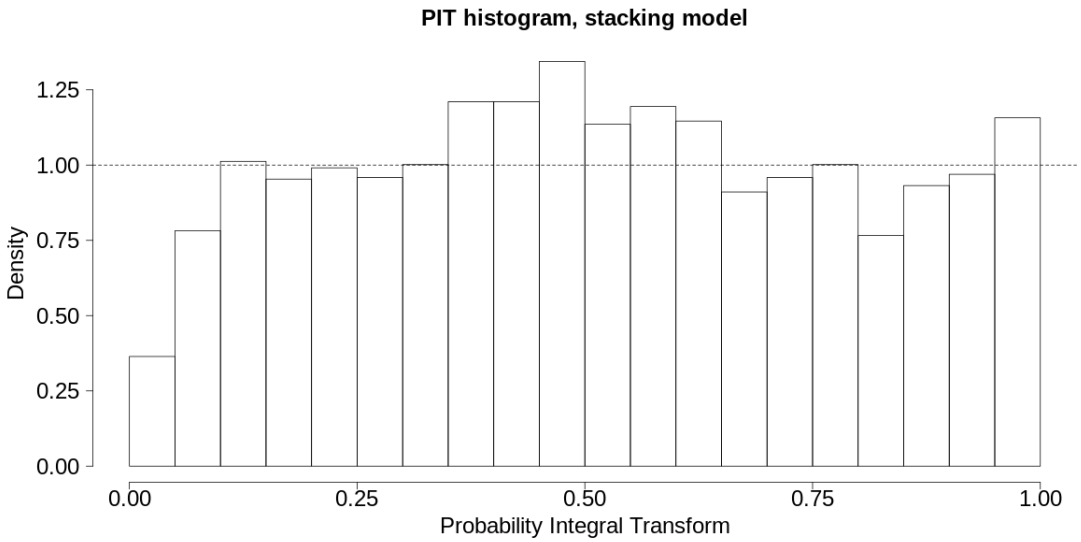


Figure D.6: PIT histogram, estimation set, stacking model

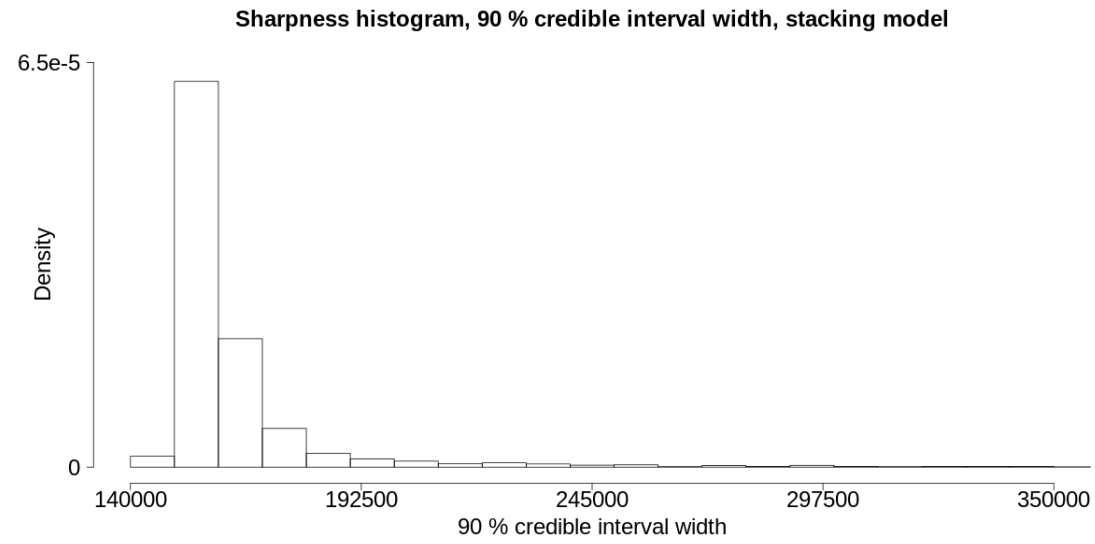


Figure D.7: Sharpness histogram, estimation set, stacking model

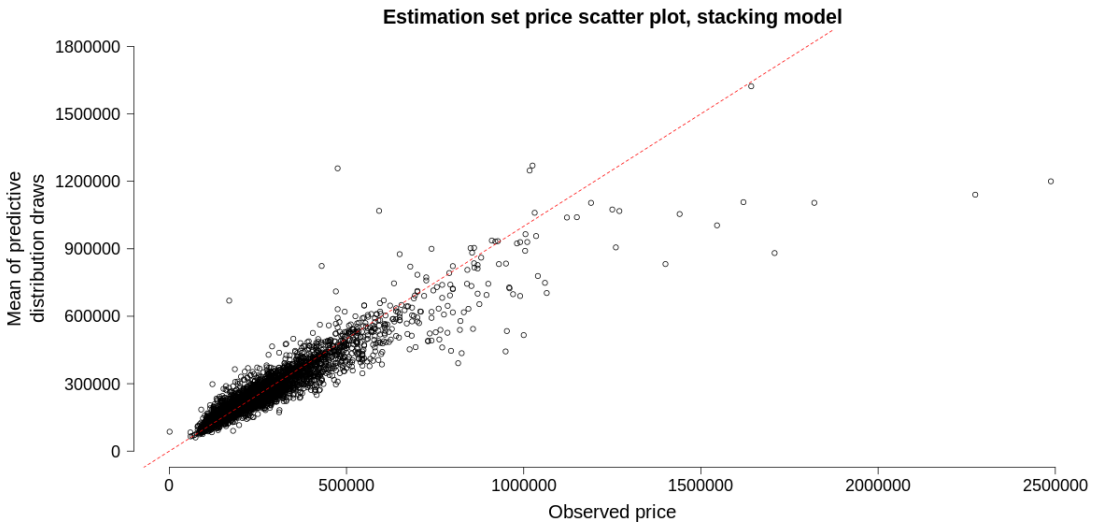


Figure D.8: Estimation set scatter plots, stacking model

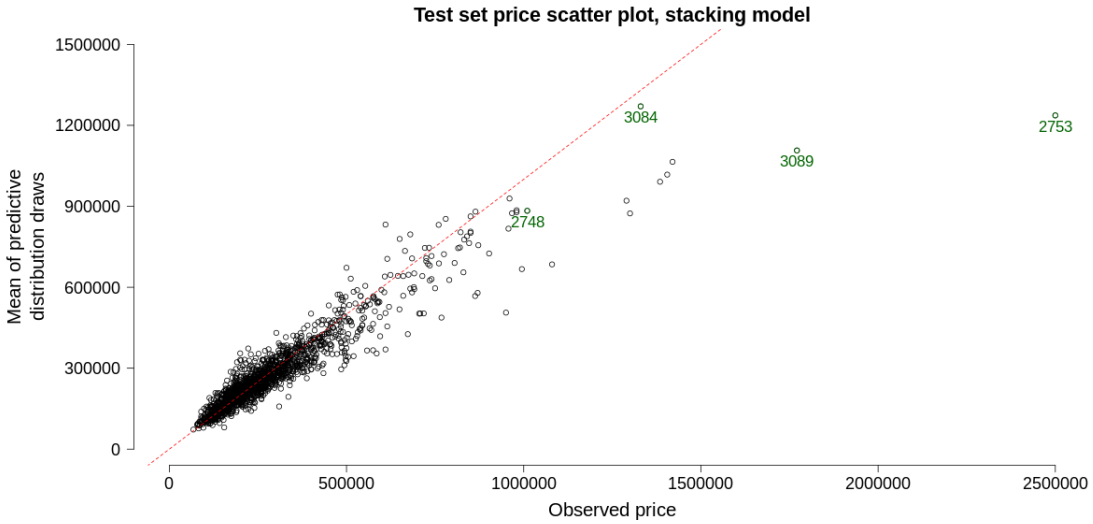


Figure D.9: Test set scatter plots, stacking model

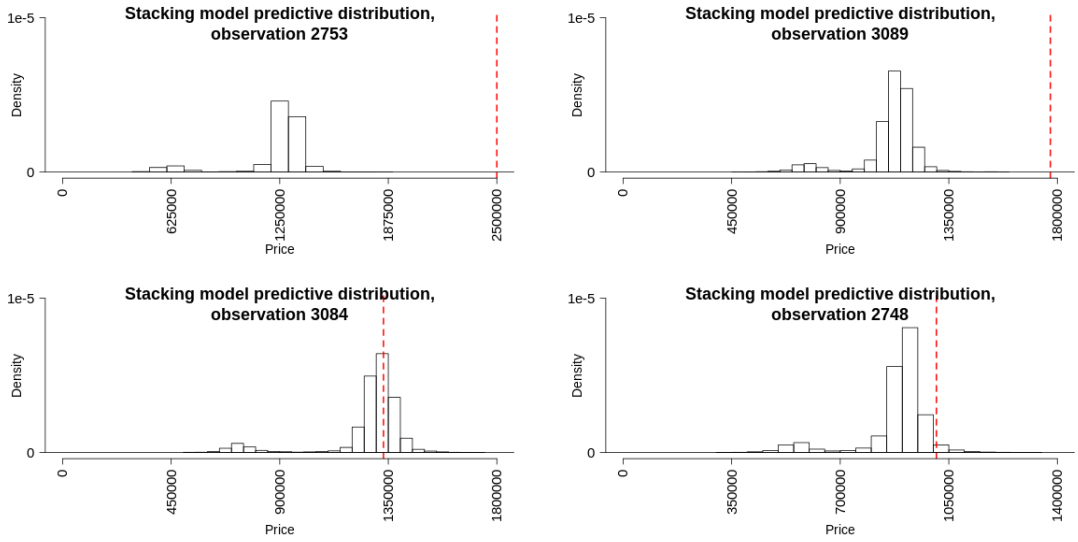


Figure D.10: Predictive distributions for chosen observations, stacking model

# Bibliography


Evgeny A. Antipov and Elena B. Pokryshevskaya. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39 (2):1772–1778, 2012. doi: 10.1016/j.eswa.2011.08.077.

Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press - Tailor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton FL, 2 edition, 2015.

Jose M. Bernardo and Adrian F. M. Smith. *Bayesian theory*. John Wiley and Sons, Ltd, 1994.

William M. Bowen, Brian A. Mikelbank, and Dean M. Prestegaard. Theoretical and Empirical Considerations Regarding Space in Hedonic Housing Price Model Applications. *Growth and Change*, 32(4):466–490, 2001. doi: https://doi.org/10.1111/0017-4815.00171.

Juhana Brotherus. Asuntojen hinnanmuodostus hedonisella menetelmällä. Master's thesis, Aalto University, 2011.

Ayse Can. Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, 22(3):453–474, 1992. doi: 10.1016/0166-0462(92)90039-4.

Robin A. Dubin. Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *The Review of Economics and Statistics*, 70(3):466–474, 1988. doi: 10.2307/1926785.

Robin A. Dubin. Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22(3):433–452, 1992. doi: 10.1016/0166-0462(92)90038-3.

Robin A. Dubin. Spatial Autocorrelation: A Primer. *Journal of Housing Economics*, 7(4):304–327, 1998. doi: 10.1006/jhec.1998.0236.

Essi Eerola and Teemu Lyytikäinen. On the role of public price information in housing markets. *Regional Science and Urban Economics*, 53:74–84, 2015. doi: 10.1016/j.regsciurbeco.2015.05.006.

Essi Eerola and Tuukka Saarimaa. Delivering affordable housing and neighborhood quality: A comparison of place- and tenant-based programs. *Journal of Housing Economics*, 42:44–54, 2018. doi: 10.1016/j.jhe.2017.12.001.

A. Stewart Fotheringham, Martin E. Charlton, and Chris Brunsdon. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, 30(11):1905–1927, 1998. doi: 10.1068/a301905.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Statistics in Society Series A*, 182(2):389–402, 2019. doi: 10.1111/rssa.12378.

Alan E. Gelfand, Hyon-Jung Kim, C. F. Sirmans, and Sudipto Banerjee. Spatial Modeling with Spatially Varying Coefficient Processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003. doi: 10.1198/016214503000170.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006. doi: 10.1214/06-BA117A.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK, 2006.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press - Tailor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton FL, 3 edition, 2014.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19 (10):555, 2017. doi: 10.3390/e19100555.

John Geweke. Bayesian treatment of the independent student–t linear model. *Journal of applied econometrics*, 8(S1):S19–S40, 1993. doi: 10.1002/jae.3950080504.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007. doi: 10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Allen C. Goodman and Thomas G. Thibodeau. Housing Market Segmentation. *Journal of Housing Economics*, 7(2):121–143, 1998. doi: 10.1006/jhec.1998.0229.

Jani Hakala. Helsinkiläisten kerrostaloasuntojen hinnanmuodostus. Master's thesis, University of Helsinki, 2018.

Oskari Harjunen. *Essays in urban economics and housing market capitalization.* PhD thesis, Aalto University, 2018.

Marcelo Hartman and Jarno Vanhatalo. Laplace approximation and natural gradient for Gaussian process regression with heteroscedastic student–t model. *Statistics and Computing*, 29(4):753–773, 2019. doi: 10.1007/s11222-018-9836-0.

Sam K. Hui, Alvin Cheung, and Jimmy Pang. A Hierarchical Bayesian Approach for Residential Property Valuation: Application to Hong Kong Housing Market. *International Real Estate Review*, 12(3):1–29, 2010.

Panu Jantunen. Täydennysrakentamisen vaikutus ympäröivien asuntojen hintoihin. Master's thesis, Aalto University, 2017.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust Gaussian Process Regression with a Student–t Likelihood. *Journal of Machine Learning Research*, 12:3227–3257, 2011.

Antti Koivuniemi. Hedoniset hinnoittelumallit asuntomarkkinoilla. Master's thesis, University of Turku, 2014.

Mika Kortelainen and Tuukka Saarimaa. Do Urban Neighborhoods Benefit from Homeowners? Evidence from Housing Prices. *The Scandinavian Journal of Economics*, 117(1):28–56, 2015. doi: 10.1111/sjoe.12090.

Lauri Kyllönen and Tarmo Räty. Asuntojen hinta-laatusuhde joensuussa, semiparametrinen estimointi. *VATT-tutkimuksia*, 2000.

Joona Luhta. Vaasan asuntomarkkinoiden alueellinen eriytyminen. Master's thesis, University of Vaasa, 2017.

Richard McElreath. *Statistical Rethinking.* CRC Press - Tailor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton FL, 2016.

Radford M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, September 1993.

Radford M. Neal. MCMC using Hamiltonian Dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. CRC Press - Tailor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton FL, 2011.

Anthony O'Hagan. On Outlier Rejection Phenomena in Bayes Inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3):358–367, 1979. doi: 10.1111/j.2517-6161.1979.tb01090.x.

Anthony O'Hagan. *Kendall's Advanced Theory of Statistics - Bayesian Inference*, volume 2B. Arnold, Hodder Headline Group, 338 Euston Road, London NW1 3BH, 1994.

Sherwin Rosen. Hedonic Prices and Implicit Markets: Product Differation in Pure Competition. *Journal of Political Economy*, 82(1):34–55, 1974. doi: 10.1086/260169.

Hasan Selim. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, 36(2):2843–2852, 2009. doi: 10.1016/j.eswa.2008.01.044.

Stephen Sheppard. Hedonic analysis of housing markets. *Handbook of Regional and Urban Economics*, 3:1595–1635, 1999. doi: 10.1016/S1574-0080(99)80010-8.

Niklas Takala. Ostajan paikkakunnan vaikutus kerrostaloasunnon kauppahintaan. Master's thesis, Aalto University, 2016.

Anniina Valaja. Raitiotien vaikutus asuntojen hintoihin tampereella. Master's thesis, Aalto University, 2018.

Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with student–t likelihood. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1910–1918. Curran Associates, Inc., 2009.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. doi: 10.1007/s11222-016-9696-4.

Lassi Vuorela. Taloyhtiön omaisuus asunto-osakekaupassa. Master's thesis, Aalto University, 2019.

Mike West. Outlier Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46 (3):431–439, 1984. doi: 10.1111/j.2517-6161.1984.tb01317.x.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using Stacking to Average Bayesian Predictive Distribution (with Discussion). *Bayesian Analysis*, 13(3):917–1007, 2018. doi: 10.1214/17-BA1091.

Danlin Yu, Yehua Dennis Wei, and Changshan Wu. Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning B: Planning and Design*, 34(6):1085–1102, 2007. doi: 10.1068/b32119.