Volume 7, Number 1, 2020

REGION

The Journal of ERSA
Powered by WU

Table of Contents

Articles

Funded by

REGION, The journal of ERSA / Powered by WU
ISSN: 2409-5370

ERSA: http://www.ersa.org
WU: http://www.wu.ac.at

Articles

REGION
The Journal of ERSA
Powered by WU

Volume 7, Number 1, 2020, 1–19
DOI: 10.18335/region.v7i1.279

journal homepage: region.ersa.org

Random Parameters and Spatial Heterogeneity using *Rchoice* package in *R*

Mauricio Sarrias[1]

[1] Universidad Católica del Norte, Antofagasta, Chile

Received: 17 September 2019/Accepted: 4 February 2020

Abstract. This document provides a brief introduction to models with spatial heterogeneity using a random parameter approach. Specifically, this paper shows how this modelling strategy can be used to capture and model spatial heterogeneity and locally varying coefficients for different latent structure. To show the main advantages of this modeling strategy, the `Rchoice` package (Sarrias 2016) in `R` is used. The examples will be focused on the ordered probit model with spatially varying coefficients using self-assessed health status as the dependent variable.

1 Introduction

Regional scientists, as well as many social researchers concerned on spatial relationships, analyze how the reciprocal geographical interaction of social agents generates spatial autocorrelation, affecting the bias and efficiency of standard econometric estimators. After Anselin (1988), a large number of papers dealt with the spatial autocorrelation using spatial versions of standard linear regression models, namely Spatial Autoregresive Regression (SAR) or Spatial Error Model (SEM) and even recent contributions extend the analysis toward Spatial Panel Data (Kelejian, Prucha 1998, 1999, Elhorst 2014). However, spatial interaction also is manifested through spatially varying coefficients referred to as: "structural instability over space, in the form of different response functions or systematically varying parameters" (Anselin 1988). In spite of the relevance of the concept, the evolution and development of econometric models that attempt to capture and model spatial heterogeneity has not been as euphoric as those focused on the spatial autocorrelation. The few attempts to capture this heterogeneity can be summarized by the spatial expansion method (SEM) (Casetti 1972), Geographically Weighted Regression (GWR) (Brunsdon et al. 1998) or assuming that the local relationship varies randomly over geographical space, a method also known as the Random Coefficient Model (RCM) (Swamy 1971). Each one of these methods enable estimation of model parameters locally, or they allow model parameter to vary as a function of location[1].

The three methods presented above share an important limitation: they require aggregating the variables at the location level. Therefore, we are prevented from using data at the individual level and capturing the spatial heterogeneity, simultaneously. This raises concerns about the misleading conclusions that can be derived at the individual level by using aggregate variables known as the ecological fallacy problem (Robinson 1950). A potential solution for this constraint is provided by multilevel modeling[2]. This

[1] For further review see for example Fotheringham, Brunsdon (1999).
[2] For other quantitative methods that avoid the ecological fallacy problem see Withers (2001).

approach separates the effect of personal and place characteristics to investigate the extent and nature of spatial variation in individual outcome measures (Goldstein 1987). The main drawback of multilevel modeling is that usually the random coefficients are assumed to be normally distributed. This makes the estimation process easier, but creates other problems. For example, this assumption implies that some locations might have positive or negative coefficients, whether or not this is true. In practice, this implies that occasionally researchers find sign reversals that are counterintuitive and difficult to explain. Furthermore, the domain of the normal distribution is $(-\infty, +\infty)$, which results in unreliable extreme coefficients and high coefficient variability. Those problems have also been found when applying the GWR approach (Jetz et al. 2005)[3].

This study focusses on models with spatially varying coefficients using simulation as in Sarrias (2019) and Train (2009). This modeling strategy is intended to complement the existing approaches by using variables at the micro level – overcoming the problem associated with spatial aggregation – and by adding flexibility and realism to the potential domain of the coefficient on the geographical space. Spatial heterogeneity is modelled by allowing the parameters associated with each observed variable to vary "randomly" across space according to some distribution $g(\cdot)$. However, it is not known how the parameters vary across space. All that is known is that they vary locally with population probability density function (pdf) $g(\cdot)$, which is assumed to be well behaved and continuous.

To show the main advantages of this modeling strategy, the `Rchoice` package (Sarrias 2016) in `R` is used. The examples will be focused on the ordered probit model with spatially varying coefficients using self-assessed health status as the dependent variable.

The remainder of this paper is organized as follows. Section 2 discusses the modelling approach for incorpo-rating continuous spatial heterogeneity using a random parameter approach. The main R packages needed for the examples are described in Section 3. The example using `Rchoice` package in `R` is presented in Section 4. Finally, Section 5 concludes.

2 Modelling approach

2.1 Continuous spatial heterogeneity

Consider the following structural model:

$$\begin{aligned} y_{ci}^{*} &= \boldsymbol{x}_{ci}'\boldsymbol{\beta}_c + \epsilon_{ci} \qquad c = 1, \ldots, C; \quad i = 1, \ldots, n_c \\ \boldsymbol{\beta}_c &\sim g(\boldsymbol{\beta}_c) \end{aligned} \tag{1}$$

where y_{ci}^{*} is a latent (unobserved) process for individual i in geographical area c (e.g, region, city, country, census track) that we are trying to explain; $\boldsymbol{x}_{ci}$ is a $K \times 1$ vector of individual and regional variables; and ϵ_i is the error term[4]. It is assumed that the vector $(y_{ci}, \boldsymbol{x}_{ci}', \boldsymbol{\beta}_c')'$ is independently and identically distributed. The conditional probability density function of the latent process, $f^{*}(y_{ic}|\boldsymbol{x}_{ci}, \boldsymbol{\epsilon}_c)$, is determined once the nature of the observed y_{ci} and the population pdf of ϵ_i is known. For example, if the observed y_{ci} is binary and ϵ_i is normal distributed, we obtain the traditional probit model. But if ϵ_i is distributed as logistic, then we obtain the binary logit model. Due to space restrictions, the applied example in this study will focus on the ordered model.

The key element in the structural model is $\boldsymbol{\beta}_c$. The notation implies that coefficients are associated with region c, representing those region-specific partial correlations on the latent dependent variable. Thus, all individuals located in the same region have the same coefficient, but there exists inter-spatial heterogeneity, i.e., the coefficients vary across regions but not within the region.

[3] There are some interesting extensions that have been recently developed. For example, Dong et al. (2015) extend the traditional multilevel models to incorporate spatial interaction effects at different level units. Dong et al. (2018) extend the GWR for ordinal categorical responses. Bayesian spatially varying coefficient models have been also suggested by Finley (2011) and Gelfand et al. (2003).

[4] Throughout this work I will use location unit, region, or geographical area interchangeably to refer to the subindex c.

However, we do not know how these parameters vary across regions. All we know is that they vary locally with population pdf $g(\boldsymbol{\beta}_c)$. Once $g(\boldsymbol{\beta}_c)$ is specified, we might have a fully parametric or a semi-parametric spatially random parameter model.

2.2 Choosing the distribution

Continuous spatial heterogeneity is introduced by assuming that the parameters vary "randomly" across regions according to some pre-specified "continuous" distribution. The pdf of the spatially random coefficients in the population is $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents, for example, the mean and variance of $\boldsymbol{\beta}_c$. The goal for the researcher is to estimate $\boldsymbol{\theta}$.

The distribution of the spatially random parameters can in principle take any shape. The researcher has to choose a priori the distribution according to his beliefs of the domain and boundedness of the coefficients.

Therefore, some prior theoretical knowledge of the spatial structure being modeled may lead to a more appropriate choice of the distribution. Below, some continuous distributions and their implications are discussed.

Normal Distribution: The normal distribution is by far the most widely used distribution for the spatially random parameters. The density of the normal parameter has mean β and standard deviation σ_β, so that $\boldsymbol{\theta} = (\beta, \sigma_\beta)'$. Thus, the coefficient for each region can be written as $\beta_c = \beta + \sigma_\beta \eta_c$, where $\eta_c \sim N(0,1)$. An important feature of the normal density is its unboundedness. This implies that every real number has a positive probability of being drawn. Thus, specifying a given coefficient to follow a normal distribution is equivalent to making the a priori assumption that there is a proportion of regions with positive coefficients and another proportion with negative ones. As an illustration, consider a normally distributed coefficient with population parameters $\beta = 0.5$ and $\sigma_\beta = 1$. The proportion of regions with positive coefficients is approximately $\Phi(\beta/\sigma_\beta) \cdot 100 \approx 70\%$. This last fact makes this distribution quite suitable when the researcher assumes that the effect of x_k on y^* can have both signs depending in the local context of each region. For example, there exists an extensive literature that uses the city population as a proxy for urbanization economies (see for example Duranton, Puga 2004). However, in some regions, a large population may suggest agglomeration economies, while in others, it may suggest congestion effects (Ali et al. 2007). In other words, β_c for the population density can take positive or negative values across space. The normal distribution can be also used as an initial exploratory analysis to determine the domain of a coefficient. For example, if the estimated parameters are $\hat{\beta} = 2$ and $\hat{\sigma}_\beta = 1$, this implies that approximately $\Phi(\hat{\beta}/\hat{\sigma}_\beta) \cdot 100 \approx 98\%$ of the regions in the sample have a positive coefficient. Therefore, the researcher may be more inclined to choose a distribution with just a positive real domain. One major disadvantage of the normal distribution is that it has infinite tails, which might result in some regions having implausible extreme coefficient values.

Triangular Distribution: This is a continuous probability distribution with probability density function shaped like a triangle. The advantage of this distribution is that it has a definite upper and lower limit, so its tails are shorter than the normal distribution and we avoid extreme coefficients that may result for some regions. The density of a triangular distribution with mean β and spread s_β is zero beyond the range $(\beta - s_\beta, \beta + s_\beta)$, rises linearly from $\beta - s_\beta$ to β, and drops linearly to $\beta + s_\beta$. The parameters $\boldsymbol{\theta} = (\beta, s_\beta)'$ are estimated.

Uniform Distribution: In this case the parameter for each location is equally likely to take on any value in some interval. Suppose that the spread of the uniform distribution is s_β, such that the parameter is uniformly distributed from $\beta - s_\beta$ to $\beta + s_\beta$. Then the parameter can be constructed as $\beta_c = \beta + s_\beta(2u_c - 1)$ where $u_c \sim U[0,1]$ and the parameters $\boldsymbol{\theta} = (\beta, s_\beta)$ are estimated. The new random draw $(2u_c - 1)$ is distributed as $U[-1,+1]$, therefore multiplying by s_β gives a uniformly distributed parameter $\pm$ s (Train 2009, Hensher, Greene 2003). The standard deviation of the uniform distribution can be derived from the spread by dividing s_β

by $\sqrt{3}$. Note also that the uniform distribution with a [0, 1] bound is very suitable when there exists spatial heterogeneity in a dummy variable. For this case the restriction is $\beta = s_\beta = 1/2$.

The normal, triangular and uniform distributions permit positive and negative coefficients. However, as I discussed above, the coefficient may present spatial heterogeneity but only in the positive or negative domain. For example, we may be confident that the coefficient for xk is positive for all regions, but still there may exist spatial heterogeneity around the mean. Some widely used distributions with domain in the positive numbers are the log-normal, truncated normal, and Johnson Sb distribution[5].

Log-normal Distribution: The support of the log-normal distribution is $(0, \infty)$. Formally, the coefficient for each region is specified as $\beta_c = \exp(\beta + \sigma_\beta \eta_c)$ where $\eta_c \sim N(0, 1)$. The parameters β and σ_β, which represent the mean and standard deviation of $\log(\beta_c)$, are estimated. The median, mean, and standard deviation of β_c are $\exp(\beta_c)$, $\exp(\beta_c + \sigma_\beta^2/2)$ and mean $\times\sqrt{\exp(\sigma_\beta^2) - 1}$, respectively (Revelt, Train 1998, Train 2009). The main drawback of the log-normal distribution is that it has a very long right-hand tail. This means that we might find regions with unreasonable extreme positive coefficients.

Truncated Normal Distribution: The domain of this distribution is $(0, \infty)$ if the normal distribution is truncated below at zero. The parameter for each region is created as $\beta_c = max(0, \beta + \sigma_\beta \eta_c)$ where $\eta_c \sim N(0, 1)$ with the share below zero massed at zero equal to $\Phi(-\beta/\sigma_\beta)$. A normal distribution truncated at 0 can be useful when the researcher has a priori belief that for some regions the marginal latent effect of the variable is null. The parameters $\boldsymbol{\theta} = (\beta, \sigma_\beta)$ are estimated.

Johnson S_b Distribution: The S_b distribution gives coefficients between 0 and 1, which is also very suitable for dummy variables. The parameter for region c is computed as $\beta_c = \frac{\exp(\beta + \sigma_\beta \eta_c)}{1 + \exp(\beta + \sigma_\beta \eta_c)}$ where $\eta_c \sim N(0, 1)$ and the parameters β and σ_β are estimated. The mean, variance, and shape are determined by the mean and variance of $\beta + \sigma_\beta \eta_c$ which is a normal distributed parameter. If the analyst needs the coefficient to be between 0 and k, then the variable can be multiplied by k. The logic behind this is the following. Since $\beta_c \times x_{ic}$ ranges between [0, 1], then $\beta_c \times k \times x_{ic}$ is the same as $k[0, 1] = [0, k]$. The advantage of the Johnson S_b is that it can be shaped like log-normal distribution, but with thinner tails below the bound.

For any distribution, all the information about the unobserved spatial heterogeneity is captured by the spread or standard deviation parameter. For example, a significant standard deviation would reveal a spatially non-stationary relationship, and the higher the standard deviation the higher the unobserved spatial heterogeneity in the parameters. Finally, it is worth noting that if only the constant is assumed to be random, then the model is reduced to the random effect model also known as the spatially constant random parameter in the multilevel context (Jones 1991). If $n_c = 1$ for all C, then the model is reduced to the RCM.

2.3 Correlated spatially random parameters and observed variations around the mean

The random parameters can be generalized to include correlation across the parameters. For example, we may be interested in whether regions with greater (lower) β_1 have also greater (lower) values for β_2. If it is true, we would say that both effects are positively correlated within regions. Furthermore, it is likely that the association between y_{ci}^* and x_{ci} is modified by unmeasured regional effects or region-specific unobserved factors. Therefore, by allowing the constant and the slope parameter to be correlated we might be able to identify whether those unobserved factors and the effect of x_{ci} are positively or negatively associated.

[5] If some coefficient is expected a priori to be negative for all the regions, one might create the negative of the variable and then include this new variable in the estimation. This "trick" allows the coefficient to be negative without imposing a sign change in the estimation procedure (Train 2009).

As an illustration of the usefulness of the correlated parameters, Wheeler, Tiefelsdorf (2005) raise the awareness of the potential dependencies (correlation) among the local regression coefficients associated with different exogenous variables in the GWR context. They use a GWR approach to explain the white male bladder cancer mortality rates in the 508 States Economic Areas of the United States. Using the population density and smoking as covariates, they find that those regions with high smoking parameter also have a low population density parameter. As they state, the important question is whether this negative correlation is real or an artifact of the statistical method. By allowing the parameters to be explicitly correlated, we are able to test whether the correlation among the parameters is in fact significant[6].

For simplicity of the notation, consider that the coefficients are distributed across space following a multivariate normal distribution, $\boldsymbol{\beta}_c \sim \text{MVN}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. In this case, the coefficient can be written as:

$$\boldsymbol{\beta}_c = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\eta}_c,$$

where $\boldsymbol{\eta}_c \sim N(\mathbf{0}, \boldsymbol{I})$, and $\boldsymbol{L}$ is the lower-triangular Cholesky factor of $\boldsymbol{\Sigma}$ such that $\boldsymbol{L}\boldsymbol{L}' = \text{var}(\boldsymbol{\beta}_c) = \boldsymbol{\Sigma}$. When the off-diagonal elements of $\boldsymbol{L}$ are zero, the parameters are independently normal distributed. If we assume that the model has only one covariate and the constant, then the extended form of the spatially random coefficient vector is

$$\begin{aligned} \begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix} &= \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \sigma_{\alpha\alpha} & 0 \\ \sigma_{\beta\alpha} & \sigma_{\beta\beta} \end{pmatrix} \begin{pmatrix} \eta_{c\alpha} \\ \eta_{c\beta} \end{pmatrix} \\ \boldsymbol{\beta}_c &= \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\eta}_c, \end{aligned}$$

where:

$$\boldsymbol{L}\boldsymbol{L}' = \begin{pmatrix} \sigma_{\alpha\alpha} & 0 \\ \sigma_{\beta\alpha} & \sigma_{\beta\beta} \end{pmatrix} \begin{pmatrix} \sigma_{\alpha\alpha} & \sigma_{\beta\alpha} \\ 0 & \sigma_{\beta\beta} \end{pmatrix} = \begin{pmatrix} \sigma^2_{\alpha\alpha} & \sigma_{\alpha\alpha}\sigma_{\beta\alpha} \\ \sigma_{\beta\alpha}\sigma_{\alpha\alpha} & \sigma^2_{\beta\alpha} + \sigma^2_{\beta\beta} \end{pmatrix} = \boldsymbol{\Sigma}$$

If we need correlated parameters with positive domain, we might create a log-normal distributed parameter. For instance, if we need β_c to be log-normal distributed, then we can transform it in the following way:

$$\beta_c = \exp(\beta + \sigma_{\beta\alpha}\eta_{c\alpha} + \sigma_{\beta\beta}\eta_{c\beta})$$

Observed spatial heterogeneity – or deterministic spatial heterogeneity – can be also accommodated in the random parameters by including region-specific covariates. Specifically, the vector of random coefficient is:

$$\boldsymbol{\beta}_c = \boldsymbol{\beta} + \boldsymbol{\pi}\boldsymbol{z}_c + \boldsymbol{L}\boldsymbol{\eta}_c \tag{2}$$

where $\boldsymbol{z}_c$ is a set of M characteristics of region c that influences the mean of the spatial random coefficients, and $\boldsymbol{\Pi}$ is a $K \times M$ matrix of additional parameters. The conditional mean becomes $E(\boldsymbol{\beta}_c|\boldsymbol{z}_c) = \boldsymbol{\beta} + \boldsymbol{\Pi}\boldsymbol{z}_c$. The main drawback of this modeling strategy – and any type of spatial heterogeneity in the form of unobserved spatial heterogeneity – is that it assumes that the coefficients are drawn from some univariate or multivariate distribution and no attention is paid to the location of the regions (Fotheringham, Brunsdon 1999). However, the previous model can be very useful to consider regions' location explicitly in the random parameters if $\boldsymbol{z}_c$ includes any function of the geographical coordinates $(\boldsymbol{u}_c, \boldsymbol{v}_c)$. Thus, if $\boldsymbol{z}_c = h(\boldsymbol{u}_c, \boldsymbol{v}_c)$, where $h()$ is any function, and $\boldsymbol{\eta}_c = \mathbf{0}$, then the model collapses into the Casetti's spatial expansion method.

[6] Those readers interested in modelling both spatial dependence and spatial heterogeneity are referred to Dong et al. (2016). They develop a spatial random slope multilevel modeling approach to account for the within-group dependence among individuals in the same area and the spatial dependence between areas simultaneously.

2.4 Estimation

Let $\boldsymbol{y}_c = \{y_{i1}, y_{i2}, \ldots, y_{in}\}$ be the sequence of choices for all individuals in region c, where n_c is the total number of individuals in that region. Assuming that individuals are independent across regions, the joint probability density function, given $\boldsymbol{\beta}_c$, can be written as

$$Pr(\boldsymbol{y}_c|\boldsymbol{X}_c, \boldsymbol{\beta}_c) = \sum_{i=1}^{n_c} f^*(y_{ic}|\boldsymbol{x}_{ic}, \boldsymbol{\beta}_c), \tag{3}$$

because, conditional on $\boldsymbol{\beta}_c$, the observations are independent. Since $\boldsymbol{\beta}_c$ is common for individuals living in the region c, within each region individuals are not independent. Thus, the unconditional pdf of $\boldsymbol{y}_c$ given $\boldsymbol{X}_c$ will be the weighted average of the conditional probability evaluated over all possible values of $\boldsymbol{\beta}$, which depends on the parameters of the distribution of $\boldsymbol{\beta}_c$:

$$P_c(\boldsymbol{\theta}) = f(\boldsymbol{y}_c|\boldsymbol{X}_c, \boldsymbol{\theta}) = \int_{\boldsymbol{\beta}_c} \left[\prod_{i=1}^{N_c} f^*(y_{ic}|\boldsymbol{x}_{ic}, \boldsymbol{\beta}_c, \boldsymbol{\theta}) \right] g(\boldsymbol{\beta}_c) d\boldsymbol{\beta}_c, \tag{4}$$

The unconditional probability has no closed form solution, therefore the log-likelihood function is difficult to compute. However, we can simulate this probability and use the simulated maximum likelihood in order to estimate $\boldsymbol{\theta}$ (Gourieroux, Monfort 1997, Hajivassiliou, Ruud 1986, Stern 1997, Train 2009)[7]. In particular, $P_c(\boldsymbol{\theta})$ is approximated by a summation over randomly chosen values of $\boldsymbol{\beta}_c$. For a given value of the parameters $\boldsymbol{\theta}$, a value of $\boldsymbol{\beta}_c$ is drawn from its distribution. Using this draw of $\boldsymbol{\beta}_c$, $P_c(\boldsymbol{\theta})$ is calculated. This process is repeated for many draws, and the average over the draws is the simulated probability. Formally, the simulated probability for region c is

$$\tilde{P}_c(\boldsymbol{\theta}) = \frac{1}{R} \sum_{r=1}^{R} \prod_{i=1}^{N_c} \tilde{P}_{icr}(\boldsymbol{\theta}) \tag{5}$$

where $\tilde{P}_{icr}$ is the probability for individual i in region c evaluated at the r^{th} draw of $\boldsymbol{\beta}_c$, and R is the total number of draws. Then, the simulated log-likelihood function is:

$$\log L_s = \sum_{c=1}^{C} \log \left[\frac{1}{R} \sum_{r=1}^{R} \prod_{i=1}^{N_c} \tilde{P}_{icr}(\boldsymbol{\theta}) \right] \tag{6}$$

Lee (1992), Gourieroux, Monfort (1991) and Hajivassiliou, Ruud (1986) derive the asymptotic distribution of the simulated maximum likelihood (SML) estimator based on smooth probability simulators with the number of draws increasing with sample size. Under regularity conditions, the estimator is consistent and asymptotically normal. When the number of draws, R, rises faster than the square root of the number of observations, the estimator is asymptotically equivalent to the maximum likelihood estimator. It is worth noting that, even though the simulated probability is an unbiased estimate of the true probability, the log of the simulated probability with fixed number of repetitions is not an unbiased estimate of the log of the true probability. This bias in the SML decreases as the number of draws increases (see for example Gourieroux, Monfort 1997, Revelt, Train 1998).

One main limitation of these modeling strategies is that the performance of the maximum likelihood estimators may not be accurate or satisfactory when the number of individuals per region is large. The problem is that the log-likelihood function involves the integration or summation over a term involving the product of the probabilities for all the individuals in each location c. Borjas, Sueyoshi (1994) were the first in noticing this problem in the context of the probit model with random effects and using Gauss quadrature. Lee (2000) also gives more insights about this problem. For example,

[7] Other methods can be used in order to approximate the integrals. For example, Gauss-Hermite quadrature procedure is another numerical method widely used. However, it has been documented that for models with more than 3 random parameters SML performs better. Bayesian estimation is also suitable for continuous spatial heterogeneity. See for example Hashiguchi, Tanaka (2014).

assuming a sample of 500 individuals per group – or regions in our case – with a likelihood contribution of 0.5 per observation, Borjas, Sueyoshi (1994) show that the value of the integrand can be as small as $\exp(500 \times \ln(0.5)) \approx \exp(-346.6)$, which is below the existing absolute value for a computer. A consequence of this might be larger standard errors, explosive estimates and/or a singular Hessian. In the worst scenario, the computation will overflow, that is, it will exceed the computer's capacity to compute the value and the maximization procedure will stop. This issue should be borne in mind when applying these methods[8].

3 Packages and dependencies

The main R packages used in the examples are the following:

Rchoice: This is the main package to estimate Binary, Poisson, and Ordered Models with Random parameters.

foreign: This package is used to read data in different formats (Stata, SPSS, etc).

car: This package will allow us to perform linear hypotheses.

lmtest: This package has generic functions that allow to perform likelihood ratio tests for nested models.

All these packages can be installed using the `install.packages()` function.

4 Application using Rchoice in R: Self-assessed health status

4.1 Ordered Probit model with spatially homogeneous parameters

Suppose we are interested in the determinants of individuals' subjective evaluation of health. We assume that the health status of individual i in municipality c, h_{ic}, follows an underlying continuous but latent health process h^*_{ic} based on a linear combination of individual and municipal covariates given by:

$$h^*_{ic} = \boldsymbol{x}'_{ic}\boldsymbol{\beta} + \epsilon_{ic} \tag{7}$$

where $\boldsymbol{x}_{ic}$ is a vector of individual and municipal characteristics; $\epsilon_{ic} \sim N(0, \sigma)$ is the error term, but since the scale of h^*_{ic} is not identified, we normalized $\sigma = 1$. Note that this model assumes that the partial correlation between the latent health status and the covariates follows a spatially stationary process.

As typical in ordered models, we do not observe h^*_{ci}, but we instead observe the self-assessed health status (SAH) for each individual, h_{ic}, which ranges between 1 (very bad health) and 5 (very good health) in our sample. The link between h_{ic} and h^*_{ic} is the following:

$$h_{ic} = \begin{cases} 1 & \text{if } \kappa_0 < h^*_{ic} < \kappa_1 \\ 2 & \text{if } \kappa_1 < h^*_{ic} < \kappa_2 \\ \vdots & \\ 5 & \text{if } \kappa_4 < h^*_{ic} < \kappa_5 \end{cases}$$

where it is assumed that $\kappa_0 = -\infty$ and $\kappa_5 = \infty$ to cover the entire real line. Since having a constant is useful in our model to accommodate random effects, we set $x_{1ic} \equiv 1$ for all $i = 1, \ldots, N$. Therefore, for identification we set $\kappa_1 \equiv 0$.

To estimate an ordered probit model with spatially homogeneous coefficients, we will use the `Rchoice` package which is loaded using the `library()` function:

[1]:
```
> # Load package
> library("Rchoice")
```

[8] For other estimation methods, such as Bayesian estimation of multi-level models, see for example Bürkner (2018).

Now, we load the dataset `sah.chile`. This dataset comes from the 2013 National Socioeconomic Characteri-zation Survey (CASEN) from Chile. CASEN is a national, population-based survey which is representative at the municipal level and is carried out by the Ministry of Planning (MIDEPLAN) to describe the socioeconomic situation as well as the impact of social programs on the living conditions of the Chilean Population[9].

In the following lines, the dataset in Stata format is downloaded. Then, the SAH variable (dependent variable) is recoded into 5 categories:

[2]:
```
> # Load data set and recode SAH variable
> library("foreign") # package to load datasets
> library("car") # package with recode function
> data <- read.dta("https://msarrias.weebly.com/uploads/3/7/7/8/37783629/sah.chile.dta")
> data$sah2 <- recode(data$shealth, "1= 1; 2 = 2; 3 = 3; 5:6 = 4; 7 = 5")
```

The vector $\boldsymbol{x}_{ic}$ includes the following controls at the individual level:

linch: log of household income.

agen: age in years / 10.

hsizen: household size / 10.

edun: years of schooling / 10.

male: =1 for men.

dcivil1: =1 if the individual is married.

dlstatus2: =1 if the individual is unemployed.

Some continuous variables are divided by 10 to improve convergence speed of the SML process and avoid singularities in the Hessian matrix.

In addition, a set of dummy variables indicating the self-perception of pollution and environmental problems is used. The dummy variables are obtained from the response to the question: *"What problems related to pollution and environmental degradation do you identify in your neighbourhood or location"*. Based on the answer, dummy variables were created for the following problems:

noise: noise pollution.

airpol: air pollution.

watpol: water pollution.

vispol: visual pollution.

waspol: garbage (rubbish) in the neighborhood.

The variables at the municipality level are:

lmdinc: log of median income (proxy for development).

lpop: log of population (size effect).

The following command lines show how to estimate the traditional ordered probit model. For other models such as the Binary (Logit and Probit) and Poisson model see Sarrias (2016)Sarrias (2016).

[3]:
```
> # Ordered probit model
> oprobit <- Rchoice(sah2 ~ linch + agen + hsizen + edun + male +
+   dcivil1 + dlstatus2 +
+        noise + airpol + watpol + vispol + waspol +
+        lmdinc + lpop,
+        family = ordinal("probit"),
+        data = data)
> summary(oprobit)
```

[9]Chile has 346 municipalities of which 324 are representative in CASEN 2013.

[3]:
```
##
## Model: ordinal
## Model estimated on: Tue Jan 07 10:31:58 2020
##
## Call:
## Rchoice(formula = sah2 ~ linch + agen + hsizen + edun + male +
## dcivil1 + dlstatus2 + noise + airpol + watpol + vispol +
## waspol + lmdinc + lpop, data = data, family = ordinal("probit"),
## method = "bfgs")
##
##
## Frequencies of categories:
## y
##       1       2       3       4       5
## 0.01087 0.01359 0.02897 0.64523 0.30133
## The estimation took: 0h:0m:5s
##
## Coefficients:
##             Estimate Std. Error z-value Pr(>|z|)
## kappa.1     0.341698   0.022434  15.231  < 2e-16 ***
## kappa.2     0.720961   0.027012  26.690  < 2e-16 ***
## kappa.3     3.015915   0.031564  95.548  < 2e-16 ***
## constant   -0.290407   0.481404  -0.603 0.546342
## linch       0.132595   0.014377   9.223  < 2e-16 ***
## agen       -0.212654   0.008214 -25.889  < 2e-16 ***
## hsizen      0.242645   0.060359   4.020 5.82e-05 ***
## edun        0.211808   0.028537   7.422 1.15e-13 ***
## male        0.154095   0.019058   8.086 6.66e-16 ***
## dcivil1    -0.028142   0.021246  -1.325 0.185316
## dlstatus2 -0.093728   0.046169  -2.030 0.042347 *
## noise      -0.139891   0.026745  -5.231 1.69e-07 ***
## airpol     -0.084563   0.026219  -3.225 0.001259 **
## watpol     -0.120485   0.037756  -3.191 0.001417 **
## vispol     -0.069064   0.060747  -1.137 0.255573
## waspol     -0.041040   0.027421  -1.497 0.134482
## lmdinc      0.147852   0.043400   3.407 0.000658 ***
## lpop       -0.016493   0.009016  -1.829 0.067359 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Optimization of log-likelihood by BFGS maximization
## Log Likelihood: -13020
## Number of observations: 16188
## Number of iterations: 190
## Exit of MLE: successful convergence
```

The argument `family = ordinal("probit")` indicates that an ordered probit model will be estimated. If the user wants an ordered logit model the argument should be `family = ordinal("logit")`. For other models, see `help(Rchoice)`.

The results show that household income and education increase the probability of reporting better health status, whereas age decreases it. Men are more likely to report better health than women and being unemployed is detrimental for health. At the neighborhood level, noise, air, and water pollution reduce health perception and vispol and waspol apparently do not matter for health. The coefficient for the logarithm of population for each municipality, which is intended to capture agglomeration effects, is negative but weakly significant, whereas the level of development is positively correlated with individuals' health evaluation.

4.2 Ordered Probit models with spatial random coefficients

The standard ordered probit model does not allow for spatial heterogeneity in the coefficients. In this section, we estimate an Ordered Probit with Spatial Random Parameters (OPSRP) model. To reduce excessive computing time, we will only assume that the variables at the level of municipalities and neighborhood vary across space.

The first and more difficult task is to choose the distribution for each of them. As explained by Hensher, Greene (2003), distributions are essentially arbitrary approximations to the real behavioral profile. The researcher chooses a specific distribution because he has

a sense that the "empirical truth" is somewhere in their domain. The most widely used distribution in the empirical literature is the normal distribution due to its properties. If unobserved spatial heterogeneity is viewed as the sum of small random influences, then the central limit theorem can be invoked to justify the normality assumption Greene, Hensher (2010). Moreover, the normal distribution is unbounded, and therefore every real number has a positive probability of being drawn. Thus, specifying a given coefficient to follow a normal distribution is equivalent to making a priori assumption that both positive and negative coefficients exits across space (Sarrias 2019).

This last property is very appealing in our case, since theoretically we might observe municipalities with positive and negative sign for the population coefficient. For instance, municipalities with a positive coefficient might be characterized for having positive urban externalities that outweigh the negative ones. In those municipalities, inhabitants, on average, enjoy better health through local positive urban externalities. If the coefficient is negative, the opposite might be expected.

A OPSRP model with normally distributed parameters is estimated as follows:

[4]:
```
> # Spatial random parameter model
> ran_1 <- Rchoice(sah2 ~ linch + agen + hsizen + edun + male + dcivil1 + dlstatus2 +
+        noise + airpol + watpol + vispol + waspol +
+        lmdinc + lpop,
+        family = ordinal('probit'),
+        data = data,
+        ranp = c(noise = "n", airpol = "n", watpol = "n", vispol = "n",
+        waspol = "n", lmdinc = "n", lpop = "n"),
+        panel = TRUE,
+        index = "idc",
+        R=30,
+        method = "bfgs")
> summary(ran_1)
```

[4]:
```
##
## Model: ordinal
## Model estimated on: Tue Dec 31 09:11:29 2019
##
## Call:
## Rchoice(formula = sah2 ~ linch + agen + hsizen + edun + male +
## dcivil1 + dlstatus2 + noise + airpol + watpol + vispol +
## waspol + lmdinc + lpop, data = data, family = ordinal("probit"),
## ranp = c(noise = "n", airpol = "n", watpol = "n", vispol = "n",
## waspol = "n", lmdinc = "n", lpop = "n"), R = 30, panel = TRUE,
## index = "idc", method = "bfgs", iterlim = 2000)
##
##
## Frequencies of categories:
## y
##       1       2       3       4       5
## 0.01087 0.01359 0.02897 0.64523 0.30133
## The estimation took: 0h:9m:48s
##
## Coefficients:
##               Estimate Std. Error z-value Pr(>|z|)
## kappa.1      0.3417117  0.0162495  21.029  < 2e-16 ***
## kappa.2      0.7209893  0.0209411  34.429  < 2e-16 ***
## kappa.3      3.0170292  0.0268859 112.216  < 2e-16 ***
## constant    -0.2905344  0.4958543  -0.586 0.557925
## linch        0.1310569  0.0143865   9.110  < 2e-16 ***
## agen        -0.2132221  0.0082107 -25.969  < 2e-16 ***
## hsizen       0.2425766  0.0604009   4.016 5.92e-05 ***
## edun         0.2116421  0.0285525   7.412 1.24e-13 ***
## male         0.1540426  0.0190620   8.081 6.66e-16 ***
## dcivil1     -0.0281834  0.0212486  -1.326 0.184717
## dlstatus2   -0.0937187  0.0462089  -2.028 0.042545 *
## mean.noise  -0.1399619  0.0269286  -5.198 2.02e-07 ***
## mean.airpol -0.0845976  0.0264687  -3.196 0.001393 **
## mean.watpol -0.1205199  0.0379779  -3.173 0.001507 **
## mean.vispol -0.0690556  0.0611251  -1.130 0.258585
## mean.waspol -0.0410117  0.0276460  -1.483 0.137953
## mean.lmdinc  0.1463353  0.0445929   3.282 0.001032 **
## mean.lpop   -0.0180293  0.0091612  -1.968 0.049067 *
```

```
## sd.noise     0.0986997  0.0261224   3.778 0.000158 ***
## sd.airpol    0.0986787  0.0254373   3.879 0.000105 ***
## sd.watpol    0.0994208  0.0398379   2.496 0.012573 *
## sd.vispol    0.0998332  0.0688382   1.450 0.146987
## sd.waspol    0.0988309  0.0273009   3.620 0.000295 ***
## sd.lmdinc    0.0050337  0.0008714   5.776 7.64e-09 ***
## sd.lpop      0.0012177  0.0011066   1.100 0.271156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Optimization of log-likelihood by BFGS maximization
## Log Likelihood: -11320
## Number of observations: 16188
## Number of iterations: 391
## Exit of MLE: successful convergence
## Simulation based on 30 Halton draws
```

The argument `ranp` indicates which variables are random in the formula and their distributions. In this example, all the random variables are assumed to be normally distributed using `"n"`. The remaining distribution discussed in Section 2.2 can be used using the following shorthands:

- Triangular = `"t"`,
- Uniform = `"u"`,
- Truncated normal = `"cn"`,
- Log-normal = `"ln"`,
- Johnson's Sb = = `"sb"`.

The number of draws for the simulation of the probability is set using the argument `R`. To keep the estimation time manageable, we use 30 draws for each individual. However, consistency of the SML requires a higher number of draws (see for example Train 2009).

The argument `index` is a string indicating the id for the municipalities in the data, whereas `panel=TRUE` allows for the spatial structure of the sample.

The previous model assumes that the coefficients has the following form:

$$\beta_k = \bar{\beta}_k + \sigma_k v_{ir}$$

where $v_{ir} \sim N(01)$. Thus, the coefficients with the `mean.` and `sd.` prefix represent the estimated mean, $\hat{\bar{\beta}}$, and standard deviation, $\hat{\sigma}$, for variable k, respectively. If $\sigma_k = 0$, then there is no evidence of systematical variation for regression coefficient over space. The output shows that there is evidence of spatial heterogeneity for most of the variables, except for `vispol` and `lpop`.

To test the joint hypothesis of coefficient homogeneity across space we can perform a Likelihood Ratio test using `lrtest` function from `lmtest` package.

[5]:
```
> # Testing spatial heterogeneity
> library("lmtest")
> lrtest(oprobit, ran_1)
```

[5]:
```
## Likelihood ratio test
##
## Model 1: sah2 ~ linch + agen + hsizen + edun + male + dcivil1 + dlstatus2 +
## noise + airpol + watpol + vispol + waspol + lmdinc + lpop
## Model 2: sah2 ~ linch + agen + hsizen + edun + male + dcivil1 + dlstatus2 +
## noise + airpol + watpol + vispol + waspol + lmdinc + lpop
## #Df LogLik Df  Chisq Pr(>Chisq)
## 1  18 -13018
## 2  25 -11318  7 3400.6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test rejects the null hypothesis providing empirical evidence of spatial heterogeneity for those variables.

Since the parameters are allowed to vary across space following a normal distribution, we can also compute the proportion of municipalities with positive coefficients using $\Phi(\hat{\beta}/\hat{\sigma})$. For example, for `noise` and `lmdinc` the results are:

[6]:
```
> # Computing proportions
> pnorm(coef(ran_1)["mean.noise"] / coef(ran_1)["sd.noise"])
```

[6]:
```
## mean.noise
## 0.07808686
```

[7]:
```
> pnorm(coef(ran_1)["mean.lmdinc"] / coef(ran_1)["sd.lmdinc"])
```

[7]:
```
## mean.lmdinc
##           1
```

Thus, we can say that for 100% of the municipalities development is positively correlated with individuals' health, whereas for around 8% of the municipalities, higher noise pollution increases health. This last result can be true or an artifact of the normality assumption.

4.3 Correlated parameters

The previous model specifies the coefficients to be independently distributed, while one would expect correlation. To show this, the model `ran_1` we will be estimated but assuming that the spatially random coefficients are correlated adding the argument `correlation = TRUE`:

[8]:
```
> # Spatially random parameters with correlated coefficients
> ran_2 <- Rchoice(sah2 ~ linch + agen + hsizen + edun + male + dcivil1 +
+        dlstatus2 + noise + airpol + watpol + vispol + waspol +
+        lmdinc + lpop,
+        family = ordinal("probit"),
+        data = data,
+        ranp = c(noise = "n", airpol = "n", watpol = "n", vispol = "n",
+        waspol = "n", lmdinc = "n", lpop = "n"),
+        panel = TRUE,
+        index = "idc",
+        R=30,
+        method = "bfgs",
+        correlation = TRUE)
> summary(ran_2)
```

[8]:
```
##
## Model: ordinal
## Model estimated on: Tue Dec 31 09:21:06 2019
##
## Call:
## Rchoice(formula = sah2 ~ linch + agen + hsizen + edun + male +
## dcivil1 + dlstatus2 + noise + airpol + watpol + vispol +
## waspol + lmdinc + lpop, data = data, family = ordinal("probit"),
## ranp = c(noise = "n", airpol = "n", watpol = "n", vispol = "n",
## waspol = "n", lmdinc = "n", lpop = "n"), R = 30, correlation = TRUE,
## panel = TRUE, index = "idc", method = "bfgs", iterlim = 2000)
##
##
## Frequencies of categories:
## y
##       1       2       3       4       5
## 0.01087 0.01359 0.02897 0.64523 0.30133
## The estimation took: 0h:9m:36s
##
## Coefficients:
##                    Estimate Std. Error z-value Pr(>|z|)
## kappa.1           0.3412626  0.0180952  18.859  < 2e-16 ***
## kappa.2           0.7195420  0.0226625  31.750  < 2e-16 ***
## kappa.3           3.4970489  0.0294076 118.917  < 2e-16 ***
## constant         -0.2901421  0.5322664  -0.545 0.585680
## linch             0.1457106  0.0149581   9.741  < 2e-16 ***
## agen             -0.2696671  0.0086621 -31.132  < 2e-16 ***
## hsizen            0.2440816  0.0629962   3.875 0.000107 ***
## edun              0.2202293  0.0300012   7.341 2.12e-13 ***
```

```
## male              0.1579018  0.0199386   7.919 2.44e-15 ***
## dcivil1          -0.0361313  0.0222373  -1.625 0.104204
## dlstatus2        -0.0930049  0.0482324  -1.928 0.053822 .
## mean.noise       -0.1446350  0.0281478  -5.138 2.77e-07 ***
## mean.airpol      -0.0855360  0.0275906  -3.100 0.001934 **
## mean.watpol      -0.1204861  0.0399076  -3.019 0.002535 **
## mean.vispol      -0.0689847  0.0636902  -1.083 0.278751
## mean.waspol      -0.0371335  0.0288215  -1.288 0.197609
## mean.lmdinc       0.1607965  0.0478379   3.361 0.000776 ***
## mean.lpop         0.0069806  0.0101792   0.686 0.492860
## sd.noise.noise    0.0050271  0.0333140   0.151 0.880055
## sd.noise.airpol   0.0058448  0.0324393   0.180 0.857013
## sd.noise.watpol   0.0539982  0.0478928   1.127 0.259539
## sd.noise.vispol   0.0739318  0.0776348   0.952 0.340943
## sd.noise.waspol   0.0182864  0.0331744   0.551 0.581483
## sd.noise.lmdinc   0.0618425  0.0086254   7.170 7.51e-13 ***
## sd.noise.lpop    -0.0813180  0.0105174  -7.732 1.07e-14 ***
## sd.airpol.airpol  0.0160025  0.0319875   0.500 0.616881
## sd.airpol.watpol  0.0559519  0.0459907   1.217 0.223760
## sd.airpol.vispol  0.0816862  0.0775592   1.053 0.292245
## sd.airpol.waspol  0.0208368  0.0330971   0.630 0.528979
## sd.airpol.lmdinc  0.0434195  0.0084910   5.114 3.16e-07 ***
## sd.airpol.lpop   -0.0588680  0.0102640  -5.735 9.73e-09 ***
## sd.watpol.watpol  0.0628180  0.0456531   1.376 0.168826
## sd.watpol.vispol  0.0818528  0.0779919   1.050 0.293946
## sd.watpol.waspol  0.0233655  0.0330245   0.708 0.479244
## sd.watpol.lmdinc  0.0408594  0.0085428   4.783 1.73e-06 ***
## sd.watpol.lpop   -0.0505877  0.0103099  -4.907 9.26e-07 ***
## sd.vispol.vispol  0.0850337  0.0779709   1.091 0.275457
## sd.vispol.waspol  0.0187284  0.0330837   0.566 0.571331
## sd.vispol.lmdinc  0.0471032  0.0083197   5.662 1.50e-08 ***
## sd.vispol.lpop   -0.0644627  0.0100842  -6.392 1.63e-10 ***
## sd.waspol.waspol  0.0224287  0.0329135   0.681 0.495591
## sd.waspol.lmdinc  0.0491523  0.0082301   5.972 2.34e-09 ***
## sd.waspol.lpop   -0.0642418  0.0099655  -6.446 1.15e-10 ***
## sd.lmdinc.lmdinc  0.0353881  0.0082862   4.271 1.95e-05 ***
## sd.lmdinc.lpop   -0.0450126  0.0099427  -4.527 5.98e-06 ***
## sd.lpop.lpop      0.0002568  0.0011373   0.226 0.821375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Optimization of log-likelihood by BFGS maximization
## Log Likelihood: -11210
## Number of observations: 16188
## Number of iterations: 315
## Exit of MLE: successful convergence
## Simulation based on 30 Halton draws
```

It is important to note that the output prints the elements of the lower-triangular Cholesky factor $\boldsymbol{L}$. The variance-covariance matrix, $\boldsymbol{\Sigma}$, can be extracted using the `vcov` function in the following way:

[9]:
```
> # Obtain Sigma
> vcov(ran_2, what = "ranp", type = "cov", se = TRUE)
```

[9]:
```
##
## Elements of the variance-covariance matrix
##
##                    Estimate Std. Error z-value  Pr(>|z|)
## v.noise.noise    2.5271e-05 3.3494e-04  0.0754   0.93986
## v.noise.airpol   2.9382e-05 2.3810e-04  0.1234   0.90179
## v.noise.watpol   2.7145e-04 1.7934e-03  0.1514   0.87969
## v.noise.vispol   3.7166e-04 2.4696e-03  0.1505   0.88037
## v.noise.waspol   9.1927e-05 6.3478e-04  0.1448   0.88486
## v.noise.lmdinc   3.1089e-04 2.0673e-03  0.1504   0.88046
## v.noise.lpop    -4.0879e-04 2.7200e-03 -0.1503   0.88053
## v.airpol.airpol  2.9024e-04 1.0858e-03  0.2673   0.78923
## v.airpol.watpol  1.2110e-03 2.4984e-03  0.4847   0.62788
## v.airpol.vispol  1.7393e-03 3.7015e-03  0.4699   0.63843
## v.airpol.waspol  4.4032e-04 1.0616e-03  0.4148   0.67829
## v.airpol.lmdinc  1.0563e-03 2.4234e-03  0.4359   0.66293
## v.airpol.lpop   -1.4173e-03 3.2248e-03 -0.4395   0.66029
```

```
## v.watpol.watpol  9.9925e-03 9.2513e-03  1.0801   0.28009
## v.watpol.vispol  1.3705e-02 9.7826e-03  1.4009   0.16124
## v.watpol.waspol  3.6211e-03 3.6712e-03  0.9863   0.32397
## v.watpol.lmdinc  8.3355e-03 4.0026e-03  2.0825   0.03730 *
## v.watpol.lpop   -1.0863e-02 5.2619e-03 -2.0644   0.03898 *
## v.vispol.vispol  2.6069e-02 2.4122e-02  1.0807   0.27982
## v.vispol.waspol  6.5591e-03 5.8208e-03  1.1268   0.25981
## v.vispol.lmdinc  1.5469e-02 7.5381e-03  2.0521   0.04016 *
## v.vispol.lpop   -2.0443e-02 9.9442e-03 -2.0558   0.03981 *
## v.waspol.waspol  2.1683e-03 2.9746e-03  0.7289   0.46604
## v.waspol.lmdinc  4.9749e-03 3.5556e-03  1.3992   0.16176
## v.waspol.lpop   -6.5438e-03 4.7027e-03 -1.3915   0.16407
## v.lmdinc.lmdinc  1.3266e-02 1.8869e-03  7.0309 2.052e-12 ***
## v.lmdinc.lpop   -1.7439e-02 2.3778e-03 -7.3341 2.232e-13 ***
## v.lpop.lpop      2.2946e-02 3.0048e-03  7.6365 2.243e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated coefficients represent the variance and covariance of the randomly distributed parameters. Their standard errors are estimated using the Delta Method. To obtain the standard deviations for the random parameters, one might use the following code:

[10]:
```
> # Obtain standard deviations
> vcov(ran_2, what = "ranp", type = "sd", se = TRUE)
```

[10]:
```
##
## Standard deviations of the random parameters
##
## Estimate Std. Error z-value Pr(>|z|)
## noise  0.0050271  0.0333140  0.1509  0.88006
## airpol 0.0170365  0.0318660  0.5346  0.59291
## watpol 0.0999627  0.0462737  2.1602  0.03075 *
## vispol 0.1614594  0.0746999  2.1614  0.03066 *
## waspol 0.0465651  0.0319403  1.4579  0.14487
## lmdinc 0.1151790  0.0081909 14.0617  < 2e-16 ***
## lpop        0.1514789        0.0099181 15.2730        < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, the correlation matrix for the estimated coefficients is:

[11]:
```
> # Obtain correlation matrix of estimated coefficients
> vcov(ran_2, what = "ranp", type = "cor")
```

[11]:
```
##             noise     airpol     watpol     vispol     waspol     lmdinc
## noise   1.0000000  0.3430771  0.5401842  0.4578971  0.3927053  0.5369247
## airpol  0.3430771  1.0000000  0.7110814  0.6323119  0.5550469  0.5383007
## watpol  0.5401842  0.7110814  1.0000000  0.8491072  0.7779253  0.7239690
## vispol  0.4578971  0.6323119  0.8491072  1.0000000  0.8724090  0.8317965
## waspol  0.3927053  0.5550469  0.7779253  0.8724090  1.0000000  0.9275751
## lmdinc  0.5369247  0.5383007  0.7239690  0.8317965  0.9275751  1.0000000
## lpop   -0.5368276 -0.5492089 -0.7173734 -0.8358492 -0.9277189 -0.9995225

##              lpop
## noise  -0.5368276
## airpol -0.5492089
## watpol -0.7173734
## vispol -0.8358492
## waspol -0.9277189
## lmdinc -0.9995225
## lpop    1.0000000
```

The results show, for example, that noise pollution is positively correlated with other forms of pollution. Therefore, in those municipalities where noise pollution is detrimental to health, so are the other forms of pollution. It is also important to note that the municipalities where noise has a negative effect are also municipalities where lower development and higher population impact negatively the self-perception of health status.

4.4 *Region-specific coefficients*

In the applied literature it is very common to map the region-specific estimates to display the spatial heterogeneity for certain coefficients. This cannot be done using just the distribution of the parameters across regions, $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$. The population distributions give us just the average affect, $\boldsymbol{\beta}$, and the spatial variation around this mean, $\sigma_{\boldsymbol{\beta}}$, when in fact we would like to know where each region's $\boldsymbol{\beta}_c$ lies in $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$. We might be able to find the likely location of a given region on the heterogeneity distribution by moving from the conditional to the unconditional distribution (Revelt, Train 2000, Brunsdon et al. 1999, Sarrias 2019). Using Bayes' theorem, we obtain:

$$f(\boldsymbol{\beta}_c|\boldsymbol{y}_c,\boldsymbol{X}_c,\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}_c|\boldsymbol{X}_c,\boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})}{f(\boldsymbol{y}_c|\boldsymbol{X}_c,\boldsymbol{\theta})} = \frac{f(\boldsymbol{y}_c|\boldsymbol{X}_c,\boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})}{\int_{\boldsymbol{\beta}_c} f(\boldsymbol{y}_c|\boldsymbol{X}_c,\boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c} \tag{8}$$

where $f(\boldsymbol{\beta}_c|\boldsymbol{y}_c,\boldsymbol{X}_c,\boldsymbol{\theta})$ is the distribution of the regional parameters $\boldsymbol{\beta}_c$ conditional on the sequence of choices of all the individuals in region c, whereas $g(\boldsymbol{\beta}_c|\boldsymbol{\theta})$ is the unconditional distribution. The conditional expectation of $\boldsymbol{\beta}_c$ is given by

$$\bar{\boldsymbol{\beta}}_c = E\left[\boldsymbol{\beta}_c|\boldsymbol{y}_c,\boldsymbol{X}_c,\boldsymbol{\theta}\right] = \frac{\int_{\boldsymbol{\beta}_c}\boldsymbol{\beta}_c f(\boldsymbol{y}_c|\boldsymbol{X}_c,\boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c}{\int_{\boldsymbol{\beta}_c} f(\boldsymbol{y}_c|\boldsymbol{X}_c,\boldsymbol{\beta}_c)g(\boldsymbol{\beta}_c|\boldsymbol{\theta})d\boldsymbol{\beta}_c} \tag{9}$$

This expectation gives us the conditional mean of the distribution of the spatially random parameter. The simulator for this expectation is:

$$\hat{\bar{\boldsymbol{\beta}}}_c = \hat{E}\left[\boldsymbol{\beta}_c|\boldsymbol{y}_c,\boldsymbol{X}_c,\hat{\boldsymbol{\theta}}\right] = \frac{\frac{1}{R}\sum_{r=1}^{R}\hat{\boldsymbol{\beta}}_{cr}\prod_{i=1}^{n_c} f^*(y_{ci}|\boldsymbol{x}_{ci},\hat{\boldsymbol{\beta}}_{cr})}{\frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{n_c} f^*(y_{ci}|\boldsymbol{x}_{ci},\hat{\boldsymbol{\beta}}_{cr})} \tag{10}$$

This estimator is the region-specific estimate, and can be computed in `Rchoice` using `effect.Rchoice` function and plotted using the function `plot`. In the following lines the municipality's coefficient for all the random parameters is plotted using a kernel approximation:

[12]:
```
> # Plot municipality-specific coefficient
> par(mfrow = c(3, 3))
> plot(ran_2, par = "noise",  type = "density", main = "Noise Pol.")
> plot(ran_2, par = "airpol", type = "density", main = "Air Pol.")
> plot(ran_2, par = "watpol", type = "density", main = "Water Pol.")
> plot(ran_2, par = "vispol", type = "density", main = "Visual Pol.")
> plot(ran_2, par = "waspol", type = "density", main = "Garbage Pol.")
> plot(ran_2, par = "lmdinc", type = "density", main = "Development")
> plot(ran_2, par = "lpop",   type = "density", main = "Population")
```

[12]:
```
Output reproduced in Figure 1
```

The red area under the kernel distribution illustrates the proportion of municipalities with a positive conditional mean. The most relevant result is that size (`lpop`) seems to be a positive externality for almost 50% of the municipalities, evidencing substantial spatial heterogeneity. This important result is obscured by the traditional ordered probit model.

We might also plot the 95% confidence interval for the conditional means of airpol and noise for the first 50 municipalities by typing:

[13]:
```
> # Plot region-specific confidence intervals.
> par(mfrow = c(1, 2))
> plot(ran_2, par = "airpol", ind =  TRUE, id = 1:50, ylab = "Municipalities")
> plot(ran_2, par = "noise", ind =  TRUE, id = 1:50, ylab = "Municipalities")
```

[13]:
```
Output reproduced in Figure 2
```

In terms of consistency of the regional-specific estimates, it is expected that $\hat{\bar{\boldsymbol{\beta}}}_c \xrightarrow{p} \boldsymbol{\beta}_c$ as $n_c \to \infty$. That is, if we have more information about the choices made by the individuals in each region, then we are in better position to identify where each region coefficient lies on $g(\boldsymbol{\beta}_c)$ (see for example Train 2009, Revelt, Train 2000, Sarrias, Daziano 2018).

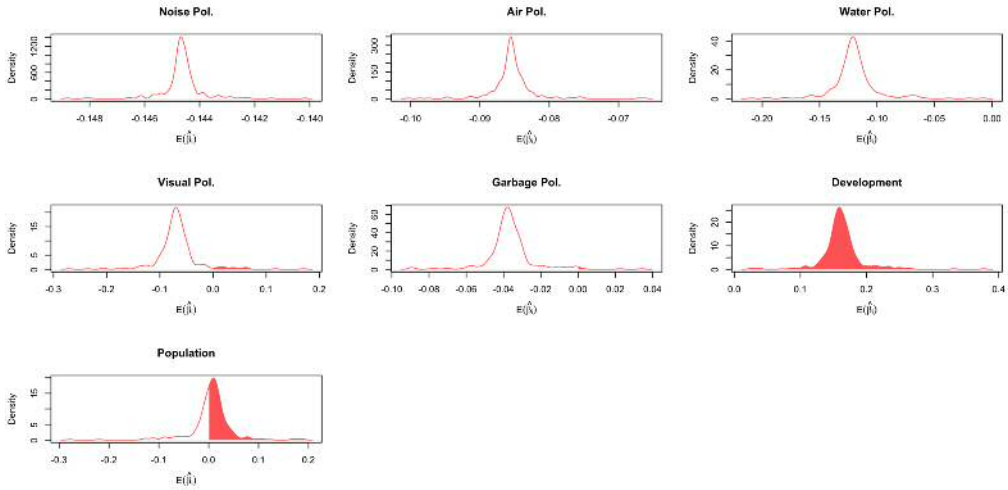


Figure 1: Plot of municipality-specific coefficients

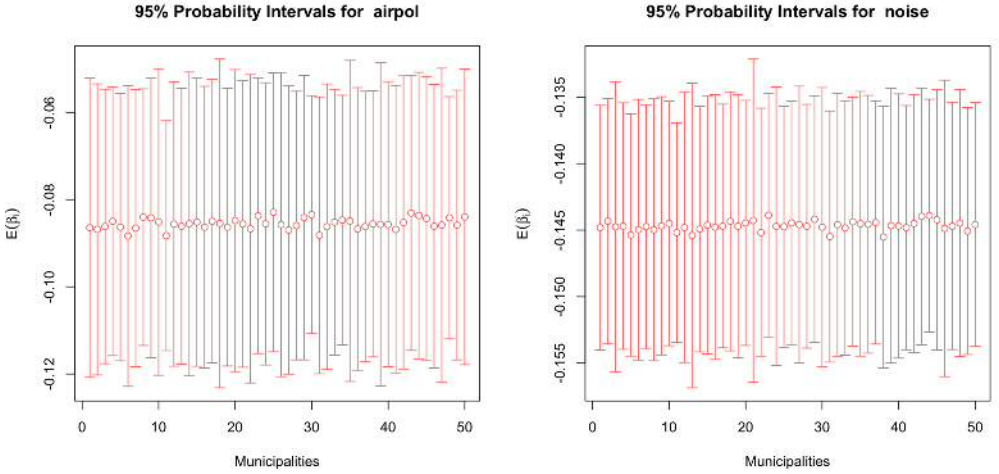


Figure 2: Plot of region-specific confidence intervals

5 Conclusion

This paper contributes to the literature of spatial econometric models that deal with spatially non-stationary process by assuming unobserved heterogeneity. This modelling approach has been widely used in discrete choice modeling, but it can also be implemented to capture and model observed and unobserved spatial heterogeneity. One of the main advantages of this modelling approach is that allows the analyst to include variables at the individual level, which mitigate the ecological fallacy problem, and to add more flexibility regarding the shape and boundedness of the coefficients.

Spatial heterogeneity is represented by some distribution $g(\boldsymbol{\beta}_c)$, which can take any continuous shape, and the analyst must choose the distribution a priori. The choice of the distribution may be guided by theoretical reasons regarding the domain and bound of the coefficients. It also discussed some extensions that can be useful in order to take into consideration the geographical location of the regions, as well as the spatial correlation of the parameters. Although the unobserved spatial heterogeneity using continuous distributions has very appealing features, the probability for each region does not have a closed form solution. Therefore, we need to simulate this probability and estimate the parameters using SML, which can be very costly in terms of computational time.

This study also shows how the `Rchoice` package can be used to estimate this type of models. To do so, we provide a simple example for ordered probit models, focusing on how the determinants of individuals' self-assessed health status might vary across space.

This work can be extended in different ways. First, one of the main concerns and limitations of the model is that the estimation requires computing the product of the probabilities for all individuals in a given region. Thus, if the number of individuals is too high, the estimation method may run into numerical difficulties. To overcome this problem some of the two methods proposed by Lee (2000) can be studied under the spatial context. These methods alleviate the numerical problems by interchanging the inner product with the outer summation. Another possible extension is to study the behavior of the parameters with small and large samples using Bayesian and EM algorithms. Finally, more empirical applications are needed in order to understand the strengths and weaknesses for estimating models with locally varying coefficients using unobserved heterogeneity.

References

Ali K, Partridge MD, Olfert MR (2007) Can geographically weighted regressions improve regional analysis and policy making? *International Regional Science Review* 30[3]: 300–329. CrossRef.

Anselin L (1988) *Spatial Econometrics: Methods and Models*, Volume 4. Springer. CrossRef.

Borjas GJ, Sueyoshi GT (1994) A two-stage estimator for probit models with structural group effects. *Journal of Econometrics* 64[1]: 165–182. CrossRef.

Bürkner P (2018) Advanced bayesian multilevel modeling with the r package brms. *R Journal* 10[1]. CrossRef.

Brunsdon C, Aitkin M, Fotheringham S, Charlton M (1999) A comparison of random coefficient modelling and geographically weighted regression for spatially non-stationary regression problems. *Geographical and Environmental Modelling* 3: 47–62

Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (the Statistician)* 47[3]: 431–443. CrossRef.

Casetti E (1972) Generating models by the expansion method: Applications to geographical research. *Geographical Analysis* 4[1]: 81–91. CrossRef.

Dong G, Harris R, Jones K, Yu J (2015) Multilevel modelling with spatial interaction effects with application to an emerging land market in beijing, china. *PloS One* 10[6]: e0130761. CrossRef.

Dong G, Ma J, Harris R, Pryce G (2016) Spatial random slope multilevel modeling using multivariate conditional autoregressive models: A case study of subjective travel satisfaction in beijing. *Annals of the American Association of Geographers* 106[1]: 19–35. CrossRef.

Dong G, Nakaya T, Brunsdon C (2018) Geographically weighted regression models for ordinal categorical response variables: An application to geo-referenced life satisfaction data. *Computers, Environment and Urban Systems* 70: 35–42. CrossRef.

Duranton G, Puga D (2004) Micro-foundations of urban agglomeration economies. In: Henderson V, Thisse J (eds), *Handbook of Regional and Urban Economics*, Volume 4. Elsevier, 2063–2117. CrossRef.

Elhorst JP (2014) *Spatial Panel Data Models*. Springer. CrossRef.

Finley AO (2011) Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution* 2[2]: 143–154. CrossRef.

Fotheringham AS, Brunsdon C (1999) Local forms of spatial analysis. *Geographical Analysis* 31[4]: 340–358. CrossRef.

Gelfand AE, Kim HJ, Sirmans CF, Banerjee S (2003) Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association* 98[462]: 387–396. CrossRef.

Goldstein H (1987) *Multilevel Models in Education and Social Research*. Oxford University Press

Gourieroux C, Monfort A (1991) Simulation based inference in models with heterogeneity. *Annales d'Economie et de Statistique* 20-21: 69–107. CrossRef.

Gourieroux C, Monfort A (1997) *Simulation-Based Econometric Methods*. Oxford University Press. CrossRef.

Greene WH, Hensher DA (2010) *Modeling Ordered Choices: A Primer*. Cambridge University Press. CrossRef.

Hajivassiliou VA, Ruud PA (1986) Classical estimation methods for ldv models using simulation. In: Engle R, McFadden D (eds), *Handbook of Econometrics*, Volume 4. Elsevier. CrossRef.

Hashiguchi Y, Tanaka K (2014) Agglomeration and firm-level productivity: A bayesian spatial approach. *Papers in Regional Science* 94[S1]: S95–S114. CrossRef.

Hensher D, Greene WH (2003) The mixed logit model: The state of practice. *Transportation* 30[2]: 133–176. CrossRef.

Jetz W, Rahbek C, Lichstein JW (2005) Local and global approaches to spatial data analysis in ecology. *Global Ecology and Biogeography* 14[1]: 97–98. CrossRef.

Jones K (1991) Specifying and estimating multi-level models for geographical research. *Transactions of the Institute of British Geographers* 16: 148–159. CrossRef.

Kelejian HH, Prucha IR (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics* 17[1]: 99–121

Kelejian HH, Prucha IR (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40[2]: 509–533. CrossRef.

Lee L (2000) A numerically stable quadrature procedure for the one-factor random-component discrete choice model. *Journal of Econometrics* 95[1]: 117–129. CrossRef.

Lee LF (1992) On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models. *Econometric Theory* 8[4]: 518–552. CrossRef.

Revelt D, Train K (1998) Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics* 80[4]: 647–657. CrossRef.

Revelt D, Train K (2000) Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier. Working Paper. Department of Economics, UCB

Robinson WS (1950) Ecological correlations and the behavior of individuals. *American Sociological Review* 15[3]: 351–357. CrossRef.

Sarrias M (2016) Discrete choice models with random parameters in R: The Rchoice package. *Journal of Statistical Software* 74[10]: 1–31. CrossRef.

Sarrias M (2019) Do monetary subjective well-being evaluations vary across space? comparing continuous and discrete spatial heterogeneity. *Spatial Economic Analysis* 14[1]: 53–87. CrossRef.

Sarrias M, Daziano RA (2018) Individual-specific point and interval conditional estimates of latent class logit parameters. *Journal of Choice Modelling* 27: 50–61. CrossRef.

Stern S (1997) Simulation-based estimation. *Journal of Economic Literature* 35[4]: 2006–2039

Swamy PAVB (1971) *Statistical Inference in Random Coefficient Regression Models*. Springer Berlin. CrossRef.

Train K (2009) *Discrete Choice Methods with Simulation*. Cambridge University Press. CrossRef.

Wheeler D, Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* 7[2]: 161–187. CrossRef.

Withers SD (2001) Quantitative methods: Advancement in ecological inference. *Progress in Human Geography* 25[1]: 87–96. CrossRef.

REGION
The Journal of ERSA
Powered by WU

Volume 7, Number 1, 2020, 21–34
DOI: 10.18335/region.v7i1.282

journal homepage: region.ersa.org

Teaching on Jupyter – Using notebooks to accelerate learning and curriculum development

Jonathan Reades[1]

[1] King's College London, London, United Kingdom

Received: 7 October 2019/Accepted: 5 January 2020

Abstract. The proliferation of large, complex spatial data sets presents challenges to the way that regional science and geography more widely is researched and taught. Increasingly, it is not 'just' quantitative skills that are needed, but computational ones. However, the majority of undergraduate programmes have yet to offer much more than a one off 'GIS programming' class since such courses are seen as challenging not only for students to take, but for staff to deliver. Using the evaluation criterion of minimal complexity, maximal flexibility, interactivity, utility, and maintainability, we show how the technical features of Jupyter notebooks particularly when combined with the popularity of Anaconda Python and Docker enabled us to develop and deliver a suite of three 'geocomputation' modules to Geography undergraduates, with some progressing to data science and analytics roles.

1 Introduction

The growth of data from sources that are both 'accidental, open, and everywhere' (Arribas Bel 2014), and characterised by volume, velocity, variety, and questions of veracity (Gorman 2013) has opened up new possibilities, and challenges, for researchers. This, in turn, calls for new conceptual, methodological, and technical approaches since 'acquiring data is no longer a strongly limiting factor to completing analytical tasks' (Bowlick, Wright 2018), working with it is. It is not particularly important whether these skills are framed as an informed empirical social science (Ruppert 2013) or as a computational social science (Lazer et al. 2009); authoritative reviews of the social sciences and humanities by The British Academy (2012), and of human geography by the Economic and Social Research Council (Ley et al. 2013), have concluded that many graduates are poorly prepared to engage with this world of 'big data'. The Royal Society (2019) has called for curriculum change at Higher Education Institutions (HEIs) with a view to encouraging interdisciplinarity and the effective integration of data science skills.

This presents something of a problem for a nascent 'geographic data science' (Singleton, Arribas Bel 2019) of the sort that regional science, and regional studies and geography more widely, require since a surprisingly large number of university programmes continue to teach proprietary, mostly point-and-click software. So many students' principal exposure to quantitative methods, let alone computational ones, comes in a standalone 'quantitative methods module' that provides little in the way of meaningful interaction with the underlying issues of spatial data and spatial data analysis at scale. And while the issue may be particularly acute for students in the U.K. (Johnston et al. 2014), even in more technically-oriented countries there is often not much more on offer than a

straightforward 'GIS course' (Wikle, Fagin 2014). Consequently, students progressing to higher levels of study or the professional realm often find that 'the skills least developed in undergraduate GIS courses are those related to programming and computer science' (Bowlick et al. 2017).

2 Dependencies

This notebook requires the GeoJSON labextension to be installed in JupyterLab. All other packages should be part of a default Python 3 installation.

3 Context

The long history of computers in geography has not been without controversy (Arribas Bel, Reades 2018, Barnes 2013, Cresswell 2014, Johnston et al. 2014), although many have actively engaged with recent developments (e.g. Torrens 2010) and expect impacts on the very fabric of the discipline (González Bailón 2013). So although our experience with teaching computational skills using Jupyter notebooks is clearly rooted in the 'geography of geography' (Bradbeer 1999) in the sense that we speak to particular challenges here in the U.K., it is part and parcel of a wider skills gap at the undergraduate level in general. In short, too few students are gaining the skills needed to engage with this deluge of data or to take advantage of cutting-edge tools developed outside of the field, either as researchers or as end-users in the public or private sectors (Singleton 2014).

This is where we believe that the pedagogical potential of Project Jupyter (Kluyver et al. 2016) is revolutionary: reflecting on our experience of trying to roll out exactly this type of programme, we seek to highlight the transformative potential of notebooks for student and researcher development. Jupyter removes significant barriers to teaching by providing a flexible and familiar interface that hides, or even postpones indefinitely, some of the complexity of managing local programming language installations whilst also allowing instructors to provide rich media and contextual information next to the code where it is needed the most. Making coding accessible is not simply about allowing students to 'hack away' at data, it can actually *help* students to better understand spatial analytic methods by linking concepts to code as Xiao's outstanding text on algorithms demonstrates (Xiao 2016).

3.1 Teaching Programming to Non Programmers

Given the interaction effects between pedagogical and subsequent practice, it is therefore worth placing the challenge of teaching programming in the context of the shifting terrain for quantitative research and researcher development. These challenges start early: many students already demonstrate what Spronken-Smith (2013) calls 'equation phobia': "students not linking numbers, and problems with visualisation of quantities." Hodgen et al. (2014) suggest just some of the reasons for this: limited prior knowledge and attainment; time elapsed since last study of maths; a failure to see relevance; and the wide range of attainment levels within each cohort (Hodgen et al. 2014). Whatever its origins, a general lack of confidence and/or competence creates a feedback loop fuelling further avoidance (Chapman 2010).

In the context of maths instruction Macdonald, Bailey (2000) have also noted the challenge inherent in delayed gratification given that 'maths is the tool, not the goal.' Given the apparent gulf between `print('Hello world.')` and being able to write useful analytical code, the issue is no less serious in programming. There is no reason why the familiarity of so-called 'Digital Natives' with computers should have any bearing on their understanding of how they actually work; indeed, today's students may well be *more* detached from the underlying processes – metaphorical and actual – thanks to 'the sophistication of modern Graphical User Interfaces' (Muller, Kidd 2014). In the long run, programming requires an ability to envision and manipulate abstract entities such as data structures sitting, in turn, on top of additional layers of abstraction such as the application and its state(s), the file system and its structure(s), the operating system and even the underlying hardware.

Figure 1: Barron Stone memorably demonstrates *for* and *while* loops (Stone 2013)

There are many differing views of how programming should be taught (Pears et al. 2007), though we come down firmly on the side of Lukkarinen, Sorva (2016) that there are advantages to 'contextualising programming practice in the field of application'. In general, it seems that introductory programming courses should strive simultaneously for richness and simplicity: richness in the 'constructs' associated with programming, and simplicity in terms of the foundation being laid (Lukkarinen, Sorva 2016). Unfortunately, the expertise of teachers is not always a plus for effective teaching (Chapman 2010) since concepts that seem intuitive and are easily connected to a range of related problems by the instructor may yield no such benefit to the novice. As we developed our teaching materials, we found that videos created by other learners could, at times, capture student attention more effectively than our own demonstrations; for example, Stone's instructional video for students at Rice University on the difference between for and while loops, shown in Figure 1. Using Jupyter notebooks this kind of content can be embedded directly in the task explanation.

3.2 Course Structure

The work reported here draws on methodological and pedagogical research conducted over the past five years in the Department of Geography at King's College London; it seeks both to position learning to code as essential to further student and staff development, and to examine the reasons why Jupyter notebooks have been selected as the best means of achieving this goal. As such, this research is necessarily caught up in a wider debate about quantitative skills amongst students; however, our undergraduate 'pathway' in *Geocomputation & Spatial Analysis* (which could be understood as an optional 'minor' in the North American tradition) seeks to go beyond the kinds of statistical skills training encouraged by funders (see brief discussion in Johnston et al. 2014) and to tackle these in conjunction with computational skills. We want to take students with a variety of social, economic, ethnic, and computational backgrounds and cultivate in (and with) them an appreciation of, and ability to undertake, interdisciplinary work with a strong computational element (see Mir et al. 2017, for a discussion of the *CS+X* format).

Based on our own experience, we felt that shoe horning exposure to 'computational geography' into a single module – as seems to occur in many American programmes (Bowlick et al. 2017) – would only reinforce student aversion to such approaches, so we opted to 'unpack' the concepts across three modules:

1. Geocomputation
2. Spatial Analysis and Modelling, and
3. Applied Geocomputation.

These modules must be taken in sequence, the preceding module acting as a pre-requisite for admission to the next, although students are free to exit the sequence at any time. We also provide an optional 'Code Camp' (Reades et al. 2019) to be undertaken over the summer before the first module begins so that students begin the term familiar with basic concepts: variables, lists/arrays, dictionaries/hashes, and functions/subroutines, provided they have done the work.

3.3 Contextualised Computing

To our knowledge, there is no other undergraduate programme like it with important differences in both style and substance from what would be covered in an Economics, Statistics, or Computer Science (CS) degree in terms of its spatial and applied focus. In this sense, the modules are an extended test of 'contextualised computing' instruction (see Lukkarinen, Sorva 2016, for a review) which seeks to emphasise relevance to 'real-world' applications and to avoid "general CS content, such as how one might go about sorting an array of any type for an unspecified purpose" (Lukkarinen, Sorva 2016). We also recognise, however, that "contextualized computing education cannot help students learn more in less time" (Guzdial 2010) and that the *transferrable* aspects of this learning need to be emphasised: in our case we try to highlight how the same approach can be applied to human and physical geography problems.

Consequently, wherever possible these exercises are grounded in spatial examples, even where these are very simple indeed, on the basis that connecting them to the learner's existing knowledge and interests will improve retention at the introductory level (Guzdial 2010). For example, a notebook on dictionaries (taken from Reades et al. 2019) can start with creating and querying a phone book of national emergency numbers where the student has to replace the `???` in `eNumbers = { ??? }` with functioning Python code:

[1]:
```
eNumbers = {
        'IS': 112,
        'US': 911
}
print(f"The Icelandic emergency number is {eNumbers['IS']}")
print(f"The American emergency number is {eNumbers['US']}")
```

[1]:
```
The Icelandic emergency number is 112
The American emergency number is 911
```

Students then progress towards a task involving a dictionary-of-dictionaries:

[2]:
```
cityData = {
        'London': {
                'population': 8673713,
                'location': [51.507222, -0.1275],
                'country': 'UK'
        },
        'Paris': {
                'population': 2140526,
                'location': [48.8567, 2.3508],
                'country': 'FR'
        }
}

for city, data in cityData.items():
  print(f"The population of {city} ({data['location'][0]:0.3f}ºN,
   {data['location'][1]:0.3f}ºE) is {data['population']:,}")
```

[2]:
```
The population of London (51.507ºN, 0.128ºE) is 8,673,713
The population of Paris (48.857ºN, 2.351ºE) is 2,140,526
```

This work is building towards a GeoJSON example in which they have to complete missing attributes in order to show a marker centred on the university's central London campus. Since GeoJSON is essentially a dictionary-of-dictionaries, this is a good test of their understanding, but with Jupyter they receive immediate feedback on this because GeoJSON can be embedded directly into the notebook: an interactive web map shows up

as soon as they've run the code, reinforcing the contextual aspect – that this is *all* about geography – of their learning.

[3]:
```
# King's College London's coordinates...
# What format are they in? Does it seem appropriate?
# How would you convert them back to numbers if you
# needed to do so?
longitude = '-0.11596798896789551'
latitude  = '51.51130657591914'

# Notice how we set up a data type and location
# here where it's easy to see where the lat/long
# values are being used we could also use these
# in a loop as a _template_ for creating many points
# from a data file! Notice too that it's a dictionary
# containing a mix of string and list values...
the_geometry = {
        "type": "Point",
        "coordinates": [longitude, latitude],
}

# Now we set up the larger 'data file' this is harder
# to read but is *still* basically a dictionary! A
# 'collection' implies more than one feature, and in this
# case the list of 'features' is nothing more than a list
# of dictionaries so that our data stays in order!
the_position = {
        "type": "FeatureCollection",
        "features": [
        {
                "type": "Feature",
                "properties": {
                        "marker-color": "\#7e7e7e",
                        "marker-size": "medium",
                        "marker-symbol": "building",
                        "name": "KCL"
                },
                "geometry": the_geometry
        }
        ]
}

# And show the points on an interactive map!
# You don't need to know what's happening here *yet*, but
# see if you can make sense of the main elements...
try:
from IPython.display import GeoJSON
from IPython.display import display
import json
parsed = json.loads(str(the_position).replace("\'", "\""))
display(GeoJSON(parsed))
except ImportError:
print("You seem to be missing either the GeoJSON extension or json library.")
```

[3]:
```
The output is shown in Figure 2
```

4 How We Reached Jupyter

Since the pathway pushes students both conceptually and technically, finding ways to take the deployment and management of the software stack out of the picture has been a priority. Our review of the pedagogical literature and practical experience gained in the private and HEI sectors—including several failures during the first few years of teaching—led us to the ultimate conclusion that a useful geospatial programming environment should possess the following characteristics:

Minimal Complexity : it does not require students to load and learn a new Operating System or large number of new applications/platforms at the same time as they are

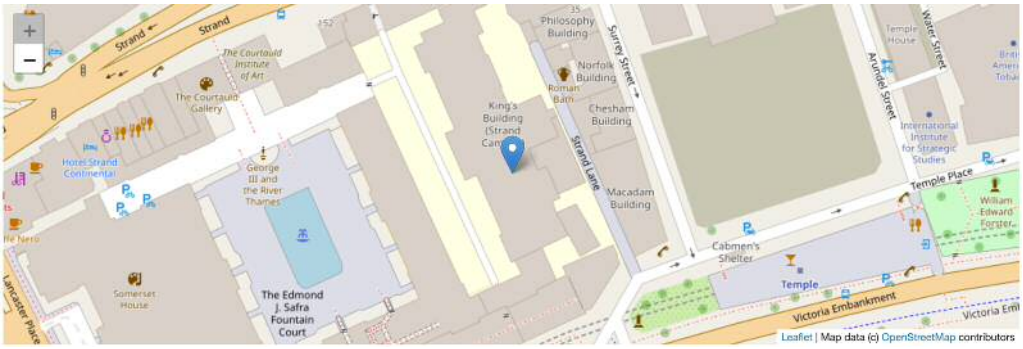


Figure 2: Output of code 3

learning to code; it should also be reasonably 'performant' on a mix of student and HEI hardware.

Maximal Flexibility : it is simple, if not always easy, to configure and install on a range of hardware, but is not 'sandboxed' or 'packaged' in ways that constrain our freedom to install what we need to teach effectively.

Interactivity : it allows us to keep commentary, 'rich' media, and other scaffolding material together with the code so that students can move between code and explanations easily, and can add their own annotations as needed.

Utility : it supports life-long learning by providing a 'real world' development environment that would be both familiar, and accessible, to students after graduation in personal and professional contexts.

Maintainability : it can be easily updated by the instructor(s) and supports version control and easy distribution mechanisms.

These five features can, at times, appear to cut against each other: maximal flexibility and minimal complexity are difficult to reconcile since the former tends to expose more 'options' to the user, while the latter seeks to mask those same options. However, a strong advantage of Jupyter is that it meets all of these criteria to some extent, and in most cases meets them fully!

4.1 Pretty Walled Gardens

The desired set of features ruled out commonly-used proprietary platforms: at the time we began developing the curriculum, MATLAB was still a *de facto* standard for many but its pricing and sandboxing approach made it both less flexible and less useful for students once they graduated and lost access to the HEI license. Like Etherington (2016), we were therefore attracted by the fact that Python presented 'no financial or hardware obstacles to teaching' and that, consequently, "students [would] always be able to use their Python programming skills..." Etherington (2016). However, in developing the early iterations of the course we also, again like Etherington (2016), encountered significant challenges in 'getting a working installation of Python together with its associated geospatial packages'.

We discovered that the existing, IT-supported Enthought Canopy Python distribution provided few of geospatial libraries, and that updating it with packages from outside of their 'walled garden' caused all manner of issues. This situation was not entirely unexpected since geospatial analysis is not a key component of Enthought's offering to universities; however, the challenges of keeping up with the state-of-the-art are such that additional barriers to software update management are undesirable. Indeed, the pace of change in the field can be gauged from Wise's review of 'geospatial technologies' in U.K. universities (Wise 2018): it not only questions the utility of 'free' programmes (presumably meaning Free Open Source Software, or FOSS) which now dominate in the data sciences and in many research projects, but it also contains not a single mention of programming—in Python or any other language.

4.2 The Wrong Kind of Flexibility

Like Muller, Kidd (2014), who sought to 'debug geographers' with an introduction to a holistic computing context alongside programming skills *tout court*, we next attempted to provide our students with virtualised Linux desktop systems in the belief that this would empower them not only with a better understanding of what was going on 'under the hood' but also with a computer on which they could experiment without fear of damaging their existing installation. For good measure, we included other useful analytics tools such as the latest version of QGIS with all of the 'bindings' for low-level packages such as GDAL (the Geospatial Data Abstraction Layer).

Using VMWare and Ubuntu 16 LTS with a full Python installation configured largely 'by hand' provided us with a fully FOSS 'solution' that students could take with them and update in the future as they gained confidence in using such software. However, we soon found that in-memory and on-disk bottlenecks, together with students' tendency to actually try to install Ubuntu's suggested updates and render their systems inoperable, made this a profoundly alienating and frustrating experience. For students already working hard to master the basics of programming, having to 'drop' into the Terminal in order to resolve installation errors when they were used to seamless updates on their host operating systems simply represented an unnecessary hassle that detracted from the real focus of the modules: learning to use code to perform spatial analyses.

4.3 Escape Velocity

While we had been tinkering with different Linux and Python distributions, a set of three connected developments had been transforming the landscape for teaching:

1. A few academics who had taken very different approaches began, rather bravely, to publish their teaching methods and materials freely for others to use (e.g. Arribas Bel 2019);

2. Data scientists not only adopted Python *en masse*, driving the rapid development of new analytical and visualisation libraries (*e.g.* pandas, seaborn, bokeh), but they had also quickly settled on the use of a then-novel technology called 'iPython notebooks' to widely share their tutorials online;

3. Since many of these data scientists were paid by firms interested in moving their work into production systems as smoothly and quickly as possible, this also led to improvements in the way that Python distributions and notebooks were managed.

Rather unexpectedly, the kinds of practical problems that data scientists were trying to solve mirrored quite closely the kinds of challenges that we, as teachers, were trying to solve in terms of being able to replicate installations across multiple systems and share code/commentary quickly and easily.

The iPython platform ultimately gained the ability to run other programming languages and was rebranded 'Project Jupyter', but this means that it has become a viable, general purpose teaching platform. So although the term 'Virtual Learning Environment' (VLE) is typically understood to refer to a full-featured, client-server system such as Moodle or Blackboard (see Britain 1999), it could also apply to Jupyter: not only does it have a client/server architecture (with the web-based interface allowing the server to run locally or on a remote system with no discernible difference to the student), but it has been progressively enriched with tools for grading and other common teaching tasks. Although we are not yet making full use of these new features, it is clear that Jupyter is well on its way to becoming an important teaching platform.

5 Discussion

Perhaps the single greatest benefit of working with Jupyter notebooks is that development is *not* being driven by educational needs: this is a full-featured development environment used day-in and day-out by professional software developers and large firms such as Netflix

(Ufford et al. 2018). So, unlike both expensive proprietary systems that are rarely used by small or innovative firms, and instructional systems whose functionality is limited to teaching purposes, students are able to seamlessly progress from learning to code, to competent coders, and on to practicing data scientists (as a few of our students have done), using a single environment. This is a platform with the capacity to grow with the student, following them out of the 'ivory tower' and into gainful employment.

An additional benefit flowing from the professional use of Jupyter is that many researchers, not least the others included in this special issue, use notebooks as a normal part of their research practice; this allows lecturers to remain abreast of technical developments on the platform without 'updating my installation' being a separate overhead in a congested working week. This pattern of usage is in sharp contrast to tools – such as SPSS or ArcGIS – that are less-used by active researchers but often still taught in standalone modules, with the quality and timeliness of teaching materials often suffering accordingly. Jupyter breaches the historical divide between computational research and teaching, not only allowing students to benefit from active research, but also for research to build on student outputs (see, for example Reades et al. 2019).

5.1 Cloning Around

Jupyter becomes particularly powerful when combined with other recent developments in the management and distribution of computing platforms. Anaconda Python's enhanced support for the configuration of virtual environments (in essence, multiple distributions of Python on the same system) allows specific versions of Python and sets of required libraries to be specified in a simple text file following the 'Yet Another Markup Language' (YAML) standard. The code below downloads and prints out part of the YAML file that we use to configure both student machines *and* our Docker container (about which more below); here the virtual environment is named `gsa2019`:

[4]:
```
import urllib

url  = 'https://raw.githubusercontent.com/kingsgeocomp/gsa_env/gsa2019/gsa.yml'
with urllib.request.urlopen(url) as resp:
file = resp.read().decode('utf8').split('\n')

# Don't output everything...
to_print = list(range(0,5))  list(range(39,48))  list(range(110,116))

print("=" * 50)
for line in to_print:
print(file[line])
print("=" * 50)
```

[4]:
```
==================================================
# OVERVIEW
# This YAML script will attempt to install a Python virtual environment able to
# support the requirements of all three of King's College London's 'Geocomputation'
# pathway in the BA/BSc Geography programme.
#
# CONFIGURATION PARAMETERS
name: gsa2019
channels:
 - conda-forge
 - defaults
dependencies:
 - python=3.7
 - pip
 - git
 - xlrd
 - xlsxwriter
 - pip:
   - six
#- git\+http://github.com/sevamoo/SOMPY#egg=sompy # Doesn't run in Python3
   - git\+http://github.com/kingsgeocomp/SOMPY#egg=sompy
==================================================
```

The use of YAML configuration files makes it easier to install a teaching instance of

Python and to expose this as a named 'iPython kernel'. The connection between virtual environments and kernels allows researchers to manage multiple research and teaching installations of Python on the *same* system, to access them through the same Jupyter interface, and to do so without changes to one Python installation impacting any others.

5.2 Docking Safely

The emergence of containerisation platforms such as Docker now makes it much simpler to distribute a pre-configured virtual machine[1] (such as a pre-packaged teaching or research environment) that will run on almost any host operating system: Mac, Windows, or Linux. Because the virtual machines are fully specified at the time of creation, students can download and install a working version with one command, while instructors can be confident that every student is working with the *same* version of every library. This year we provided students with a Docker image that leveraged the work of Arribas Bel (2019) but that had been customised to provide only the features that we wished to teach.

The combined popularity of Python and Docker has led to the creation of novel, web-based platforms such as Binder (mybinder.org); these take notebooks stored on the GitHub code-sharing web site to build a Docker image serving those notebooks on Binder's servers. Students may now learn to code without installing any software whatsoever. Local installation can be deferred to the point at which specialist requirements or load on the server require it. In a stroke, one of the most pernicious barriers to entry, needless technical issues associated with installation and configuration of programming software, has been eliminated.

5.3 Houston, We Have a Problem

Of course, no single solution is without drawbacks and Jupyter is no exception. It is worth noting that there are quite specific technical, conceptual, and development issues raised by Jupyter that are difficult to circumvent without both know-how and some careful thinking about assessment and teaching. The principal technical challenge relates to user permissions on managed machines (*e.g.* in computer clusters) since Python, Jupyter, and Docker all struggle to different degrees with 'locked down' Windows systems. Indeed, Docker does not currently run at all without administrator privileges. We worked closely with university-level IT staff to install and provision Anaconda Python and Jupyter. Provision of the YAML configuration script assisted with both installation and isolation of our teaching environment from their existing installation, easing institutional barriers to adoption.

From a teaching standpoint, an additional issue is that Git – the dominant version control software that we use to manage and share notebook changes – sees notebooks in a way that means just re-running code registers as a local modification of the file that needs to be committed to the version control system. So although 'GitHub' provides support for the online display of Jupyter notebooks, the use of Git can lead to a large number of essentially meaningless commits. This can make tracking meaningful content changes over time more difficult, and it means that we've shied away from teaching students about version control on the basis that they may not perceive the value of commits that seem to record little of value.

A final and rather unexpected disbenefit was uncovered the year after we moved from the Spyder IDE to Jupyter: weaker student understanding of execution flow. Unlike a traditional script that clearly executes from top-to-bottom (typically in its entirety), Jupyter notebooks freely intermingle code blocks and text/rich media blocks allowing – and even encouraging – the user both to jump between widely separated blocks without executing intervening code and to edit and re-run earlier blocks. This leads to: a) difficult-to-diagnose bugs because the code *looks* like it should execute properly but doesn't, and b) to a weaker student understanding of system 'state' in terms of instantiated variables, loaded libraries, and available functions. We typically seek to cultivate this understanding by stressing that the *real* test, whether directly assessed or not, of whether their code

[1] It should be noted that, technically, Docker containers are not virtual machines in the traditional sense.

Table 1: Evaluating Jupyter

	Pros	Cons
Minimal Complexity	Deploying a full geographic data science 'stack' requires installing one application (Docker or Anaconda Python) and running two lines of code in a Terminal/Shell to install and configure Jupyter, its dependencies, and the analytical libraries. Environment requires no configuration.	Persistent challenges with student understanding of file system interaction and paths. Some confusion around multiple Python instances manifesting as different 'kernels' in notebooks.
Maximal Flexiblity	Combination of Binder, Docker, and Anaconda Python allows us to install on nearly any hardware/operating system mix. Docker uses same YAML configuration script as Anaconda Python so maintaining compatibility and consistency is straightforward.	Students cannot update Docker containers and do not gain understanding of package management or dependency conflict resolution.
Interactivity	Students can view/edit/add rich media, code, and other content directly within the Jupyter notebook environment. Textual and graphical outputs from code cells in notebooks are saved between restarts of Jupyter.	Students do not develop a strong understanding of execution flow and system state.
Utility	Growth of Jupyter has made it the 'tool of choice' for data scientists, and students are able to continue working with a fully functioning development environment. Students can edit installation and configuration scripts incrementally, as expertise grows.	Relative ease of installation may not prepare students for managing their own development and production environments. Students remain unfamiliar with IDEs and code-completion.
Maintainability	Docker and Anaconda update mechanisms are straightforward. GitHub works well for distribution, previewing, and (to a lesser extent) version control.	Nature of notebooks makes it harder for instructors to track incremental changes in version control, and for students to see value of such an approach.

'works' is that a notebook can be run in full (`Restart Kernel and Run All Cells`) without user intervention.

We should note that, in the absence of an Integrated Development Environment (IDE), students are unlikely to benefit from test suites and other tools that support developer best-practice. However, such an approach can also have the effect of deterring new students by pushing back the point at which they appear to be achieving anything concrete: "Because learning in computer science and programming is challenged by numerous barriers, students need to be motivated about the purpose, value, and utility of concepts within course work" (Bowlick et al. 2017) So while knowledge of professional tools and practices is desirable, we nonetheless feel that these kinds of ideas and issues are best tackled when students have progressed further with their studies and are motivated to tackle more abstract challenges.

6 Conclusion: Back Here on Earth

In order to understand why the practical benefits of teaching with Jupyter notebooks outweigh the technical and conceptual challenges encountered, it is worth returning to the evaluation criteria outlined near the start of this work. Table 1 summarises the pros and cons observed across the five dimensions identified by our review of the state-of-the-art nearly six years ago.

From this, the principal technical recommendation is that a flexible mix of platforms should be used to deliver Jupyter-based learning. We recommend Binder to deliver foundational material using few non-core Python libraries, and now strongly recommend that students use Docker in subsequent modules. However, a critical issue is that Windows 10 Home Edition does not support Docker, and it is therefore *still* necessary to support direct installation of Anaconda Python and associated configuration of the 'kernel' using a

YAML text file. We are also investigating the use of a containerised JupyterHub running on our own hardware: this would allow students to mimic using Binder while benefiting from the ability to save work and make full use of Python's capabilities. All of the code supporting these configurations is available as a Github repository, as is Arribas-Bel's resource.

6.1 And Back to the Future

A failure to engage directly with computational approaches and tools poses long-term risks: while ours 'has always been a following discipline' (Burton 1963), what is new is that other disciplines have now taken an interest in cities and regions (O'Sullivan, Manson 2015). Ruppert (2013) warns, "if social scientists do not step forward, then computational social science risks becoming the exclusive domain of ... computing scientists" (Ruppert (2013). However, there is also an enormous opportunity for students equipped with both domain knowledge and programming skills to act as 'knowledge brokers' (Bowlick, Wright 2018). As Mir et al. (2017) note: "truly transformative work at the intersection of computing and ... other disciplines requires ... people with heterogeneous skill-sets (both computational and non-computational) who, despite their differences in training, can work collaboratively." In other words, facing the future requires both translators and explorers: individuals who understand the broader terrains across which knowledge moves and the frontiers at which new knowledge is generated.

We have also come to believe that the use of Jupyter-like platforms in non-STEM disciplines may have a role to play in addressing a deeper problem: the widening participation challenge in computationally-oriented disciplines such as data science (The Royal Society 2019). A particular contribution is these other disciplines' capacity to provide an applied context for computational training that helps to motivate further study and engagement (see Bort et al. 2015, for a creative application in literary studies). It should not be the responsibility of Geography and allied fields to plug the so-called 'leaky pipeline' (Berryman 1983), but they may yet create novel pathways for a more diverse cohort of students to enter computationally intensive fields. Such an outcome would not only be to the benefit of Computer Science, it would very much be to the benefit of an innovative Regional Science as well.

Acknowledgements

This work builds on the input of many – staff and students – to the Geocomputation and Spatial Analysis pathway at King's College London; however, I wish to particularly acknowledge the critical contributions of Dr. James Millington, Michele Ferretti, Dr. Chen Zhong, and Dr. Yijing Li. Finally, Dr. Arribas-Bel has donated many hours of his time – directly and by example – to helping me to develop and migrate our teaching environment.

References

Arribas Bel D (2014) Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography* 49: 45–53. CrossRef.

Arribas Bel D (2019) A course on geographic data science. *The Journal of Open Source Education* 2: 42. CrossRef.

Arribas Bel D, Reades J (2018) Geography and computers: Past, present, and future. *Geography Compass* e12403

Barnes T (2013) Big data, little history. *Dialogues in Human Geography* 3: 297–302

Berryman S (1983) Who will do science? Trends, and their causes in minority and female representation among holders of advanced degrees in science and mathematics. A special report. Rockefeller Foundation, New York, NY

Bort H, Czarnik M, Brylow D (2015) Introducing computing concepts to non majors: A case study in gothic novels. Proceedings of the 46th ACM Technical Symposium on Computer Science Education, 132–137. ACM. CrossRef.

Bowlick F, Goldberg D, Bednarz S (2017) Computer science and programming courses in geography departments in the United States. *The Professional Geographer* 69: 138–150. CrossRef.

Bowlick F, Wright D (2018) Digital data centric geography: Implications for geography's frontier. *The Professional Geographer* 70: 687–694. CrossRef.

Bradbeer J (1999) Barriers to interdisciplinarity: Disciplinary discourses and student learning. *Journal of Geography in Higher Education* 23: 381–396. CrossRef.

Britain S (1999) A framework for pedagogical evaluation of virtual learning environments. Report, Joint Information Systems Committee. https://www.webarchive.org.uk/wayback/archive/20140613220103/http://www.jisc.ac.uk/media/documents/programmes/jtap/jtap-041.pdf

Burton I (1963) The quantitative revolution and theoretical geography. *The Canadian Geographer/Le Géographe Canadien* 7: 151–162. CrossRef.

Chapman L (2010) Dealing with maths anxiety: How do you teach mathematics in a geography department? *Journal of Geography in Higher Education* 34: 205–213. CrossRef.

Cresswell T (2014) Déjà vu all over again: Spatial science, quantitative revolutions and the culture of numbers. *Dialogues in Human Geography* 4: 54–58

Etherington T (2016) Teaching introductory GIS programming to geographers using an open source python approach. *Journal of Geography in Higher Education* 40: 117–130. CrossRef.

González Bailón S (2013) Big data and the fabric of human geography. *Dialogues in Human Geography* 3: 292–296. CrossRef.

Gorman S (2013) The danger of a big data episteme and the need to evolve geographic information systems. *Dialogues in Human Geography* 3: 285–291. CrossRef.

Guzdial M (2010) Does contextualized computing education help? *ACM Inroads* 1: 4–6. CrossRef.

Hodgen J, McAlinden M, Tomei A (2014) Mathematical transitions: A report on the mathematical and statistical needs of students undertaking undergraduate studies in various disciplines. Report, The Higher Education Academy

Johnston R, Harris R, Jones K, Manley D, Sabel C, Wang W (2014) Mutual misunderstanding and avoidance, misrepresentations and disciplinary politics: Spatial science and quantitative analysis in (United Kingdom) geographical curricula. *Dialogues in Human Geography* 4: 3–25. CrossRef.

Kluyver T, Ragan Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team (2016) Jupyter notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B (eds), *Positioning and power in academic publishing: Players, agents and agendas*. IOS Press, 97–90

Lazer D, Pentland A, Adamic L, Aral S, Barabási A, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Life in the network: The coming age of computational social science. *Science* 323: 721–723. CrossRef.

Ley D, Braun B, Domosh M, Elliott S, Le Heron R, Peake L, Willekens F, Yeoh B (2013) International benchmarking review of UK human geography. Report, Economic and Social Research Council, in partnership with the Royal Geographical Society (with IBG) and the Art and Humanities Research Council. https://esrc.ukri.org/files/research/research-and-impact-evaluation/international-benchmarking-review-of-uk-human-geography/

Lukkarinen A, Sorva J (2016) Classifying the tools of contextualized programming education and forms of media computation. Proceedings of the 16th Koli Calling International Conference on Computing Education Research, 51–60. ACM. CrossRef.

Macdonald R, Bailey C (2000) Integrating the teaching of quantitative skills across the geology curriculum in a department. *Journal of Geoscience Education* 48: 482–486. CrossRef.

Mir D, Mishra S, Ruvolo P, Pollock L, Engen S (2017) How do faculty partner while teaching interdisciplinary CS+X courses: Models and experiences. *Journal of Computing Sciences in Colleges* 32: 24–33

Muller C, Kidd C (2014) Debugging geographers: Teaching programming to non computer scientists. *Journal of Geography in Higher Education* 38: 175–192. CrossRef.

O'Sullivan D, Manson S (2015) Do physicists have geography envy? and what can geographers learn from it? *Annals of the Association of American Geographers* 105: 704–722. CrossRef.

Pears A, Seidman S, Malmi L, Mannila L, Adams E, Bennedsen J, Devlin M, Paterson J (2007) A survey of literature on the teaching of introductory programming. *ACM SIGCSE Bulletin* 39: 204–223. CrossRef.

Reades J, De Souza J, Hubbard P (2019) Understanding urban gentrification through machine learning. *Urban Studies* 56: 922–942. CrossRef.

Reades J, Ferretti M, Millington J (2019) Code camp: 2019. Github repository, King's College London

Ruppert E (2013) Rethinking empirical social sciences. *Dialogues in Human Geography* 3: 268–273. CrossRef.

Singleton A (2014) Learning to code. *Geographical Magazine* 77

Singleton A, Arribas Bel D (2019) Geographic data science. *Geographical Analysis*: 1–15. CrossRef.

Spronken-Smith R (2013) Toward securing a future for geography graduates. *Journal of Geography in Higher Education* 37: 315–326. CrossRef.

Stone B (2013) Differences between for & while loops (in Python). Video, YouTube. https://www.youtube.com/watch?v=9AJ0uoxtdCQ

The British Academy (2012) Society counts. Report, The British Academy, https://www.-thebritishacademy.ac.uk/sites/default/files/BA Position Statement - Society Counts.pdf

The Royal Society (2019) Dynamics of data science skills: How can all sectors benefit from data science talent? Report, The Royal Society, https://royalsociety.org/-/media/policy/projects/dynamics of data science/dynamics of data science skills report.pdf

Torrens P (2010) Geography and computational social science. *GeoJournal* 75: 133–148. CrossRef.

Ufford M, Pacer M, Seal M, Kelley K (2018) Beyond interactive: Notebook innovation at Netflix. Blog post, Netflix, https://netflixtechblog.com/notebook-innovation-591ee3221233. [last checked: 3 October 2019]

Wikle T, Fagin T (2014) GIS course planning: A comparison of syllabi at US college and universities. *Transactions in GIS* 18: 574–585. CrossRef.

Wise N (2018) Assessing the use of geospatial technologies in higher education teaching. *European Journal of Geography* 9

Xiao N (2016) *GIS Algorithms: Theory and Applications for Geographic Information Science & Technology*. Research Methods. SAGE. CrossRef.

REGION
The Journal of ERSA
Powered by WU

Volume 7, Number 1, 2020, 35–48
DOI: 10.18335/region.v7i1.268

journal homepage: region.ersa.org

The Impact of Migration on a Regulated Rental Market

Adam Alexander Tyrcha[1]

[1] University of Cambridge, Cambridge, United Kingdom

Received: 21 June 2019/Accepted: 3 February 2020

Abstract. Throughout the 20th century, the Swedish rental market has generally been heavily regulated, with both a rental queue in place, as well as fixed rents, with a limited ability to vary these. Though these systems remain in place, in the 21st century, a number of deregulatory measures have been taken. Meanwhile, evolving migration flows and strong humanitarian migration in particular have continued. These developments combined mean that now more than ever, the impacts of migration on the rental housing market are increasingly likely. This paper investigates the relationship between foreign-born and internal migration and rents on the housing market. Findings suggest that foreign-born migration, and refugees in particular, impact rents, especially in major cities.

JEL classification: J61, R23, R31, R11

Key words: immigration; internal migration; housing; rents; rental market

1 Introduction

Rental markets constitute a crucial segment of most housing markets worldwide due to the flexibility that this type of housing provides, allowing people who lack the means or desire to purchase housing an easily accessible alternative. When considering the rental market, drawing on examples from Sweden are particularly interesting, owing to the differences in the country's rental system when compared to most other systems. Though public and private actors construct rental housing, the Swedish rental system is rent controlled, with the ability of the landlord to set rent being limited. However, some deregulation has occurred over the past decade or so, transforming the nature of this market (Section 2).

Meanwhile, migration flows coming into Sweden also continue to transform. In 2017, 144,489 people migrated to Sweden, where 27% of all new residence permits approved were granted to humanitarian migrants, with 35% corresponding to labour migrants, and 24% to family reunification migrants (SCB 2017). Hence, it is clear that a diverse range of migrants continue to seek opportunities for themselves in Sweden.

Substantial amounts of research have been conducted into the relationship between migration and the owner-occupied housing market (Saiz 2007, Degen, Fischer 2009, Gonzalez, Ortega 2012, Tyrcha, Abreu 2019). Some research has also been conducted into the relationship between migration and the rental market, generally finding a 1% increase in population leads to between 0-1% increase in rents (Ottaviano, Peri 2005, Saiz 2007, Latif 2015, Tumen 2016, Mussa et al. 2017). This is an important relationship

to understand, owing to the large proportion of the global population that does live in rented accommodation. Both migrants and natives alike, particularly those who are less affluent, tend to be overly reliant on rental accommodation (World Bank 2019), meaning that the global relevance of this topic is clear.

However, no research has been conducted in Sweden, despite 59 per cent of apartments in Sweden being rental apartments (SCB 2018). Furthermore, the unique institutional context of the Swedish market, where rents have historically been highly regulated, but underwent somewhat substantial deregulation in 2011 (SABO 2011), provides an interesting context for the analysis of the impact of migration on a recently deregulated housing market. Thus, in this paper I extend the literature by investigating the impacts of migration on rent in the highly regulated Swedish context. The paper will examine both general and regional impacts, while accounting for internal migration as well as different forms of foreign-born migration, allowing for further extension of the literature in this manner.

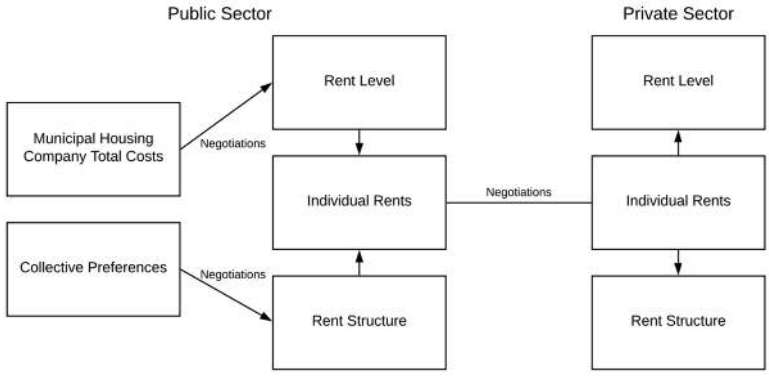
2 Background

2.1 Literature Review

As alluded to in Section 1, there has been some research going into the impacts of migration on the rental market previously, though primarily in unregulated markets. This research has found that migration has an impact on rents, owing primarily to the demand-driven effects that an influx of population creates, pushing up prices by 0-1% (Ottaviano, Peri 2005, Saiz 2007, Latif 2015, Tumen 2016, Mussa et al. 2017). Indeed, perhaps the seminal study on the impact of immigration on the rental market is Saiz (2003). He finds that following an exogenous migration shock to Miami's renter population in the 1970s and 1980, rents rose by "8% to 11% more in Miami than in comparison groups between 1979 and 1981." Saiz finds that the shock to unskilled migration had a greater impact on rents in poor areas and persisted in the years following the migration shock. His study extended to the fairly unique migration boost provided to Miami by the Mariel Boatlift from Cuba, and thus it is possible that impacts were different from more gradual migration impacts.

However, the trends found by Saiz (2003) also hold in Saiz's own follow-up (Saiz 2007), where he finds that immigration inflows equal to 1% of a city's size lead to a 0.6% increase in rents, confirming this on the national scale. Mussa et al. (2017), though, use US data to find an impact of 0.8% in the immediate area, rising to 1.6% in surrounding MSAs indicating the presence of spillover effects expected where there is spatial sorting. A recent study by Tumen (2016) in Turkey confirms the above findings, and although a different methodology is employed, finds a 5.5% increase in house rents overall resulting from the natural experiment of Syrian refugees coming to Turkey. Results here could be significantly stronger owing to the substantially varying nature of the migration flow, as well as the slightly less developed nature of the receiving country. The same reasoning could be applied to an even more recent study, though of a relatively more dated time period in the fall of the Berlin Wall. This study also finds that an increase in migration corresponding to 1% of western Berlin's population resulted in rents rising between 3.3-4.8% (Kürschner 2017).

Nevertheless, impacts are not uniformly positive across all studies. "An inflow of immigrants equal to 1 percent of the initial population is associated with a 0.14-0.18 percent increase in average housing rent" (Aitken 2014, p. 13) in England and Wales, indicating that a moderate effect remains a possibility in this regard. Differences found in this study could also broadly be a result of the generally unique nature of the UK migration and housing context. However, similar findings are made by Latif (2015) in Canada, where an increase in migration of 1% is found to correspond to a 0.14-0.17% increase in rents, although a lack of key controlled variables could serve to explain this. Indeed, this corresponds well with the findings of Sharpe (2015), who after introducing a range of new control variables, is unable to ascertain any causal effects of migration on the US rental market. This is important to keep in mind when conducting analysis, and underlines that diverging impacts could be found depending on controlled variables,



Source: Adapted from: Bengtsson 1994.

Figure 1: The Swedish use-value system

techniques taken, and countries chosen to study. This is particularly relevant in Sweden, where the rental market operates under a fairly unique rental system, described in the following section (Section 2.2).

2.2 The Swedish Rental Market

The rental system in Sweden has its groundings in the Rent Regulation Act of 1942 and the Rental Act of 1968. In essence, the aim of the law is: "First, regardless of the market situation, the landlord should be prevented from raising the rent of a flat in order to get rid of an undesired sitting tenant. Second, in times of housing shortage, the landlord should be prevented from raising the rent to market level the sitting tenant cannot afford. Third, the landlord should be prevented from raising the rent without the sitting tenant having a real chance to look after his interests, individually or with the assistance of a tenants association." Nevertheless, the rent should, in theory, still "reflect market rents on a market in long-run equilibrium, regardless of whether the market at any given time is actually in equilibrium or not" (Bengtsson 1994, p. 3).

Though the current outcome was not necessarily the intent upon creation, Figure 1 demonstrates the state of the rental system in practice, until the deregulatory reform in 2011.

The original intention of the system was for the public and private sectors to have input in negotiations (Bengtsson 1994). However, as Figure 1 indicates, the private sector had effectively become a price-taker in the equation, with "tenant associations [having] the formal right to collective negotiations, even against the landlord's will" (Bengtsson 1994, p. 4), effectively removing bargaining from the process. Instead, "claimed rents [were] compared with the local rent level of dwellings judged to have the same use-value in terms of size, standard, service, location and more" (Lind 2001, p. 11), where "reasonableness of a rent in the private sector is directly related to the rent set by the public sector" (Lind 2001, p. 11). Further, newer apartments tend to have higher rents, owing to a lack of regular rent reviews in older apartments, and not accounting for various factors in the rent-setting process e.g. the role of refurbishment. Indeed, urban renewal and re-urbanization has resulted in "the gap between the actual rent and the market rent increasing dramatically in old stock in attractive areas" (Lind 2001, p. 3).

Though some of the above still holds, 2011 saw a wide-reaching reform, enabling a higher degree of negotiation and competitiveness in the rent-setting process, while still maintaining elements of rent regulation. Since the 1st of January 2011, the municipal housing companies' role has been vastly diminished, and they are no longer primarily involved in rental negotiations, as shown in Figure 1. Instead, rents can be negotiated with any relevant party (SABO 2011). This means that there is scope for the private sector to not just be a price-taker, but be more actively involved and even leading in the negotiations. Though rents must still be tested against the use-value of comparable properties, enabling the private sector to negotiate has resulted in the process taking on

more free market characteristics. Partly as a result of this, as well as a form of "free market creep" resulting from other minor reforms enacted throughout much of the 21st century, rents rose by 19% between 2008-2018, a markedly higher increase than previous decades – which can be compared to 9% CPI inflation over the same time period (SCB 2018). Hence, migration having an impact on rents is a highly feasible relationship(?). In 2019, a government proposal which would enable rents to be set completely freely among newly produced housing was introduced and is currently undergoing investigation. This could substantially transform the housing market in the long-term.

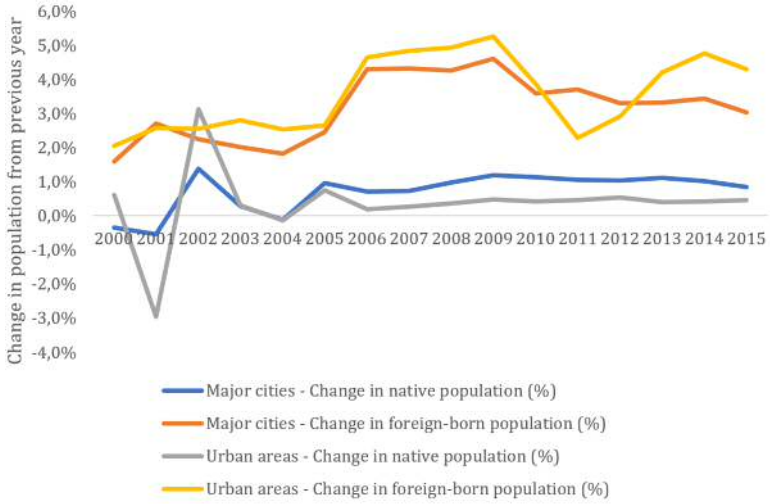
Additional complexities arise owing to the rental queue system in Sweden, where rental properties are not advertised on the free market. Instead, a queuing mechanism exists, with properties allocated to people in the queue as they become available. Nevertheless, internal and international migration have been shown to impact this rental queue (Tyrcha 2019), and thus, it is feasible that migration could impact rental levels, too. Sweden is a particularly interesting case to study, owing to the widely unexplored, regulated or semi-regulated nature of the rental market. The aforementioned deregulatory reform means that the impacts of migration on the rental market could now feasibly be stronger than in past decades – though, indeed, even the previously more regulated nature of the market does not necessarily preclude migration's impact on rents.

The mechanism by which migrants create an impact on rents is by and large the same as in an unregulated rental market. An inflow of migrants causes upward pressure on rents, given constant supply and ceteris paribus, as demand for rental housing exceeds supply of rental housing. Under the new 2011 regulations, this mechanism is allowed to function in a more unimpeded way than previously, as rental housing owners are more able to adjust rents upon rental review than previously (though remain subject to certain use value restrictions which limit the permissible annual growth rate of rents, as well as the absolute value of rents, relative to other comparable properties). However, even under the regulated system, public sector owners would often feel forced to adjust rents a little more than the average in a given location, depending on demand, owing to outward pressures placed on them from governments or other public and private actors. Reasons for this could include the desire to combat the parallel black market, or limit the size of the rental queue by increasing turnover of apartments (Bengtsson 1994, Lind 2001, SCB 2019). Such impacts could, both before and after the reform, manifest partly owing to the aforementioned ability of certain migrant groups to bypass the rental queue and thus indirectly impact the rental market. The impact of such heterogeneities, and the subsequent differences in impact of different migrant groups, are discussed further in Section 2.3 below.

Nevertheless, given the unique institutional context, it also must be seen as relatively unlikely that the Swedish rental market will see the same impacts as markets in less stringently regulated contexts. The relationship between migration and rental levels on the unique Swedish rental market will be explored further in this paper.

2.3 The Role of Heterogeneities

Meanwhile, it is important to also acknowledge directly that the human capital differences among migrant groups could indirectly serve to influence their motivations to migrate, and thus also their ultimate impact on the housing market. Eichholtz, Lindenthal (2014) find that in the context of domestic migration in England, human capital is a key driver of housing demand. To some degree, it could therefore also be theorized that human capital may be an underlying factor influencing the scale of impact of both internal and foreign-born migration on house prices. This is despite evidence of internal migration constituting an interesting migration flow which differs substantially in nature to foreign-born migration, as well as an abundance of evidence of the impacts of internal migration on the labour market (e.g. Friedberg, Hunt 1995, Borjas 2006, Hammarstedt, Palme 2006, Gerdes, Wadensjö 2010, Kerr, Kerr 2011, Dustmann, Frattini 2014), internal migration is very rarely included or analysed in the housing market context (the only paper that does this comprehensively is Wang et al. (2017), in China). Given the size and variation of internal migration flows, it appears clear that inclusion of this variable, separate to foreign-born migration, in analysis would be advisable. The heterogeneity of the variable



Source of data: SCB, 2018

Figure 2: Change in population in different areas over time

and wide differences when compared to foreign-born migration in the descriptive statistics in Figure 2 mean it is likely that the impacts of internal migration could differ to foreign-born migration. As a result, the two forms of migration will be analysed separately in this paper.

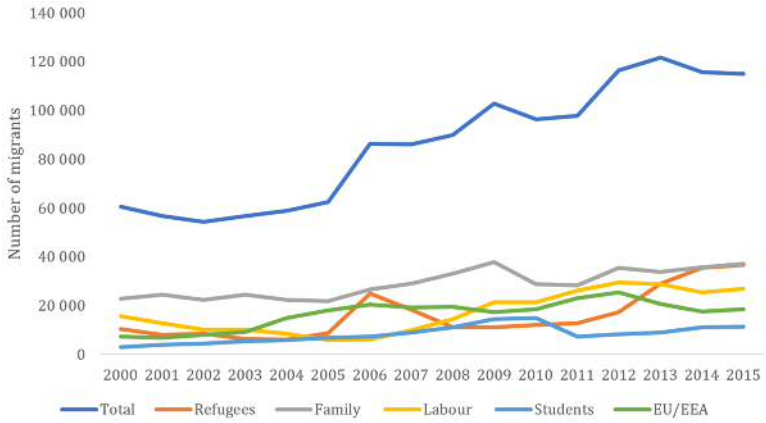
To investigate further along these lines, Figure 2 shows the differences in migration trends between major cities and smaller urban areas, for different kinds of migrants. This is a relevant distinction to make owing to the vastly differing urban characteristics of major cities and smaller urban areas in Sweden. The former consists of three large metropolitan areas (Stockholm, Gothenburg, Malmö), the smallest of which has a population of over 700,000, while the smaller urban areas in Sweden can have populations as small as 50,000 (SCB 2019). Further, the different types of areas are likely to attract different kinds of migrants, owing to differences in characteristics as a result of the size of regional labour markets, demographical differences, migrant preferences, differences in the nature and likelihood of the presence of certain network effects, and more (SCB 2019).

Figure 2 shows migration trends in Swedish major cities and urban areas over time. It is clear that aside from a few anomalous years in the early 2000s, native people have behaved relatively similarly over time in major cities and urban areas, while foreign-born migrant settlement patterns have varied a little more. Most clear, however, is that the migration flows of natives compared to foreign-born migrants are very different in nature when compared to one another.

Looking more closely at the different dimensions of foreign-born migration also reveals a number of interesting trends. These are shown in Figure 3.

Figure 3 shows the different spikes in foreign-born migration over time since 2000. As can be seen, migration has been rising relatively steadily over the studied period, but with identifiable spikes in different forms of migration in different years. These include a spike in refugee migration around 2006 and from 2012 onwards, as well as spikes in family reunification migration, labour migration, student migration and EU migration in different years.

Among these foreign-born migrant groups, further heterogeneity arises. In the studied period, between 60 to 80% of refugees (depending on the year) have not been able to obtain their own housing, but have instead been allocated housing by the government (SCB 2018). The government has mandated for all municipalities to provide refugees with housing, with requirements for each municipality depending on a complex set of criteria, including labour market opportunities and housing availability (Wennström, Öner 2015). As a result of this, refugees are unlikely to push up rental levels on their own, but instead will have an indirect impact on rents, as municipalities must find housing for refugees.



Source of data: Migrationsverket, 2018

Figure 3: Foreign-born migration to Sweden by reason for migration

In this sense, refugees are able to bypass the aforementioned rental queue, and instead indirectly impact the rental market by lowering the supply of rental properties. Despite the lack of a direct impact, and the inability of refugees to spend much on housing, the indirect impact, with migration ultimately impacting rents through an indirect mechanism, is clear.

Meanwhile, family reunification migrants, primarily associated with labour migrant and refugee flows, also present a unique form of migration flow. These forms of migrants are expected to find their own housing, but may have a diminished impact on the rental market as they are inherently less likely to have to attain their own housing, owing to the family links that have allowed them into Sweden. As a result, such migrants may often live at relatively higher densities than other migrants (Migrationsverket 2018), and therefore diminish the impacts that the wider foreign-born group may have on the housing market. Foreign-born labour migrants, meanwhile, are left to their own devices and impact the rental or housing markets in the traditional free market way, as do internal migrants. They may thus also have a stronger impact, owing to their relative willingness and ability to spend, though they are also more able to choose other, more expensive forms of housing. This may in turn limit their impact on relatively more affordable housing types such as rental housing. Hence, it is clear that broadly speaking, the impacts created by foreign-born and internal migrants may differ widely, owing to the wide heterogeneity between and within the two groups. This paper also aims to investigate this.

3 Data and Methodology

3.1 Data Overview and Initial Analysis

SCB (2018), the Swedish statistical agency, has provided data relating to rents, migration-related variables, as well as all controlled variables. The rental data constitutes the mean rent per square meter in January each given year, on the municipal level. The data comes from Sweden's largest, publicly operated statistical agency and thus can be considered reliable. However, this data is only available for a select number of municipalities. Complete data is available for 55 municipalities, and thus, 55 municipalities will be used in the rental analysis.

The model will take the following form:

$$\Delta \ln(r)_{k,t} = \alpha + \theta \frac{\Delta\text{foreign-born}_{k,t}}{\text{population}_{k,t-1}} + \theta \frac{\Delta\text{Swedish-born}_{k,t}}{\text{population}_{k,t-1}} + \beta_1 \ln(In_{k,t-1}) + \beta_2 Em_{k,t-1} + \beta_3 \ln(T_k) + \beta_4 \ln(B_k) + \beta_5 A_k + \beta_6 NS_{k,t-1} + \beta_7 L_{k,t-1} + \sum_{t=2}^{16} \delta_t Y_t + \epsilon_{k,t} \quad (1)$$

where $\ln(r)$ is the natural logarithm of the rent, Δforeign-born$_{k,t}$ is the change in the foreign-born population in location k between $t-1$ and t, ΔSwedish-born$_{k,t}$ is the change in Swedish-born (internal) migration in location k between $t-1$ and t, where Swedish-born rather than all internal migration is used to avoid double counting of foreign-born migrants. By population$_{k,t-1}$ we denote the total population in location k and time $t-1$, In is real income, Em is employment, T is a temperature average from 1961-1990, B is % of population with a bachelor's degree in 1984 per 10,000 inhabitants, A is the percentage of the population aged 20-64 in 1984, NS is the new housing stock that becomes available in every year per 1,000 residents (a supply-side variable), L is a dummy variable allowing me to account for the differing impacts of new Planning and Building Legislation that took effect in 2011 (another supply-side variable) where 1 = 2000-2010 and 0 = 2011-2015, Y is years from 2000 to 2015 $t = 2, \ldots, 16$, and ϵ is the error term.

Following Saiz (2007), a shift-share instrumental variable approach will be taken, to control for potential endogeneity in the data. The approach makes use of migrant settlement patterns in the past, which tend to influence the settlement patterns of current migrants, but not be correlated with local economic conditions in the present, since they are based on exogenous push factors in the country or region of origin. The approach will take the following form:

$$\overline{\Delta\text{foreign-born}_{k,t}} = \overline{\Delta\text{foreign-born}_{Sweden,t,o}} * \frac{\overline{\Delta\text{foreign-born}_{k,1984}}}{\overline{\Delta\text{foreign-born}_{Sweden,1984}}} \quad (2)$$

$$\overline{\Delta\text{Swedish-born migration inflow}_{k,t}} = \overline{\Delta\text{Swedish-born migrations}_{Sweden,t}} * \frac{\overline{\Delta\text{Swedish-born migration inflow}_{k,1984}}}{\overline{\Delta\text{Swedish-born migrations}_{Sweden,1984}}} \quad (3)$$

where Δforeign-born$_{k,t}$ is the change in the foreign-born population in location k between $t-1$ and t, and ΔSwedish-born migrations$_{k,t}$ is the change in Swedish-born (internal) migration in location k between $t-1$ and t. Where the year listed is 1984, the variable refers to the share of the respective migrant population in that year.

The instrumental variable approach relies on settlement patterns in 1984, the earliest year for which data is available, to attempt to predict current settlement patterns among migrant groups. As such, it functions well as an instrument, since the past settlement patterns are correlated with current migration flows, owing to network and path dependency effects, but not with current rental values[1]. The approach assumes a wide variation in the composition and settlement patterns of different migrants over the studied period, which as noted in the earlier sections, is met in the case of Sweden.

Complete data is available from 2000 to 2015. As a robustness check, and to ensure that any migration effects being picked up are truly measuring the impact of the migration variable, the regression will first be run for 2000-2015, but then also run for 2000-2010 and 2011-2015, separately. This is because the deregulatory reform, detailed in Section 2, should mean that migration is likely to have a stronger impact on rents after 2011, while impacts before 2011 should be smaller. A structural chow test will be used to test whether a structural break does indeed manifest in 2011.

[1]The instrument is also found to produce sufficiently sized Cragg-Donald and Kleibergen-Paap F statistics.

3.2 Regional Analysis

Following the above regression, the municipalities will be classified based on their characteristics and ran in separate regressions. This should enable identification of potential differential effects between municipalities. This analysis will take the following form:

- Major Cities (24 municipalities)
- Urban Areas (25 municipalities)

This split is conducted owing to the descriptive statistics shown in Figure 2. As outlined in association with that diagram, the difference in the population size of major city and urban municipalities is substantial. Further, access to labour markets is much stronger in major city municipalities, as well as the ability to commute across municipalities for employment purposes (where as in smaller urban areas, this is much more condensed to the home municipality) (SCB 2019). As such, there are a number of fundamental differences between major cities and smaller urban areas that make these interesting for study separate to one another.

3.3 Reason for Migration Analysis

Finally, in one last regression, the migration-related variables will be broken down, based on reason for migration. This will also enable identification of differential impacts between migrants of different origins. This will take the following form:

- Labour migration (labour migrants, EU migrants, students)
- Family reunification migration
- Refugee migration

In this paper, I elect to distinguish between foreign-born and internal migration, owing to the differences seen in Figure 3 in terms of migration patterns, as well as the broad differences between different forms of foreign-born and internal migration. It is valuable to also break down foreign-born migration into the above groups, partly owing to the different experiences that these groups are likely to be faced with on the rental market, as outlined in Section 2.3. Since it is clear that migration patterns have varied broadly for different migration groups in major cities and urban areas, it is also interesting to note whether corresponding differences can be seen in the impacts of different forms of migration on rents.

4 Results

The results of the first regression are shown in Table 1. It shows a relatively weak, but significant impact of foreign-born migration on rents, significant at the 5% level, producing a coefficient of 0.249 in the IV regression. The fact that there is a significant, positive impact is relatively noteworthy, given that the largely regulated rent structure means the dependent variable is not able to vary freely, yet migration is still able to impact rental levels, even if on a diminished scale when compared to previous literature. This indicates that migration may be having a relatively large impact on rents in Sweden, compared to other factors. Nevertheless, other controlled variables such as income and new stock are also significant, indicating that it is not just migration alone that is driving up the cost of rent. Concurrently, the fact that Swedish-born internal migration is not significant at all is also noteworthy. These results are very interesting as they seem to indicate wide heterogeneity in the degree of impacts that different forms of migrants have on the rental market in Sweden. This could stem from relative interest in the rental market (as compared to other markets), as well as relative willingness and ability to spend. Swedish-born migrants, who generally have higher wealth than foreign-born migrants, may be relatively less interested in the rental market, while foreign-born migrants may be forced into this market to a higher degree. Further, preferences between the two groups may vary, as the latter group values the flexibility offered by the rental market more than the former group, resulting in a higher willingness to spend on rental accommodation.

Table 1: The relationship between migration and rental levels

	OLS	IV
ΔForeign-born$_t$/Population$_{t-1}$	0.236**	0.249**
	(0.107)	(0.109)
ΔSwedish-born$_t$/Population$_{t-1}$	0.090	0.102
	(0.071)	(0.110)
Log income$_{t-1}$	0.077***	0.094***
	(0.028)	(0.031)
Employment$_{t-1}$	0.041	0.044
	(0.037)	(0.038)
Log January temperature	0.000	-0.001
	(0.000)	(0.002)
New stock$_{t-1}$	0.021**	0.022**
	(0.009)	(0.009)
Legislation	-0.001	-0.005
	(0.002)	(0.007)
Year fixed effects	Yes	Yes
Region fixed effects	Yes	Yes
Observations	880	880
R-Squared	0.224	0.226

Notes: * ... significant on the 10%, ** ... 5%, *** ... 1% level.

To verify whether the effect being picked up is truly that of migration, and not some other effect that is driving up house prices, I conduct separate regressions for the data from 2000-2010, and for data from 2011-2015 (as detailed in Section 3.1). The results are shown in Table 2.

Taken at face value, the results in Table 2 appear to indicate that the migration variable's impacts are indeed stronger between 2011-2015, while impacts seen between 2000-2010 are decidedly weaker (though not non-existent). Indeed, the foreign-born variable produces a coefficient of 0.254, significant at the 1% level, while the impact between 2000 and 2010 is 0.180, significant only at the 10% level. The income variable is also less significant between 2000-2010, while new stock is not significant at all. The differences in these results would suggest that a structural break is likely. The chow test confirms this – producing an f-value of 9.372, which is larger than the critical value of 1.5987. The p-value is also 0.000, further confirming the strength of these results. This would seem to indicate that the regressions are indeed capturing the impacts of foreign-born migration.

Having established the above, Table 3 allows me to study these trends in further detail, looking at these municipality characteristics. The regressions are only run from 2000-2015, in order to improve sample size. Also this table shows relatively few significant variables. Only foreign-born migration is significant in major cities, with a coefficient of 0.234 significant at the 10% level. This is consistent with Table 1 displayed earlier, and confirms that the theory regarding the generally stronger impact of foreign-born migration on the rental market, as well as generally, may hold true, at least in major cities. As such, it appears that foreign-born migration has a particularly strong impact on rents in major cities. In smaller urban areas, neither migration variable is significant, which could be explained by less pressure, in absolute terms, being placed on the rental market, which is less constrained owing to a larger amount of space and resources in smaller urban areas. Rental housing could perhaps also be less demanded owing to relative preferences toward other private housing cooperatives and owner-occupied housing, which is relatively more readily available than in major cities. However, the role of migrant preferences when looking for housing and the intersection between this and other relevant trends, such as economic opportunities and the role of small-town revival, could also be playing into the less significant impacts.

Adjusting policy to reflect the fact that foreign-born migration is likely to push up

Table 2: The relationship between migration and rental levels over different time periods

	2000-2015		2000-2010		2011-2015	
	OLS	IV	OLS	IV	OLS	IV
ΔForeign-born$_t$/Population$_{t-1}$	0.236**	0.249**	0.180*	0.195*	0.254***	0.284***
	(0.107)	(0.109)	(0.104)	(0.109)	(0.088)	(0.093)
ΔSwedish-born$_t$/Population$_{t-1}$	0.090	0.102	0.024	0.039	0.018	0.015
	(0.071)	(0.110)	(0.140)	(0.144)	(0.087)	(0.094)
Log income$_{t-1}$	0.077***	0.094***	0.056*	0.057*	0.162***	0.166***
	(0.028)	(0.031)	(0.031)	(0.031)	(0.041)	(0.041)
Employment$_{t-1}$	0.041	0.044	0.012	0.021	0.074	0.060
	(0.037)	(0.038)	(0.065)	(0.068)	(0.103)	(0.105)
Log January temperature	0.000	-0.001	0.002	0.002	0.000	0.000
	(0.000)	(0.002)	(0.004)	(0.004)	(0.001)	(0.001)
New stock$_{t-1}$	0.021**	0.022**	0.009	0.010	0.027**	0.029**
	(0.009)	(0.009)	(0.021)	(0.021)	(0.010)	(0.011)
Legislation	-0.001	-0.005				
	(0.002)	(0.007)				
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	880	880	605	605	275	275
R-Squared	0.224	0.226	0.203	0.204	0.275	0.277

Notes: * ... significant on the 10%, ** ... 5%, *** ... 1% level.

rental values more so than internal migration could be advisable in certain areas. Access to the rental market is vital, as it is often the first port of call for many more vulnerable groups or individuals in society, who lack access to the considerable funds required to access the private housing cooperate or owner-occupied housing market. Hence, targeted initiatives to avoid such vulnerable groups struggling to gain access to the rental market should perhaps be considered in response. This could include initiatives to alleviate pressure on the rental market, by e.g. encouraging more building, or instituting a targeted queue-jumping scheme. The results in Table 3 highlight that any policy adjustments targeted at the rental market should likely be focused to major cities, and particularly those which have received the largest relative influxes of foreign-born migrants.

Finally, to shine further light on any trends and potential requirement for policy adaptations, I look at the impacts of migration flows segmented by migrant background on rental levels in Table 4. In this table, I note generally weak or no impacts of migration flows on rents. The only significant impacts are produced by refugee migration, with 0.287 significant at the 10% level overall, and 0.374 significant at the 5% level in major cities. The lack of significance for other variables suggests broad heterogeneity in migrant impacts, perhaps owing to the system of rent regulation that is in place. It would, however, appear that refugee migrants are capable of creating a substantial shock to the rental market. This could be a result of a lack of competition stemming from other groups for this type of housing, with other forms of migrants instead favouring other forms of housing where possible.

It is also likely that the Swedish government's refugee placement policy is contributing to the impacts produced by refugees. The policy consists of mandating that all Swedish municipalities provide housing to a certain number of refugees (Wennström, Öner 2015), and this housing stock is likely to be taken from the rental market. Hence, although refugees may not directly be causing pressure on the rental market, indirect impacts could be resulting in refugees appearing to be the most impactful group on this market, as municipalities reserve rental housing for refugees, causing pressure on the rental market. A natural conclusion is to focus initiatives which alleviate pressure on the rental market to areas which have received an influx of refugees, particularly in major cities, where other migrants or natives may be pushed out of or struggle to gain access to the rental market. However, further initiatives could include looking into changing the allocation pattern of refugees and encouraging municipalities where the rental market has not been as

Table 3: The relationship between migration and rental levels in different types of municipalities

	Major Cities		Smaller Urban Areas	
	OLS	IV	OLS	IV
ΔForeign-born$_t$/Population$_{t-1}$	0.216*	0.234*	0.288	0.361
	(0.116)	(0.121)	(0.305)	(0.333)
ΔSwedish-born$_t$/Population$_{t-1}$	0.144	0.121	0.151	0.114
	(0.123)	(0.136)	(0.199)	(0.233)
Log income$_{t-1}$	0.034	0.036	-0.067	-0.079
	(0.081)	(0.099)	(0.098)	(0.145)
Employment$_{t-1}$	0.032	0.023	0.072	0.082
	(0.037)	(0.034)	(0.079)	(0.077)
Log January temperature	0.001	0.002	0.001	0.002
	(0.002)	(0.002)	(0.001)	(0.002)
Percentage with bachelor's degree (1984)	0.023	0.026	-0.081	-0.099
	(0.024)	(0.027)	(0.071)	(0.074)
Percentage working age (1984)	-0.019	-0.005	0.112	0.116
	(0.034)	(0.021)	(0.091)	(0.097)
New Stock$_{t-1}$	0.014	0.015	0.025	0.027
	(0.016)	(0.017)	(0.029)	(0.029)
Legislation	-0.011	-0.016	-0.016	-0.017
	(0.027)	(0.029)	(0.033)	(0.034)
Year fixed effects	Yes	Yes	Yes	Yes
Observations	384	384	400	400
R-Squared	0.191	0.192	0.280	0.282

Notes: * ... significant on the 10%, ** ... 5%, *** ... 1% level.

affected by migration to take in a larger share of refugees in future. This is a controversial initiative, though, and would have to be weighed against the wider societal impacts which such policy could have, e.g. through cost-benefit analysis.

Further, I note that the impacts of endogeneity appear to be fairly limited for the rental market. The only significant variables for the rental market appear to be affected by endogeneity to some degree – but the impacts are being very slightly underestimated, rather than overestimated, yet results do not appear to be overly affected by this.

5 Conclusions

In this paper I examine the impacts of foreign-born and internal migration on house prices on the rental market in Sweden. This extends the literature by analysing the effects of migration on a subset of the housing market that is highly unique, owing to its regulated nature, and has not been studied previously. Analysis is disaggregated on the regional level, and special emphasis is also placed on different forms of migration, including internal migration and a number of subsets of foreign-born migration. This allows one to highlight the impacts, or lack thereof, that a diverse range of migrant groups have on the rental market.

The results indicate a generally limited impact of migration on the rental market in Sweden – somewhat expected, given the regulated nature of the market. However, recent deregulation moves could be contributing to the fact that some significant impacts are found, for foreign-born migration, in particular. Indeed, disaggregation on the regional level, as well as by reason for migration, shows that foreign-born migration consisting of refugees in particular appears to be impactful in terms of housing rents in Sweden, primarily in major cities.

This is likely to be at least partly a result of the Swedish government's refugee placement policy, mandating that municipalities accept refugees regardless of the availability of housing. However, the rental market being most readily accessible to this group is also likely to play a role, pushing up prices for other groups wishing to access the market. This could have significant societal impact, as many vulnerable and less wealthy groups rely

Table 4: Results by Reason for Migration

	Overall		Major Cities		Smaller Urban Areas	
	OLS	IV	OLS	IV	OLS	IV
ΔLabour migration$_t$/	0.188	0.286	1.388	1.144	0.618	0.733
Population$_{t-1}$	(0.511)	(0.736)	(1.743)	(1.905)	(1.344)	(1.905)
ΔFamily reunification migration$_t$/	0.300	0.219	0.188	0.195	0.700	0.411
Population$_{t-1}$	(0.508)	(0.671)	(0.344)	(0.355)	(1.390)	(1.739)
ΔRefugee migration$_t$/	0.246*	0.287*	0.388**	0.374**	0.301	0.236
Population$_{t-1}$	(0.133)	(0.155)	(0.166)	(0.167)	(0.361)	(0.415)
ΔSwedish-born$_t$/Population$_{t-1}$	0.087	0.099	0.131	0.167	0.181	0.194
	(0.088)	(0.101)	(0.104)	(0.133)	(0.139)	(0.166)
Log income$_{t-1}$	-0.066	-0.078	-0.114	-0.117	-0.158	-0.167
	(0.040)	(0.057)	(0.076)	(0.077)	(0.071)	(0.086)
Employment$_{t-1}$	-0.041	-0.051	-0.011	-0.015	-0.107	-0.151
	(0.082)	(0.097)	(0.073)	(0.078)	(0.148)	(0.161)
Log January temperature	0.001	0.001	0.002	0.002	0.002	0.002
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.002)
Bachelor's degree (%, 1984)	0.016	0.009	0.021	0.025	-0.073	-0.087
	(0.022)	(0.017)	(0.026)	(0.028)	(0.073)	(0.075)
Working age (%, 1984)	-0.021	-0.028	-0.010	-0.015	0.123	0.137
	(0.028)	(0.035)	(0.036)	(0.039)	(0.093)	(0.099)
New stock$_{t-1}$	0.015	0.012	0.015	0.017	0.025	0.027
	(0.014)	(0.017)	(0.016)	(0.016)	(0.029)	(0.031)
Legislation	-0.001	-0.002	-0.003	-0.004	-0.002	-0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	880	880	384	384	400	400
R-Squared	0.150	0.150	0.145	0.146	0.244	0.244

Notes: * ... significant on the 10%, ** ... 5%, *** ... 1% level.

on the rental market, unable to access other forms of housing markets. Hence, targeted initiatives to alleviate pressure on the rental market, particularly in major city areas which have been most affected by refugees and other migrants accessing the rental market, could be advisable. In addition, an adjustment to future refugee allocation policy to reflect the above trends and alleviate pressure further could be investigated. Further research could investigate these trends further, as well as look at the impacts of migration on less regulated markets, particularly in a disaggregated manner, as in this paper.

References

Aitken A (2014) The effects of immigration on house prices and rents: Evidence from England and Wales. Royal Holloway, University of London. Mimeo

Bengtsson B (1994) Swedish rental policy – A complex superstructure with cracking foundations. *Scandinavian Housing and Planning Research* 11: 182–189. CrossRef.

Borjas G (2006) Native internal migration and the labor market impact of immigration. *Journal of Human Resources* XLI: 221–258. CrossRef.

Degen K, Fischer AM (2009) Immigration and Swiss house prices. CEPR Discussion Paper No. DP7583

Dustmann C, Frattini T (2014) The fiscal effects of immigration to the UK. *The Economic Journal* Feature Issue: F593–F643. CrossRef.

Eichholtz P, Lindenthal T (2014) Demographics, human capital, and the demand for housing. *Journal of Housing Economics* 26: 19–32. CrossRef.

Friedberg R, Hunt J (1995) The impact of immigrants on host country wages, employment and growth. *The Journal of Economic Perspectives* 9: 23–44. CrossRef.

Gerdes C, Wadensjö E (2010) Post-enlargement migration and labor market impact in Sweden. In: Kahanec M, Zimmermann K (eds), *EU Labor Markets After Post-Enlargement Migration*. Springer, Berlin Heidelberg, 163–179. CrossRef.

Gonzalez L, Ortega F (2012) Immigration and housing booms: Evidence from Spain. *Journal of Regional Science* 53: 37–59. CrossRef.

Hammarstedt M, Palme M (2006) Intergenerational mobility, human capital transmission and the earnings of second-generation immigrants in Sweden. IZA Discussion Paper

Kerr S, Kerr W (2011) Economic impacts of immigration: A survey. NBER working paper. CrossRef.

Kürschner K (2017) Immigration and rental prices of residential housing: Evidence from the fall of the Berlin wall. Verein für Socialpolitik, annual conference 2017 (Vienna)

Latif E (2015) Immigration and housing rents in Canada: A panel data analysis. *Economic Issues* 20: 91–108

Lind H (2001) Rent regulation: A conceptual and comparative analysis. *European Journal of Housing Policy* 1: 41–57. CrossRef.

Migrationsverket (2018) Selected statistics. https://www.migrationsverket.se/Om-Migrationsverket/Statistik.html

Mussa A, Nwaogu UG, Pozo S (2017) Immigration and housing: A spatial econometric analysis. *Journal of Housing Economics* 35: 13–25. CrossRef.

Ottaviano GIP, Peri G (2005) Rethinking the gains from immigration: Theory and evidence from the U.S. NBER Working Paper. CrossRef.

SABO – Sveriges Allmännyttiga Bostadsföretag (2011) Allmännyttan. https://www.allmannyttan.se/historia/tidslinje/hyresratten-och-lagstiftningen/

Saiz A (2003) Room in the kitchen for the melting pot: Immigration and rental prices. *The Review of Economics and Statistics* 85: 502–521. CrossRef.

Saiz A (2007) Immigration and housing rents in American cities. *Journal of Urban Economics* 61: 345–371. CrossRef.

SCB – Statistics Sweden (2017) Selected statistics. http://www.scb.se/hitta-statistik/

SCB – Statistics Sweden (2018) Selected statistics. http://www.scb.se/hitta-statistik/

SCB – Statistics Sweden (2019) Selected statistics. http://www.scb.se/hitta-statistik/

Sharpe J (2015) Three essays on the economic impact of immigration. UKnowledge, University of Kentucky

Tumen S (2016) The economic impact of syrian refugees on host countries: Quasi-experimental evidence from Turkey. *American Economic Review* 106: 456–60. CrossRef.

Tyrcha A (2019) Why does the queue keep growing? The relationship between migration and rental housing queues in Sweden. *Economics Bulletin* 39: 1251–1258

Tyrcha A, Abreu M (2019) Migration diversity and housing prices – evidence from Sweden. https://ssrn.com/abstract=3394234

Wang X, Hui E, Sun J (2017) Population migration, urbanization and housing prices: Evidence from the cities in China. *Habitat International* 66: 49–56. CrossRef.

Wennström J, Öner (2015) Den geografiska spridningen av kommunplacerade flyktingar i Sverige. http://www.nationalekonomi.se/sites/default/files/NEFfiler/43-4-jwöö.pdf

World Bank (2019) Rental housing. http://documents.worldbank.org/curated/en/810681-468339259949/pdf/Rental-Housing-lessons-from-international-experience-and-policies-for-emerging-market.pdf

REGION
The Journal of ERSA
Powered by WU

Volume 7, Number 1, 2020, 49–70
DOI: 10.18335/region.v7i1.261

journal homepage: region.ersa.org

Measures of labour market accessibility. What can we learn from observed commuting patterns?*

Arnstein Gjestland[1], Liv Osland[1] and Inge Thorsen[1]

[1] Western Norway University of Applied Sciences

Received: 13 February 2019/Accepted: 28 February 2020

Abstract. It is well known that measures of labour market accessibility explain spatial variation in housing prices, even in markets with polycentric labour market structures. This paper examines whether data on observed commuting patterns can replace or supplement gravity-based measures representing the commuting potential at specific locations. We use data from a region in Western Norway, and we find that measures based on observed commuting flows and commuting time cannot replace a gravity-based measure of labour market accessibility. Based on, inter alia, the spatial Durbin estimator we find that measures of observed commuting flows increase the explanatory power of a hedonic house price model.

1 Introduction

The relationship between house prices and access to workplaces is a central theme in both theoretical and empirical housing market research. There are many reasons why this relationship is important. Travelling to work is a regular and bounded trip. According to Vågane et al. (2011), travelling to work constitutes 18 per cent of all travels in Norway. During workdays, approximately 25 per cent of the travels are journey to work. Most commuters travel at the same point in time every day, which also creates congestion. Road transportation infrastructure is often given a capacity to deal with such traffic peaks. At the same time, investments in the road network affect labour market accessibility, which is in turn capitalized into house prices. In order to reduce many of the transportation problems related to commuting, planners may seek to locate houses in areas where job accessibility is assumed to be high.

According to Handy, Niemeier (1997), there is no consensus in the literature on a good measure of accessibility. In explaining housing prices, gravity-based accessibility measures have been suggested as a generalization of modern polycentric labour market structures. Although more recent research has shown that gravity-based accessibility measures explain significant spatial variation in housing prices, Handy, Niemeier (1997) show that the gravity-based measures are not the only weighted measures that can be used to capture the job opportunity density of a given area.

Gravity-based measures of labour market accessibility reflect the potential for commuting from a specific residential area. In this paper we introduce three other measures, based on actual commuting patterns. One measure is origin specific: the percentage of the total working population living in the zone and working in a different zone. The second measure

*We are grateful to two anonymous reviewers for very helpful comments.

is destination specific: the percentage of people working in the zone that are living in another zone. The third is based on the calculation of actual commuting time in each zone. The proposed measures are related to labour market accessibility. Our ambition has been to test empirically whether the measures can replace a gravity-based measure of labour market accessibility, or whether they should supplement such a potential measure, adding relevant information on spatial characteristics in order to explain housing prices.

The measures based on observed commuting flows have some advantages relative to the potential measures of job opportunities. The new measures can be more easily explained to non-experts in the field, and they are computationally simpler than the gravity-based measures, which may involve non-linear methods of estimation. Some of the measures of observed commuting flows are also less demanding in terms of data requirements, for example, travelling times and the transportation network.

Simplicity and data requirements are not, however, the most important issues in favour of incorporating measures based on observed commuting flows. Such measures potentially offer a kind of market-based evaluation of characteristics relevant for explaining housing prices. The values of a gravity-based measure represent the potential of making favourable labour market decisions, in terms of the traveling time between the residential location and the job location. In contrast, the other measures we consider are based on labour market decisions that have actually been made. Our basic hypothesis is that locations offering favourable labour market opportunities capitalize into the housing market, see e.g. Gjestland et al. (2014). A main motivation of this paper is to study whether observation-based information can substitute, or maybe supplement, the gravity-based potential measure in explaining spatial variation of housing prices. According to Handy, Niemeier (1997), "The fundamental issue is that an accessibility measure is only appropriate as a performance measure if it is consistent with how residents perceive and evaluate their community. In other words, a practical definition of accessibility must come from the residents themselves." (p. 1176). In view of this citation, estimation of hedonic house prices can be a useful tool. Assuming market equilibrium, this method can be characterized as a revealed preference approach. It enables the measurement of the implicit prices of goods and amenities that are not directly traded in any markets. As such, this approach can be an appropriate framework for evaluating how alternative measures of potential and observed labour market interaction contribute to explain house prices.

The paper is outlined as follows. Section 2 contains a brief literature review. Thereafter, in Section 3, we present the study area and formulate explicit hypotheses to be tested. In Section 4 we present the data while the empirical results and the analyses are presented in Sections 5–7. Finally, conclusions based on our findings follows in Section 8.

2 A brief literature review

The most widely accepted theory that links residential location to the price of housing is given by urban economic theory represented by the monocentric city model. The relevant prediction of this model is that households living far from the centre of employment are compensated for higher costs of commuting by way of a lower price for housing.

In the housing market literature, accessibility has traditionally been accounted for by the simple measure of distance to the central business district (CBD) (see e.g., Ball, Kirwan 1977, Dubin 1992). It is, however, well acknowledged in the literature that the monocentric model frequently has not been supported by empirical evidence. Many reasons have been suggested for the disparity between theory and empirical results. One obvious suggestion is the polycentric pattern of employment (Anas et al. 1998). In spite of this, there are in fact relatively few papers that focus on how polycentrism may affect property values. One natural suggestion to cope with polycentrism is found in Waddell et al. (1993), who include both the distance to the CBD and the distances to secondary employment centres. One potential problem with this approach is that the researcher has to choose which employment centres to include. Because of problems with spatial multicollinearity and interpretation of partial effects, it may not be straightforward to include distances to many employment nodes as separate variables in an empirical hedonic

house price model. See, however, Heikkila et al. (1989) for possible ways of dealing with this issue.

The potential, gravity-based, measures of accessibility (Handy, Niemeier 1997) are frequently used in the literature. To cite Anselin (2002, p. 250), these variables are specified so that "the potential for interaction between an origin i and all destinations j was formulated as a sum of 'mass' terms in the destination, suitably downscaled by a distance decay function". For a useful general discussion on the use of the accessibility concept in spatial analysis, see Kwan et al. (2003). Farber et al. (2013) discuss metrics based on the time-geographical concept of joint accessibility for measuring the spatial interaction potential of a region. However, in this paper, we focus on the use of gravity-based accessibility measures that have been suggested as a generalization of modern polycentric labour market structures (Heikkila et al. 1989). Nevertheless, there are not many papers that relate gravity-based accessibility measures to housing prices.

The evaluation of gravity-based measures differs in the literature. Jackson (1979) does not find evidence of the dominance of either the CBD-gradient or the gravity-based employment index. Adair et al. (2000) find heterogeneous results. In the overall Belfast Urban Area, the gravity-based accessibility had small or negligible effect, while stronger effects were found by repeating the analysis at the sub-market level. By using Norwegian housing price data from a wider labour market area, the accessibility measure was clearly significant in Osland, Thorsen (2008). Osland, Pryce (2012) use housing price data from Glasgow. The employment data were from all Scottish data zones. They found a highly significant non-monotonic relationship between house prices and access to employment. According to this research, house prices would fall as we move very close to an employment node if there are significant negative externalities from the firms located at the employment node. The result that there are negative externalities related to high levels of accessibility is in line with results found in Li, Brown (1980), although this paper measures access to employment by way of distance to the CBD. Ahlfeldt (2011) studies land prices and finds that a gravity-based accessibility measure can explain residential land prices. According to this paper, the measure is able to disentangle positive accessibility effects from negative congestion effects related to transportation infrastructure.

3 Study area and hypotheses to be tested

Our study area is situated in the south-west of Norway. The population is approximately 230 000, most of it concentrated in the north-western corner in the twin cities of Stavanger and Sandnes. Because of natural barriers, the study area is clearly delimited from neighbouring markets. This is also reflected in data on commuting flows, and contributes to making the market appropriate for an empirical analysis of the relationship between observed commuting flows, labour market accessibility, and spatial variation in housing prices.

This paper presents results from a regression model where the price of homogenous housing at a given location is related to a range of variables. These variables are either structural variables related to the house itself, or to its specific location in the geography as follows:

$$P_{it} = f(z_{sit}, z_{lit}) \tag{1}$$

where P_{it} is the price of house i in year t, z_{sit} is the value of structural dwelling-specific attributes, and z_{lit} represents location-specific attributes.

In this paper, focus is in particular on location-specific labour market attributes. The ambition is to introduce alternative measures reflecting the prospects of finding favourable combinations of residential location and job location. The first type of measure to be considered focuses on the spatial dimension: short commuting trips are preferred to longer distance commuting, ceteris paribus, since commuting involves both time costs and pecuniary, distance dependent, costs. Jobs are not distinguished in terms of e.g. positions, career opportunities or wages. Both jobs and workers are considered to be homogeneous in such respects. The commuting literature offers solid support for the use of a labour market accessibility measure to explain commuting flows between different

zones of a geography. One frequently used measure is the Hansen measure (Hansen 1959). Incorporating such an accessibility measure into a doubly constrained gravity modelling framework gives a so called competing destinations model (Fotheringham 1983, Gitlesen, Thorsen 2000). Osland, Thorsen (2008) introduced this measure as an attribute of a specific location in a housing market study. In Osland, Thorsen (2008) the specific formulation of the labour market accessibility measure was given by:

$$S_j = \sum_{k=1}^{98} E_k \exp(\sigma d_{jk}) \tag{2}$$

In this expression E_k represents employment in postal zone k, d_{jk} represents minutes driving time between zones j and k, σ is a parameter estimated by maximum likelihood estimation. In this way the distance deterrence parameter is estimated simultaneously with the other parameters in the models to be presented in Section 5. There are 98 postal delivery zones in the region.

For a theoretical interpretation of this measure in a commuting context, see Gitlesen, Thorsen (2000), where the rationale of a labour market accessibility measure is argued to follow from a two-stage household decision-making process. The first step involves the selection of a set of relevant location alternatives. In a search theoretical framework, distance appears as an information filter, increasing the probability of choosing combinations with short distance between job and residence. The accessibility measure is capturing relevant information on the spatial distribution of jobs. A location of high labour market accessibility is attractive, for instance because it increases the likelihood that household members can coordinate their journeys-to-work. In this context, the labour market accessibility measure is interpreted as a job opportunity density measure, and it also makes good sense to introduce such a measure in hedonic housing market studies. It is according to standard urban economic theory that houses for sale in highly accessible labour market locations can be expected to attract high bids, reflecting a high willingness to pay for residential locations involving low expected commuting costs. Another possibility would be to apply a network modelling approach to measuring accessibility. Xiao et al. (2016) demonstrate that this approach adds explanatory power in an urban setting. However, in our slightly more macroscopic framework we proceed with a more transparent and easily available information to measure accessibility.

As an alternative, or supplement, to the measure S_j, we suggest the following intuitive indicator of spatial labour market interaction in an explanation of housing prices. Let $\boldsymbol{X}$ be a commuting flow matrix where a typical element x_{ij} denotes the number of people living in zone i and working in zone j. The variable *OUT-COM* is then defined as the proportion of people living in zone i and working in another zone in the study area as follows:

$$OUT\text{-}COM_i = \frac{\sum_{j=1, j\neq i}^{N} x_{ij}}{\sum_{j=1}^{N} x_{ij}} 100 \tag{3}$$

The variable *IN-COM* is defined as the proportion of people working in zone j and living in one of the other zones:

$$IN\text{-}COM_j = \frac{\sum_{i=1, i\neq j}^{N} x_{ij}}{\sum_{i=1}^{N} x_{ij}} 100 \tag{4}$$

These measures represent a computationally simpler way to account for spatial labour market interaction than the non-linear accessibility measure. In addition, they do not require data on distances, or travelling times, between all the zones.

The two measures provide information on the spatial structure in the region. Labour market accessibility, represented by S_j, can be interpreted as a potential measure, representing the job opportunity density of a residential location. The measure reflects the degree to which a worker is able to take advantage of spatial variations in wage offers and the supply of career-enhancing jobs, within a reasonable commuting time. A reasonable hypothesis is that *OUT-COM* and *IN-COM* measure to what degree the workers actually

takes advantage of a favourable labour market accessibility, being based on observed rather than potential labour market behaviour. Individual heterogeneities in qualifications and preferences may result in a high level of spatial labour market interaction, with a correspondingly high level of excess commuting in densely populated urban areas. This further can be expected to correspond to high observed levels of out- and in-commuting in the centrally located zones of the region, representing a rationale for incorporating the measures *OUT-COM* and *IN-COM* in a model focusing on the relationship between the housing market and labour market interaction.

Despite a relatively wide scaling, Figure 1 indicates that there is a significant and positive correlation between *OUT-COM* and *IN-COM*. For the 98 postal delivery zones, the correlation coefficient between the two measures is 0.772. Both *OUT-COM* and *IN-COM* are further positively related to S_j, represented by correlation coefficients of 0.717 and 0.837, respectively. To the degree that the observation-based measure *IN-COM* is presupposed to represent labour market accessibility, it should a priori be expected to have a positive impact on housing prices. In particular for centrally located zones, the situation is similar for *OUT-COM*. The high level of out-commuting might result from a matching process, where heterogenous workers take advantage of attractive opportunities and job offers in a reasonable commuting distance outside the residential zone. In general, however, a high value of *OUT-COM* might also indicate that few jobs are available within the zone, contributing to a low level of labour market accessibility and low housing prices. Hence, it is not obvious what sign should be expected for *OUT-COM* in a hedonic regression model of housing prices.

In addition to this labour market accessibility perspective, it is important to account for the fact that there is a key difference between S_j and the other two measures. S_j is essentially capturing the existing spatial distribution of jobs, whereas *OUT-COM* and *IN-COM* in addition reflect the residential location choices of people. This means that the two observation-based measures are reflecting the (general) attractiveness of a place, including other perspective than the potential for labour market interaction. A high local value of *OUT-COM* might for instance reflect local amenities and/or attractive neighbourhood characteristics, making the zone appealing as a residential location. Hence, a positive estimate of the parameter attached to *OUT-COM* can be interpreted to capture positive neighbourhood externalities, in addition to the somewhat ambiguous effect of variations in labour market accessibility.

On the other hand, a substantial level of commuting into an area could produce congestion and other negative externalities that might have a significant effect on housing prices, see for instance Hughes, Sirmans (1992). The fact that job concentration and traffic in itself can be connected with negative externalities is also a major point in, for instance, Li, Brown (1980), Wilhelmsson (2000) and Osland, Pryce (2012). Hence, the parameter attached to *IN-COM* can be influenced by negative externalities, in addition to the positive effect stemming from labour market accessibility. This means that expected estimated sign of this parameter is also ambiguous, as it is a result of two counteracting effects. The possibility that *OUT-COM* and *IN-COM* capture different kinds of externalities is an argument in favour of including both measures in the model formulation, in addition to the labour market accessibility measure, S_j.

Following standard urban economic theory commuting time should reflect the actual commuting costs for households. It is also to be expected that commuting time rather than distance is a proper measure of commuting costs (Ma, Banister 2006). The third measure is, hence, based on actual mean commuting time in each zone:

$$MCT_i = \frac{\sum_{j=1}^{J} x_{ij} d_{ij}}{\sum_{j=1}^{J} x_{ij}} \tag{5}$$

This measure is calculated by first computing the total commuting time (TCT) from zone i as:

$$TCT_i = \sum_{j=1}^{J} x_{ij} d_{ij}$$

where i is the residential zone, j is the destination zone or job-zone, x_{ij} is the number of people residing in zone i and working in zone j, and d_{ij} is the travelling time between zone i and j. If $i = j$, individuals live and work in the same zone. In these cases, internal travelling time have been calculated as half of the travelling time to the nearest zone. In this way, we adjust for the fact that the areal size of the zones varies. The TCT for zone i is, finally, divided by the total size of the workforce living in zone i, given by $\sum_{j=1}^{J} x_{ij}$.

Assuming perfectly competitive housing market and following standard urban economic theory, the impact of commuting time on housing prices should be negative. However, for various reasons, workers do not minimize commuting distances (Hamilton 1982). Residential decision making is not merely about minimizing transportation costs and different structures of the urban or regional spatial structure could give different results regarding the extent of excess commuting (Ma, Banister 2006). A priori, it is therefore not obvious what sign should be expected for the impact of variations in MCT on housing prices. A high average commuting time might result for peripheral locations, where long distance commuting is the only relevant option for many workers. Such cases pull in the direction of a negative impact of MCT on housing prices. On the other hand, a high MCT can be observed in very centrally located zones, with a high level of spatial interaction, and significant labour market opportunities in many industries. As mentioned above, the high level of spatial interaction might reflect a situation with highly heterogeneous jobs and workers, where workers take advantage of attractive job opportunities in neighbouring zones. If such cases are dominating, then MCT should be expected to have a positive effect on housing prices.

In this paper we will test the hypotheses that:

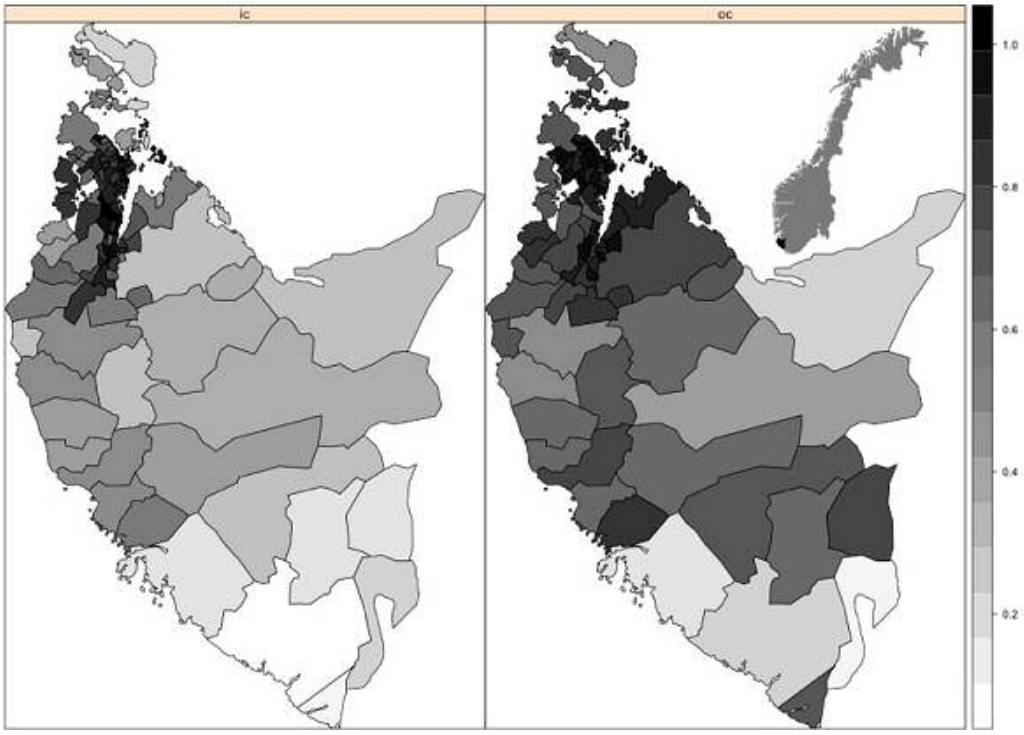
- $\mathbf{H_0^A}$: *OUT-COM* and *IN-COM* can replace the accessibility measure S_j in explaining spatial variation in housing prices
- $\mathbf{H_0^B}$: *OUT-COM* and *IN-COM* supplement the accessibility measure, and contribute with additional relevant information in explaining spatial variation in housing prices.
- $\mathbf{H_0^C}$: *MCT* can replace the accessibility measure, and contribute with additional relevant information in explaining spatial variation in housing prices.
- $\mathbf{H_0^D}$: *MCT* supplement the accessibility measure, and contribute with additional relevant information in explaining spatial variation in housing prices.

An illustration of the observed variation of the pattern of in- and out-commuting in the study area is given in Figure 1. The general tendency is a high degree of in-commuting in the central municipality of Stavanger and surrounding zones in the north. The further south and away from the CBD we move, the lower the degree of in-commuting. The percentage of out-commuting is also high in the most central cities, and is at a smaller but still high level in the postal zones surrounding these cities. One has to move to the most eastern and southern postal zones to find low levels of out-commuting.

It should be noted that the zones vary in size, and the largest postal code zones are located in rural, sparsely populated, areas. In such zones, it is to be expected that the percentage of commuting will be lower because significant distances must, on average, be travelled before the boundaries of these zones are crossed.

4 Overview of data

The results to be presented are based on housing price data from the second half of 2003 to 2007. The sample consists of 4392 observations from 13 municipalities, and 98 postal delivery zones. Only privately-owned single-family houses are included. One important reason for this restriction is that this is about the only house-type available on the market in rural areas. The housing data comes from two sources: Finn.no (Finn), a web-based service used by the main real-estate franchises in Norway, and GAB, the National Building Register. The data from Finn are used to compute the national housing price index. It



Note: The map of all the postal delivery zones in the study area. Darker areas signify high levels of commuting. In the top right-hand corner, the study area is indicated on a map of Norway.

Figure 1: The 2006-values of *IN-COM* (left) and *OUT-COM* (right)

includes the actual selling price, the year and month of sale, a measure for the size of the house, the type of house, the year the house was built, and an identification code for each property.

Statistics Norway started collecting this data in 2002 and the completeness of the information available for each observation improves over time. Complete identification codes are, for instance, missing from all observations in 2002 and the first half of 2003. By 2007 the identification codes are nearly complete. According to Statistics Norway, the Finn data covers about 40% of the house sales in Norway. From 2004 this percentage is probably higher, because data from one of the largest real-estate franchises (Notar AS) was added.

The GAB register is a combination of three registers: the official land property register, an address register, and a building register. A lot of information is missing on

Table 1: Descriptive statistics

Variable	n	Mean	Median	Std. dev.	Min	Max	Source
Price (in 1000 NOK)	4392	2630.7	2450	1188.1	280	16100	Finn and GAB
Age	4392	41.1	34	32.2	0	307	Finn (GAB)
LivingArea	4392	166.8	160	56.9	33	714	Finn
LotSize	4392	656.4	573	679	8.9	24700	GAB
Garage dummy	4392	0.355	-	-	-	-	GAB
DistCBD (minutes by car)	4392	19.05	12.74	18.40	0	104.02	-
IN-COM	4392	0.68	-	0.22	0.10	0.99	-
OUT-COM	4392	0.79	-	0.20	0.12	0.97	-
MCT (Mean Travelling Time)	4392	10.00	-	3.30	5.97	29.47	-

buildings constructed before 1983. For buildings and additions constructed after 1983 information is quite extensive and accurate. The GAB register plays a central part in the formal registration of a property transaction in Norway. This formal registration is not compulsory. Statistics Norway reports that about 94% of the house sales that are formally registered are registered within six months. GAB and Finn contain different information about housing characteristics. In order to obtain as much information as possible on housing attributes, we have combined the two data sources. Matching is based on property identification code and selling price. The implication is that the prices are available from both sources. In GAB, only the last selling price of a house appears in the register. In cases where a house has been sold more than once in the study period, we have to rely on price information from only the Finn database. Houses on leased lots (about 2-3% of the houses) have been excluded.

The hedonic or micro variables we were able to obtain are presented in Table 1, which also shows quite a big dispersion in lot size. Observations with lot size equal to zero have been excluded. We have also excluded observations with a useful floor space below 30 square meters. Observations with missing exogenous variables are not included in the regressions.

Statistics Norway states the following about the Finn sample: "The statistics (...) cover a majority of all used dwellings sold in Norway. Nonetheless it is possible that to some degree there is systematic sample skewness with regard to geography." To be more specific, the sample seems to be relatively smaller in most rural areas. This is in addition to the fact that the population of sold houses is smaller in the rural zones. In spite of this, the Finn data is used to compute the official national price index for used dwellings. Accordingly, we do not believe that this issue will have any impact on estimated results. To our knowledge this data is the best information available in Norway.

In addition to the micro data described above, we use some variables that are grouped according to postal zone. These zones vary greatly in areal size, and the urbanized zones in northwest are smaller than the most rural parts of the areas located in the south and east. In addition, there are topological differences. The terrain is far more mountainous in the east and south with a more limited road network. In these zones some of the habitation is concentrated in small hamlets in the valleys, but a substantial part is more dispersed stemming from small farms and holdings no longer used for agricultural production.

For each zone we have defined travel time to the CBD, in addition to travelling distances between all zones in the area and the number of jobs in each zone. The matrices of travelling times were calculated by the Norwegian Mapping Authority. The estimations were based on the specification of the road network into separate links, with known distances and speed limits existing in 2006. Information on speed limits and road categories is converted into travelling times through instructions from the Institute of Transport Economics. The centre of each (postal delivery) zone is found through detailed information on residential densities and the road network. Finally, the matrix of travelling times is constructed from a shortest-route algorithm.

The study area is markedly different from metropolitan areas in other countries. The region we are studying is one of the most affluent in Norway. The crime rate is relatively low, and the variation in the quality of public schools is small. The last point is due to an extensive egalitarian regional policy in Norway. However, some amenities such as provision of a range of services, closeness to open land and nature, etc. is expected to change when moving towards the CBD. Thus, the variable, distance to CBD, is important and is interpreted as urban attraction.

5 Alternative empirical model specifications

There are many examples in the literature where the hedonic methodology is used in empirical studies of housing markets. A review of some contributions can be found in Anselin, Lozano-Gracia (2009). There is no agreement, however, on what is the correct specification of a hedonic house price model, and the question of functional form remains an empirical problem that must be determined for each market under scrutiny. Hence we start by finding a parsimonious base model (M0). We then use the RESET test and

semi-parametric regression as aids to determine the correct functional form. The resulting base model is then tested for spatial effects. In the next section, the results from tests of the hypotheses from Section 3 and corresponding models is presented. We also illustrate how the predicted house price will vary with the relevant variables using the technique of a standard house. Finally, we explore the possibility of verification of our results by the spillover impacts from a spatial Durbin model.

Ignore first the possibility that labour market accessibility and commuting flows contribute to explaining housing prices. Based on previous empirical research from the study area (Osland et al. 2007, Osland, Thorsen 2008), we start with the following formulation of an empirical hedonic price model:

$$\ln(P) = a_0 + a_1 \ln(DistCBD) + a_2(\ln(DistCBD))^2 + \boldsymbol{bA} + c\,YearDummy_t + \epsilon_t \qquad (6)$$

where P is the observed real selling price of house i (1998 is the base year), $\boldsymbol{A}$ is a vector of the dwelling attributes listed in Table 1, *DistCBD* is the travelling time to the CBD, measured in minutes of car driving, and t represents the year of sale. All variables appear in logarithmic form except for the dummy variables. In the following discussion, equation (6) represents our base model, M0.

In order to estimate the housing price gradient, it is necessary to identify the centre of the geography. Following Plaut, Plaut (1998), much of the empirical literature in the field assumes that the location of the centre is known in advance. In our study the zone representing the CBD is found endogenously. We have experimented with different centrally located zones and used the descriptive measures of R^2 and SRMSE/APE (defined in Table 2) to find the zone appearing as the CBD of this geography. The result corresponds to a priori knowledge of the city of Stavanger. The inclusion of a quadratic term of the CBD account for the fact that the CBD-house price gradient are more elastic with increased distance to CBD (Osland et al. 2007). The variable *DistCBD* is interpreted as accounting for the effect of urban attraction, and reflects that households value urban amenities found in the city centre of the region. The inclusion of a gravity-based accessibility measure (equation (2)) can be interpreted as representing a more general labour market accessibility effect on housing prices (Osland, Thorsen 2008).

As mentioned above, the modelling procedure was motivated by previous estimation results from the same study area. The new data used in this paper are, however, from a more recent time period, with less information on housing attributes than was the case in Osland et al. (2007) and Osland, Thorsen (2008). To avoid model mis-specifications we therefore initially apply a semi-parametric approach, the RESET test (Ramsey 1969), and tests for spatial effects (Anselin 1988). The chosen modelling procedures are advocated and applied in Osland (2010).

The RESET test is a mis-specification test related to the functional form of the variables included in the model. In this case the test is based on powers of the fitted values and the fourth power is the highest. We test the null hypothesis that the model has no omitted variables. The alternative hypothesis is that the model is mis-specified.

The estimation was mainly performed in Stata, but we also use the program R combined with related packages (see Bivand et al. 2008).

Semi-parametric regression analysis is a flexible approach that is used as an exploratory tool to detect non-linearity in the data. There exist some hedonic studies that use similar approaches (see for instance Coulson 1992, Pace 1998, Bao, Wan 2004). In this paper a variant of the generalized additive models based on Hastie, Tibshirani (1990) is applied in combination with iterative penalized regression-smoothing splines. The method is explored in detail in Wood (2006). We estimate the model represented by equation (6) and include each continuous variable in turn into the smooth function $s(z)$, so that z is a variable vector not included in $\boldsymbol{A}$. The estimations have been made by using the mgcv (multiple generalized cross-validations) package (version 1.7-28) in R.

Consider for example the inclusion of lot size as a variable in the hedonic regression model, represented by the variable $\ln(LotSize)$ in the smooth function. The graphical result is illustrated in the right-hand side of Figure 2. This graph is based on a thin plate regression spline. The values on the caption of the y-axis denote efficient degrees of

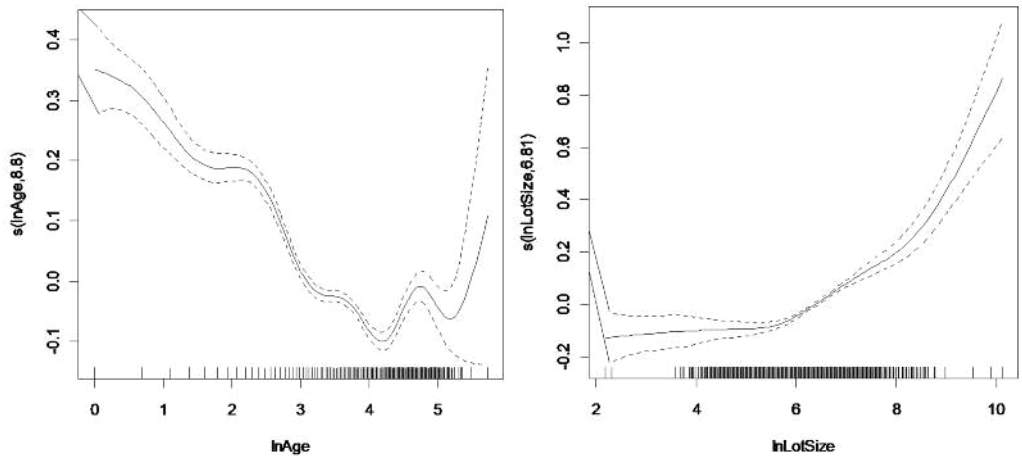


Figure 2: Age (left) and LotSize (right) are in turn included in the semi-parametric smooth function

freedom of the plotted term. The interpretation is that the equivalent of 6.81 degrees of freedom is used in estimating the smooth function (see Wood 2006, p. 170-172). The solid line in the figure represents the variation around the mean predicted value of the dependent variable. The dashed lines represent the approximately 95% confidence regions of the predicted values. The figure illustrates that the square of the variable $\ln(LotSize)$ should be included in the model. Finally, we have also included the square of $\ln(Age)$ in the hedonic regression model. A priori it is to be expected that these variables should be included as a non-linear relationship with housing prices.

The most important question in this paper is how to represent labour market accessibility and the characteristics of commuting flows in the model. The evaluation is based on traditional specification tests, such as Wald tests and log-likelihood ratio tests, in addition to the other measures described above.

Finally, a range of descriptive measures is included; see Table 1. Starting with a relatively parsimonious model formulation, more comprehensive model specifications are based on the results of the documented tests and descriptive measures.

Table 2 offers results of the following model specifications:

M1: M0 extended by the gravity-based accessibility indicator defined by equation (2).

M2: M0 extended by observed in and out-commuting, defined by equations (3) and (4).

M3: M2 extended by the gravity-based accessibility indicator defined by equation (2).

M4: M0 extended by mean commuting time defined by equation (5).

M5: M4 extended by the gravity-based accessibility indicator (2).

All the model specifications have been tested for spatial effects. We use the spdep (spatial dependence) package, from the R statistical programming environment. The robust Lagrange-multiplier (RLM) tests (see Florax, Nijkamp 2003) are reported in Table 2. The RLM tests asymptotically follows a chi-squared (1) distribution. The RLM-error statistics test the null hypothesis of no significant spatial error correlation, correcting for the presence of local spatial lag dependence in the dependent variable. Similarly, the RLM-lag statistics test the null hypothesis of no spatial autocorrelation in the dependent variable, correcting for the presence of local spatial error dependence. The used row-standardized spatial weight matrices, is based on a k-nearest neighbour structure. The k-nearest neighbour is chosen on the basis of distances in meters. Based on the log-likelihood values we use $k = 3$ for the spatial error model, so that each observation have the minimum of three neighbours.

For all model specifications, the null hypothesis of no spatial autocorrelation had to be rejected. When the RLM-error test statistic is the largest, a spatial error model with the above-mentioned weight structure would normally, remedy the problem (Anselin 1988). An important consequence of this is that the ordinary least squares estimator is consistent (Anselin 1988). Based on M0, the spatial error model is formulated as follows:

$$\begin{aligned} \ln(P) &= a_0 + a_1 \ln(DistCBD + a_2(\ln(DistCBD))^2 + \boldsymbol{bA} + c\,YearDummy_t + \tilde{\epsilon}_t \quad (7) \\ \tilde{\epsilon} &= \lambda \boldsymbol{W}\tilde{\epsilon} + u \end{aligned}$$

where $\boldsymbol{W}$ is the weight matrix, and λ is the spatial autoregressive parameter (Bivand et al. 2008, p. 284). The estimation of the spatial error model variants of M0–M5 does not change any of the results. The results from the spatial error models are presented in Appendix A.

6 Results on potential and observed measures of commuting pattern

Notice first from Table 2 that labour market accessibility (*ACCESS*) has a significantly positive impact on housing prices in all the models where it is taken into account (M1, M3 and M5). This is consistent with previous empirical analysis of the housing market in the region (Osland, Thorsen 2008). According to Table 2 the results are also consistent with the findings in Osland, Thorsen (2008) that spatial variation in housing prices is explained by a labour market accessibility effect and an urban attraction effect (*DistCBD*).

As mentioned in Section 3, it would be convenient, both from a computational and from a data collection point of view, if the variables *IN-COM* and *OUT-COM* could replace the accessibility measure S_j. By comparing the results from the model specifications M1 and M2 in Table 2, however, the hypothesis H_0^A has to be rejected. The negative estimate of the coefficient attached to *IN-COM* in M2 means that this variable cannot be interpreted as representing labour market accessibility. According to model specification M2, the variable *OUT-COM* has no significant impact on housing prices. In addition, the value of the log-likelihood function is clearly higher in M1 than in M2. Most of the other reported descriptive statistics favour M1.

As labour market accessibility cannot be replaced by the observed characteristics of commuting flows, M2 has an important spatially defined variable omitted from the model specification. Hence, the parameter estimates related to *IN-COM* and *OUT-COM* will be biased. M3 accounts for both labour market accessibility and the relevant characteristics of observed commuting flows.

The results based on M3 support the hypothesis H_0^B that the variables *IN-COM* and *OUT-COM* contribute additional relevant information in explaining spatial variation in housing prices. According to Table 2, both variables appear to be significant in the hedonic regression model M3. The value of the likelihood ratio test statistic is approximately 40 when M3 is compared with M1. This value clearly exceeds the critical value of the chi-square distribution ($\chi^2_{0.05}(2) = 5.991$). The p value of the Wald test is 0.000, given a null hypothesis of no joint significance of these two variables.

How should the results related to *IN-COM* and *OUT-COM* be interpreted? Consider, for instance, a zone located a short distance from the CBD, with a high value of the gravity-based labour market accessibility measure. It follows from the results in Table 3, that housing prices are predicted to be high in this zone; the labour market accessibility effect and the urban attraction effect operate in the same direction. As mentioned in Section 3, *IN-COM* at least to some degree captures the effect of variations in labour market accessibility. However, the result from M2 means that this is not the dominating effect of variations in *IN-COM*. The estimate of the coefficient attached to *IN-COM* is significantly negative. This means that negative externalities related to job concentrations, stemming for example from traffic, is dominating the effect explained by spatial variation in labour market accessibility. This conclusion is supported by the results following from M3, where labour market accessibility is explicitly accounted for by the variable S_j.

According to results for M3, a high value of *OUT-COM* is predicted to have a positive impact on house prices, adding, for instance, to the effects of the variables representing

labour market accessibility effect and urban attraction. The parameter estimates reported in Table 2 is small, however, and the positive impact of variations in *OUT-COM* is relatively marginal. *OUT-COM* is not found to have a significant impact on housing prices when S_j is not accounted for, in the model formulation M2. Recall from Section 3 that a possible negative impact of *OUT-COM* on housing prices can be argued to reflect a situation where few jobs are available within the zone, corresponding to a low level of labour market accessibility. Altogether, the results reported in Table 2 mean that this effect is dominated by the combined effect of positive neighbourhood externalities and a generally high spatial labour market interaction in centrally located areas.

Another hypothesis is that observed characteristics of in- and out-commuting may in particular be relevant for rural areas. To test this hypothesis, the variables representing commuting flows were interacted with a dummy variable, taking the value 1 if the zone is a rural zone, otherwise taking the value 0. This model extension did not alter significantly the parameter estimates related to the variables *IN-COM* and *OUT-COM*.

The labour market accessibility measure S_j offers information on the spatial distribution of jobs. It does not take into account the residential location pattern which would reflect the number of competing workers. Similarly, it does not account for the possibility that spatial labour market interaction is influenced by heterogeneities both in the working force and in the supply of jobs. Jobs for different categories of workers may for instance be clustered in specific zones of the geography, and this may influence commuting flows, residential location choices, and the willingness to pay for houses in different locations. Hence, information on the spatial distribution of different categories of jobs and workers may prove relevant in studying both commuting flows and house prices, but such data are not in general available at a sufficiently disaggregate subdivision into zones. In such a scenario, observation-based measures can, to some degree, capture the effect of labour market heterogeneities and characteristics of the residential location pattern. These issues are definitely not captured by the potential measure of labour market accessibility, S_j.

The Figures 3 and 4 illustrate how the predicted price of a so-called standard house varies along with variation in *IN-COM*, *OUT-COM*, and the gravity-based accessibility variable. A standard house is defined as a house that was sold in 2007, has a garage, and has not been sold in a rural area. Except for this, all the continuous variables are set to their average values for the sample (Osland, Thorsen 2008). The dependent variable has been transformed from its logarithmic form to prices in accordance with the following transformation rule:

$$P = \exp\left(\widehat{\ln(P)}\right) \exp\left(\frac{\widehat{\sigma^2}}{2}\right) \tag{8}$$

Here, σ^2 denotes an unbiased estimator of the residual variance (see Wooldridge 2003, p. 208).

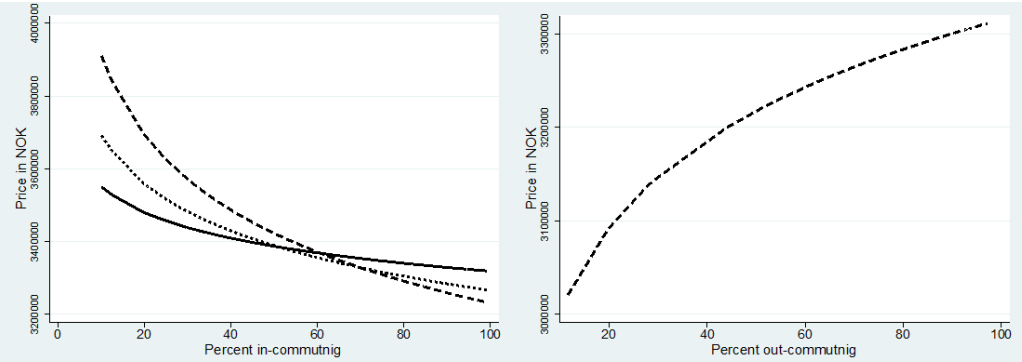
Notice from the left part of Figure 3 that incorporating the gravity-based accessibility measure, in a more adequate model formulation, contributes to increase the partial effect on housing prices of variations in the variable *IN-COM*. The model formulation M2 gives a biased, undervalued, estimate of the negative externalities associated with *IN-COM*. Figure 4 shows that the results related to the gravity-based accessibility measure are not sensitive to whether we include the commuting variables or not. As the value of the accessibility measure increases, so does the price of a standard house, albeit at a decreasing rate.

We have also experimented by introducing other variables related to observed commuting flows. The results on the impact of average commuting time are reported in Table 2. Figure 5 indicates that the relationship between the mean commuting time (MCT) and house prices is nonlinear. Experiments proved that the nonlinearities are satisfactorily represented by a quadratic term in a simple polynomial regression. A Wald test of the joint significance of the inclusion of the variable MCT and MCT^2 clearly has to be rejected in M5. The relevant p-value is 0.000.

Figure 6 provides an illustration of how different model formulations estimate the impact of variations in MCT on housing prices. At a first glance, the estimates resulting from the models M4 and M5, might seem relatively similar in Table 2. According to

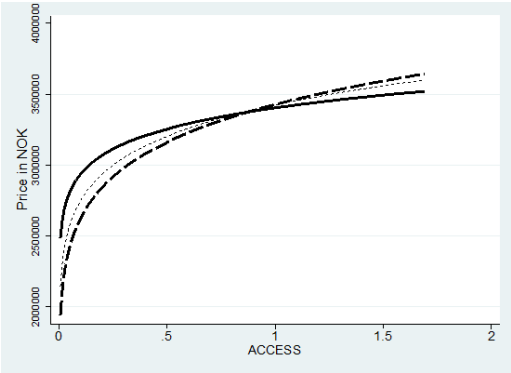
Table 2: Estimated results from alternative hedonic house price models

Variable name	M0	M1	M2	M3	M4	M5
Constant	12.970	12.417	12.942	12.085	12.543	11.430
	(71.89)	(63.52)	(71.75)	(59.50)	(49.53)	(38.87)
LotSize	-0.173	-0.204	-0.167	-0.203	-0.177	-0.221
	(-3.00)	(-3.62)	(-2.91)	(-3.55)	(-3.1)	(-3.92)
LotSize^2	0.022	0.025	0.022	0.025	0.023	0.026
	(4.62)	(5.27)	(4.55)	(5.22)	(4.72)	(5.59)
RurLotSize	-0.025	-0.021	-0.027	-0.023	-0.025	-0.019
	(-10.28)	(-8.83)	(-10.15)	(-9.64)	(-10.11)	(-8.36)
Age	-0.2017	-0.192	-0.205	-0.199	-0.204	-0.196
	(-11.75)	(-11.38)	(-11.84)	(-11.72)	(-11.75)	(-11.47)
Age^2	0.017	0.016	0.018	0.017	0.018	0.016
	(5.75)	(5.23)	(5.92)	(5.65)	(5.85)	(5.38)
Garage	0.042	0.043	0.040	0.042	0.042	0.040
	(7.21)	(7.44)	(6.93)	(7.16)	(7.19)	(7.06)
LivingArea	0.510	0.505	0.510	0.503	0.511	0.504
	(42.77)	(42.69)	(42.80)	(42.66)	(42.99)	(42.88)
YearDum04	0.097	0.098	0.097	0.097	0.097	0.097
	(8.89)	(9.03)	(8.90)	(9.05)	(8.89)	(8.98)
YearDum05	0.206	0.207	0.205	0.205	0.205	0.205
	(18.75)	(18.92)	(18.68)	(18.90)	(18.68)	(18.84)
YearDum06	0.370	0.371	0.370	0.370	0.370	0.370
	(33.55)	(33.67)	(33.49)	(33.82)	(33.51)	(33.82)
YearDum07	0.558	0.558	0.558	0.557	0.558	0.555
	(53.95)	(54.17)	(53.85)	(54.44	(53.93)	(54.39)
DistCBD	-0.056	-0.104	-0.043	-0.114	-0.076	-0.171
	(-3.14)	(-5.36)	(-2.28)	(-5.57)	(-3.90)	(-7.40)
DistCBD^2	-0.046	-0.027	-0.051	-0.024	-0.043	-0.0150
	(-13.18)	(-5.84)	(-12.22)	(-4.53)	(-10.86)	(-2.85)
ACCESS		0.064		0.096		0.114
		(6.48)		(7.79)		(7.06)
IN-COM			-0.030	-0.083		
			(-2.06)	(-4.90)		
OUT-COM			-0.043	0.001		
			(0.03)	(2.31)		
MCT					0.393	0.399
					(2.45)	(2.29)
MCT^2					-0.081	-0.054
					(-2.41)	(-1.45)
n	4392	4392	4392	4392	4392	4392
R^2	0.822	0.824	0.823	0.826	0.823	0.826
R^2(adj)	0.822	0.824	0.822	0.825	0.822	0.825
Log-likelihood	1184.02	1207.05	1187.48	1226.90	1187.60	1228.14
VIF	16.36	17.56	16.06	17.60	45.87	46.20
Ramsey reset	0.767	0.760	0.464	0.316	0.8003	0.5364
APE	510405	508469	510366	507978	510486	508223
SRMSE	0.289	0.288	0.290	0.288	0.289	0.288
RLM-lag	1.70	0.86	1.38	0.17	1.70	0.39
RLM-error	323.59	313.09	324.15	307.78	320.36	313.01



Note: The dashed lines refer to M3, while the solid line refers to a corresponding model specification without a gravity-based accessibility measure (M2). Predicted housing prices for variation in *OUT-COM* based on M2 is not shown in the Figure to the right, given that this variable does not contribute to explain the variation in housing prices in these two models.

Figure 3: Predicted house prices of a standard house



Note: The solid line refers to M1 where the commuting variables are excluded. The dotted line refers to M3, and the dashed line refers to M5. The values of the accessibility variables is mean-normalized.

Figure 4: Predicted house prices of a standard house

Figure 6, however, there is a substantial difference between the model M5, and the more parsimonious model M4. Technically, this is due to the different parameter values estimated for the quadratic term, but it can also be argued that the results reflect changes in the characteristics of the urban structure.

As pointed out in Section 3, a high *MCT* may reflect either a peripheral location, or a centrally located area with heterogenous agents and considerable excess commuting. For the model M5, the effects of labour market heterogeneity and a high level of spatial interaction in the central parts of the urban area seem to dominate, since *MCT* is estimated to have a positive impact on housing prices. Over time, the job growth in the Stavanger urban area has come in areas that used to be the outskirt of the city, mainly in suburban industrial parks. As a consequence, workers living in residential areas close to the city centre, see Figure 1, no longer have on average shorter commuting times than workers living in some of the suburban areas. Still, some of the residential areas close to the city centre have very high housing prices. These areas have traditionally been fashionable residential locations, with neighbourhoods that are popular, beyond the pure urban attraction effect. This is one possible explanation why housing prices are predicted to be an increasing function of *MCT*.

The first, rising part of the curve resulting from model M4 in Figure 6 can be explained from the same line of reasoning as above. In this model, however, the labour market accessibility measure is not incorporated. Since *MCT* is closely and negatively correlated to labour market accessibility (Pearson's correlation coefficient is -0.8677), the labour

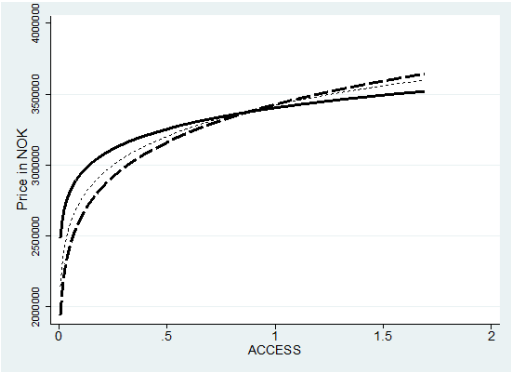
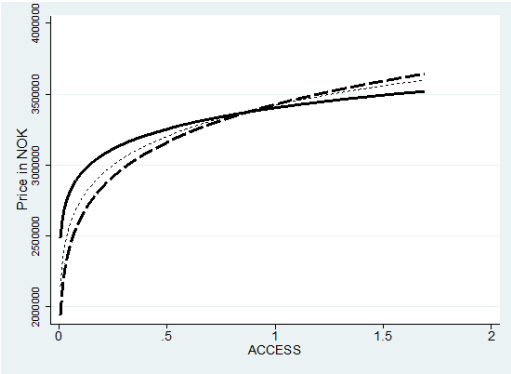


Figure 5: The variable ln MCT is included in the semi-parametric smooth function



Note: The solid line refers to M4, which includes the variable mean commuting time. The dotted line refers to M5, which also includes the gravity based accessibility measure in addition to mean commuting time.

Figure 6: Predicted house prices of a standard house

market accessibility effect dominates for variations in MCT higher than 10 minutes. The effect of MCT is, hence, negatively biased in this model formulation.

7 Issues of endogeneity and robustness checks

Empirical studies using the hedonic house prices model necessitates considerations of endogeneity, which may bias the estimated implicit prices. First, there should not be an omitted variable bias. This problem is discussed in Section 7.1. Another potential source of endogeneity bias is reverse causality, which is discussed in Section 7.2.

7.1 Omitted variable bias

An important example of left out spatially related missing characteristic is negative and positive externalities such as noise, local air quality or the physical and social surroundings. We do not control for these types of variables in our model specifications. The reason for not including the variables is lack of data. If these variables have a significant impact on housing prices, and if they correlate with the studied variables, they may, create a bias of the studied coefficients. However, the bias may go in many directions, given the potential of a large number of missing spatial characteristics.

We have performed two robustness checks in order to study if an important omitted variable bias is present. Given that we do not have information on specific environmental variables, we use the average value of houses sold in an area as control for a range of left-out characteristics. We define a neighbourhood either at the postal code level or at the municipality level. There are 98 postal codes and 11 municipalities in the study area. The inclusion of the control variable does not change any of the conclusions regarding

sign of coefficients for the studied variables in any of the models. However, the impact on the coefficients vary according to which definition of neighbourhood we use. Using the most disaggregate neighbourhood level reduces the absolute value of the relevant coefficients, whereas the inclusion of the most aggregate control variable increases the absolute value of the coefficients. The difference in estimated coefficients is largest when using a control variable at the disaggregated level.

It is difficult to get information about the direction of a potential bias based on this robustness check. Given the changes in the coefficients, the bias is relatively small, still most of the results show that the estimated coefficients of the accessibility variables documented in Table 2, are outside a 95 % confidence region of the models which include a control variable.

The second approach is to use an alternative estimator, the spatial Durbin model. This estimator is robust to omitted variables reflecting spatial characteristics (LeSage, Pace 2009). Description of estimation procedures and general interpretations of the spatial Durbin model will follow the presentations found in these publications. According to LeSage, Pace (2009), this model-estimator is robust to omitted variables reflecting spatial characteristics. The results presented in Table 2 are based on relatively parsimonious model specifications, and most empirical hedonic house price models are encumbered by omitted variables. Hence, the spatial Durbin model could reveal additional information in this respect. The spatial Durbin variant of the model is specified as follows:

$$P = \rho \boldsymbol{W} \beta_0 + \rho \boldsymbol{W} \boldsymbol{X} \beta_1 + \epsilon \tag{9}$$

In equation (9), P is a vector of observed prices, $\boldsymbol{X}$ is a matrix of observations on independent variables, and $\boldsymbol{W}$ is the $n \times n$ matrix of exogenous spatial weights. This model allows a spatial lagging of the dependent variable, in addition to a spatial lagging of the independent variables (see Bivand 1984, LeSage, Fischer 2008). The spatial Durbin model has been estimated by using the same weight matrix as described for the spatial error model in Section 5, except that we use $k = 4$ in the weights, because this gives the highest log-likelihood values in this case.

According to LeSage, Fischer (2008) the estimated parameters related to the spatial Durbin model have no straightforward interpretation. For this reason, we only report the spillover impacts, estimated by the procedure in LeSage, Pace (2009, p. 38). The covariance matrix of the coefficients has been calculated by using numerical methods (LeSage, Pace 2009, p. 56-59). This matrix and traces of powers series of the weights matrix, estimated by Monte Carlo approximations, were used to derive impact measures, and tests of significance (LeSage, Pace 2009, p. 96-104 and 114-115).

In the spatial Durbin model, represented by equation (9), the price of a house i is a function of the neighbouring house prices through the lagged dependent variable. Neighbouring house prices are a function of the values of the houses' own attributes. Changing these attributes has an effect on its own price, and hence also on the price of house i. In addition, the price of house i depends on the attribute values of its neighbours, as expressed through the spatially lagged independent variables. The dimension of the spillover effects depends upon the size of the estimated spatial autocorrelation parameters and the specification of the neighbourhood matrix (see LeSage, Fischer 2008, LeSage, Pace 2009, Kirby, LeSage 2009). Even when the lagged independent variables are statistically not significant, there may still exist some significant spillover effects occurring through a spatially autocorrelated dependent variable.

The estimated average impacts from the spatial Durbin model are presented in Table 3. The results are based on M3 and M5. Excluding the polynomial variants of all the commuting variables that we are studying makes the interpretation of the results easier. In our case the direct impacts are calculated as the average effect on a house price i of a change in each of the explanatory variables related to that house. The average total impact is the estimated effect on the price, followed by a change in each of the variables, respectively, over all observations. The indirect impact is represented by the difference between the total and direct impacts. In this way the indirect impact captures the average effects on the price of house i from the change in the variables of other houses.

Table 3: Estimated direct, indirect, and total impact from variants of M3 and M5 using the spatial Durbin Estimator

	M3			M5		
Variable Name	Direct	Indirect	Total	Direct	Indirect	Total
Lotsize	-0.201	0.150	-0.051	-0.173	0.012	-0.161
	(-6.47)	(2.21)	(-0.70)	(-4.92)	(0.15)	(-1.84)
LotSize^2	0.025	-0.013	0.013	0.024	-0.005	0.020
	(9.69)	(-2.20)	(2.01)	(8.73)	(-0.67)	(2.64)
RurLotsize	-0.028	0.007	-0.021	-0.041	0.024	-0.017
	(-8.89)	(1.76)	(-8.10)	(-3.92)	(2.21)	(-5.24)
Age	-0.196	0.038	-0.158	-0.161	-0.097	-0.285
	(-11.93)	(1.05)	(-4.13)	(-9.39)	(-2.51)	(-6.22)
Age^2	0.016	-0.004	0.012	0.009	0.020	0.028
	(5.89)	(-0.71)	(1.84)	(2.95)	(3.09)	(4.13)
Garage	0.040	0.026	0.066	0.039	0.009	0.048
	(6.48)	(2.06)	(4.70)	(6.78)	(0.63)	(2.86)
LivingArea	0.503	-0.037	0.466	0.487	0.094	0.581
	(52.91)	(-1.84)	(21.11)	(53.64)	(3.72)	(20.94)
YearDum04	0.099	-0.029	0.070	0.097	-0.001	0.096
	(9.58)	(-1.24)	(2.86)	(9.05)	(-0.04)	(2.22)
YearDum05	0.209	-0.027	0.182	0.207	0.038	0.245
	(19.93)	(-1.16)	(7.48)	(19.40)	(1.03)	(5.72)
YearDum06	0.373	-0.014	0.359	0.371	-0.019	0.353
	(35.41)	(-0.57)	(14.40)	(35.31)	(-0.53)	(8.50)
YearDum07	0.560	-0.022	0.538	0.557	0.001	0.558
	(56.92)	(-0.92)	(21.51)	(54.24)	(0.04)	(14.02)
DistCBD	-0.081	-0.073	-0.153	-0.036	-0.103	-0.139
	(-3.45)	(-2.09)	(-4.86)	(0.40)	(-0.97)	(-4.23)
DistCBD^2	-0.024	0.008	-0.016	0.014	-0.033	-0.019
	(-4.20)	(0.95)	(-1.94)	(0.37)	(-0.87)	(-2.66)
ACCESS	0.084	0.035	0.120	0.161	-0.051	0.109
	(7.23)	(1.87)	(6.99)	(2.00)	(-0.63)	(5.22)
IN-COM	-0.073	-0.064	-0.136			
	(-4.67)	(-2.62)	(-6.18)			
OUT-COM	0.040	0.033	0.073			
	(2.36)	(1.87)	(2.72)			
MCT				0.247	-0.122	0.125
				(1.80)	(-0.88)	(3.43)

Note: Z-values in parentheses.

The spatial Durbin model yields the same sign on all the estimated parameters, as for the ordinary least squares estimator. The direct impacts yield results that are within the 95% confidence region of the ordinary least squares regression results for the model variant based on M3. For the specification based on M5, this is not always the case. For most of these variables the total impact is within the 95% confidence region. The impacts related to age variables is not within the mentioned 95% confidence region. For M1, not reported, the total impact of *ACCESS* is not within the 95% confidence region. The indirect impact is negative, but not significant.

Indirect impacts are significant at the 5% level for the variables *LivingArea* and *Age*^{2} for the model based on M5. The same is the case for the spatial Durbin variant of M1. In the model based on M3, only *IN-COM* has a significant indirect impact at the 5% significance level. The indirect impact is negative. This means that an increase in *IN-COM* of the houses neighbouring a house i, will on average have a negative impact on the price of house i. Hence, there are negative spillover effects of having a house located in an area with a high relative level of in-commuting. This result is in line with our interpretations of regression results found in previous sections. The average indirect

impacts of *OUT-COM* and *ACCESS* are positive, but not significant at the 5% level. The results of this robustness check also show that all the direct impacts related to the studied variables have the same sign as the ones documented in Table 2. Given the fact that the correlation between the variables are relatively high, it is to be expected that the inclusion and exclusion of variables will have an impact on the value of the estimated coefficients. The direction of any existing bias is not possible to determine based on any of the robustness checks.

7.2 Reverse Causality

The second relevant type of endogeneity problem could be reverse causality. By way of example, neighbourhoods with higher housing prices could attract high-income earners, perhaps also with higher educational attainments. Locations close to these places could be relevant for specific firms in order to get access to an attractive pool of labour. This type of reasoning is based on the hypothesis that "jobs follow people", which subsequently may improve labour market accessibility in places with higher housing prices. Hoogstra et al. (2017) provide a useful meta-analysis of the related literature regarding this question. Their main finding is that the evidence is inconclusive. The causality between jobs and people could run in different directions (pages 371-372). If this holds true, for our data, the resulting bias could be minor, and unpredictable.

If there exists reverse causality, identification of the impact of accessibility would necessitate an instrument. One possible instrument could be an exogenous change in measures of accessibility. However, we do not have access to any such instrument. Moreover, we focus on several variables, which are potentially endogenous. Testing for exogeneity is, hence, not straightforward according to e.g. Baum et al. (2007). We have to use several instruments, and in these cases, the traditional IV-estimators could be biased and inconsistent (see also Nordvik et al. 2019, for further discussions of this issue).

Finally, it is also possible to argue that the need for an instrument is less important in our case, given the results from the spatial Durbin model. This model accounts for indirect spill over impacts and the spatial Durbin model, in general, is robust to omitted systematic spatial variation of characteristics (LeSage, Pace 2009).

8 Conclusions

In explaining spatial variation in housing prices, gravity-based accessibility measures have been suggested as a generalization of modern polycentric labour market structures. From a computational and data collection perspective, it would be convenient if easily available information on actual commuting patterns could replace a more complex measure of commuting potential. According to our results, two of the hypotheses formulated in Section 3 have to be rejected, however. The labour market accessibility effect is not adequately represented by the proposed characteristics of observed commuting patterns. We used a wide range of different methods to obtain robust conclusions. Labour market accessibility in relation to housing prices is best captured by the gravity based potential variable.

Our results, on the other hand, provide support for the hypotheses H_0^B and H_0^D. Observed measures of commuting patterns are found to contribute with information that adds to the effect of a potential measure in explaining spatial variation in housing prices. In particular, the results from the ordinary least squares and the spatial Durbin estimator support the hypothesis that a relatively high level of commuting into a zone corresponds to negative externalities, such as noise, pollution, or other negative effects of heavy traffic and/or industrial activities. There is only weak support for the hypothesis that a relatively high level of out-commuting from a zone corresponds to attractive neighbourhood characteristics, which are positively related to housing prices. Average commuting time, a priori, reflect something about the actual commuting costs for households. However, our interpretation of the positive impact of this variable is that a high level of spatial interaction in the central parts of the urban area seem to dominate. Over time, the job growth in the Stavanger urban area has come in areas that used to be the outskirt of the city. Still, some of the residential areas close to the city centre have very high

housing prices. These areas have traditionally been fashionable residential locations, with neighbourhoods that are popular, beyond the pure urban attraction effect.

The results presented in this paper contribute to modifying predictions of how changes in labour market accessibility affect housing prices. Assume, as an example, that a number of jobs are relocated from a zone. This means that the labour market accessibility is reduced for this zone, and house prices decreases. On the other hand, the reduced number of jobs might lead to a reduced commuting flow into the zone, contributing to increased house prices. To some extent, this offsets the effect of a reduction in labour market accessibility. We will not be more specific on the overall effect in this paper. This depends, for instance, on where the jobs are relocated, and how the labour market accessibility is affected.

The introduction of alternatives to the potential measure of accessibility was not only motivated by considerations of simplicity and data requirements. We have been arguing that commuting flows, the residential location pattern, and house prices may result from a complex mixture of labour market heterogeneities, characteristics of the residential location pattern, and different kinds of externalities. Ideally, such heterogeneities, as well as the conditions causing negative and positive externalities should be explicitly controlled for in the model formulation. In most cases, however, relevant information is not available without a massive data collection effort. It is a useful result that observation-based measures of commuting flows to some degree capture the effect of heterogeneities and externalities. This information on commuting flows is often easily available. Referring to the main ambition and motivation of our analysis, we find that this information is adding to the explanatory power of the hedonic model of housing prices. Hence, we conclude that observation-based measures of commuting flows can supplement, but not substitute, the gravity-based potential measure of accessibility in explaining spatial variation of housing prices.

References

Adair A, McGreal S, Smyth A, Cooper J, Ryley T (2000) House prices and accessibility: The testing of relationships within the Belfast urban area. *Housing studies* 15: 699–716. CrossRef.

Ahlfeldt G (2011) If Alonso was right: modelling accessibility and explaining the residential land gradient. *Journal of Regional Science* 51: 318–338. CrossRef.

Anas A, Arnott R, Small K (1998) Urban spatial structure. *Journal of Economic Literature* 36: 1426–1464

Anselin L (1988) *Spatial econometrics: Methods and models*. Kluwer, London. CrossRef.

Anselin L (2002) Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics* 27: 247–267. CrossRef.

Anselin L, Lozano-Gracia N (2009) Spatial hedonic models. In: Mills T, Patterson K (eds), *Palgrave Handbook of Econometrics*, Volume 2. Palgrave Macmillan. CrossRef.

Ball M, Kirwan R (1977) Accessibility and supply constraints in the urban housing market. *Urban Studies* 14: 11–32. CrossRef.

Bao H, Wan A (2004) On the use of spline smoothing in estimating hedonic housing price models: Empirical evidence using Hong Kong data. *Real Estate Economics* 32: 487–507. CrossRef.

Baum CF, Schaffer ME, Stillman S (2007) Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal* 7: 465–506. CrossRef.

Bivand R (1984) Regression modelling with spatial dependence: An application of some class selection and estimation methods. *Geographical Analysis* 16: 25–37. CrossRef.

Bivand R, Pebesma E, Gómez-Rubio V (2008) *Applied spatial data analysis with R.* Springer. CrossRef.

Coulson N (1992) Semiparametric estimates of the marginal price of floorspace. *Journal of Real Estate Finance and Economics* 5: 73–83. CrossRef.

Dubin R (1992) Spatial autocorrelation and neighbourhood quality. *Regional Science and Urban Economics* 22: 433–452. CrossRef.

Farber S, Neutens T, Miller HJ, Li X (2013) The social interaction potential of metropolitan regions: A time-geographic measurement approach using joint accessibility. *Annals of the Association of American Geographers* 103: 483–504. CrossRef.

Florax R, Nijkamp P (2003) Misspecification in linear spatial regression models. *Technical report, Tinbergen Institute Discussion Paper, (TI 2003-081/3)*

Fotheringham AS (1983) A new set of spatial interaction models: The theory of competing destinations. *Environment and Planning A* 15: 1121–1132. CrossRef.

Gitlesen JP, Thorsen I (2000) A competing destinations approach to modeling commuting flows: A theoretical interpretation and an empirical application of the model. *Environment and Planning A* 32: 2057–2074. CrossRef.

Gjestland A, McArthur D, Osland L, Thorsen I (2014) The suitability of hedonic models for cost-benefit analysis: Evidence from commuting flows. *Transportation Research Part A: Policy and Practice* 61: 136–151. CrossRef.

Hamilton B (1982) Wasteful commuting. *Journal of Political Economy* 90: 1497–1504. CrossRef.

Handy S, Niemeier D (1997) Measuring accessibility: an exploration of issues and alternatives. *Environment and Planning A* 29: 1175–1194. CrossRef.

Hansen W (1959) How accessibility shapes land use. *Journal of the American Institute of Planners* 25: 73–76. CrossRef.

Hastie T, Tibshirani R (1990) *Generalized additive models.* Chapman and Hall, London. CrossRef.

Heikkila E, Gordon P, Kim J, Peiser R, Richardson H (1989) What happened to the CBD-distance gradient? Land values in a polycentric city. *Environment and Planning A* 21: 221–232. CrossRef.

Hoogstra GJ, van Dijk J, Florax RJ (2017) Do jobs follow people or people follow jobs? A meta-analysis of Carlino-Mills studies. *Spatial Economic Analysis* 12: 357–378. CrossRef.

Hughes J, Sirmans C (1992) Traffic externalities and single-family house prices. *Journal of Regional Science* 32: 487–500. CrossRef.

Jackson J (1979) Intraurban variation in the price of housing. *Journal of Urban Economics* 6: 464–479. CrossRef.

Kirby D, LeSage J (2009) Changes in commuting to work times over the 1990 to 2000 period. *Regional Science and Urban Economics* 39: 460–471. CrossRef.

Kwan MP, Murray AT, O'Kelly ME, Tiefelsdorf M (2003) Recent advances in accessibility research: Representation, methodology, and applications. *Journal of Geographical Systems* 5: 129–138. CrossRef.

LeSage J, Fischer M (2008) Spatial growth regressions: Model specification, estimation and interpretation. *Spatial Economic Analysis* 3: 275–304. CrossRef.

LeSage J, Pace R (2009) Introduction to spatial econometrics. Chapman and hall/crc, boca raton. CrossRef.

Li M, Brown H (1980) Micro-neighbourhood externalities and hedonic prices. *Land Economics* 56: 125–140. CrossRef.

Ma K, Banister D (2006) Excess commuting: A critical review. *Transport Reviews* 26: 749–767. CrossRef.

Nordvik V, Osland L, Thorsen I, Thorsen IS (2019) Capitalization of neighbourhood diversity and segregation. *Environment and Planning A: Economy and Space* 51: 1775–1799. CrossRef.

Osland L (2010) An application of spatial econometrics in relation to hedonic house price modelling. *Journal of Real Estate Research* 32: 289–320

Osland L, Pryce G (2012) Housing prices and multiple employment nodes: Is the relationship nonmonotonic? *Housing Studies* 27: 1182–1208. CrossRef.

Osland L, Thorsen I (2008) Effects on housing prices of urban attraction and labor market accessibility. *Environment and Planning A* 40: 2490–2509. CrossRef.

Osland L, Thorsen I, Gitlesen J (2007) Housing price gradients in a geography with one dominating center. *Journal of Real Estate Research* 29: 321–346

Pace R (1998) Appraisal using generalized additive models. *Journal of Real Estate Research* 15: 77–99

Plaut P, Plaut S (1998) Endogenous identification of multiple housing price centers in metropolitan areas. *Journal of Housing Economics* 7: 193–217. CrossRef.

Ramsey J (1969) Tests for specification errors in classical linear least squares regression analysis. *Journal of Royal Statistical Society B* 31: 350–371. CrossRef.

Vågane L, Brechan I, Hjorthol R (2011) Den nasjonale reisevaneundersøkelsen – 2009. The institute for transport economics, TØI report 1130

Waddell P, Berry B, Hoch I (1993) Residential property values in a multinodal urban area: New evidence on the implicit price of location. *Journal of real estate finance and economics* 7: 117–141. CrossRef.

Wilhelmsson M (2000) The impact of traffic noise on the values of single-family houses. *Journal of Environmental Planning and Management* 43: 799–815. CrossRef.

Wood S (2006) *Generalized additive models. An introduction with R* (2nd ed.). Chapman and Hall/CRC. CrossRef.

Wooldridge J (2003) *Introductory econometrics. A modern approach.* South-Western, Mason, OH

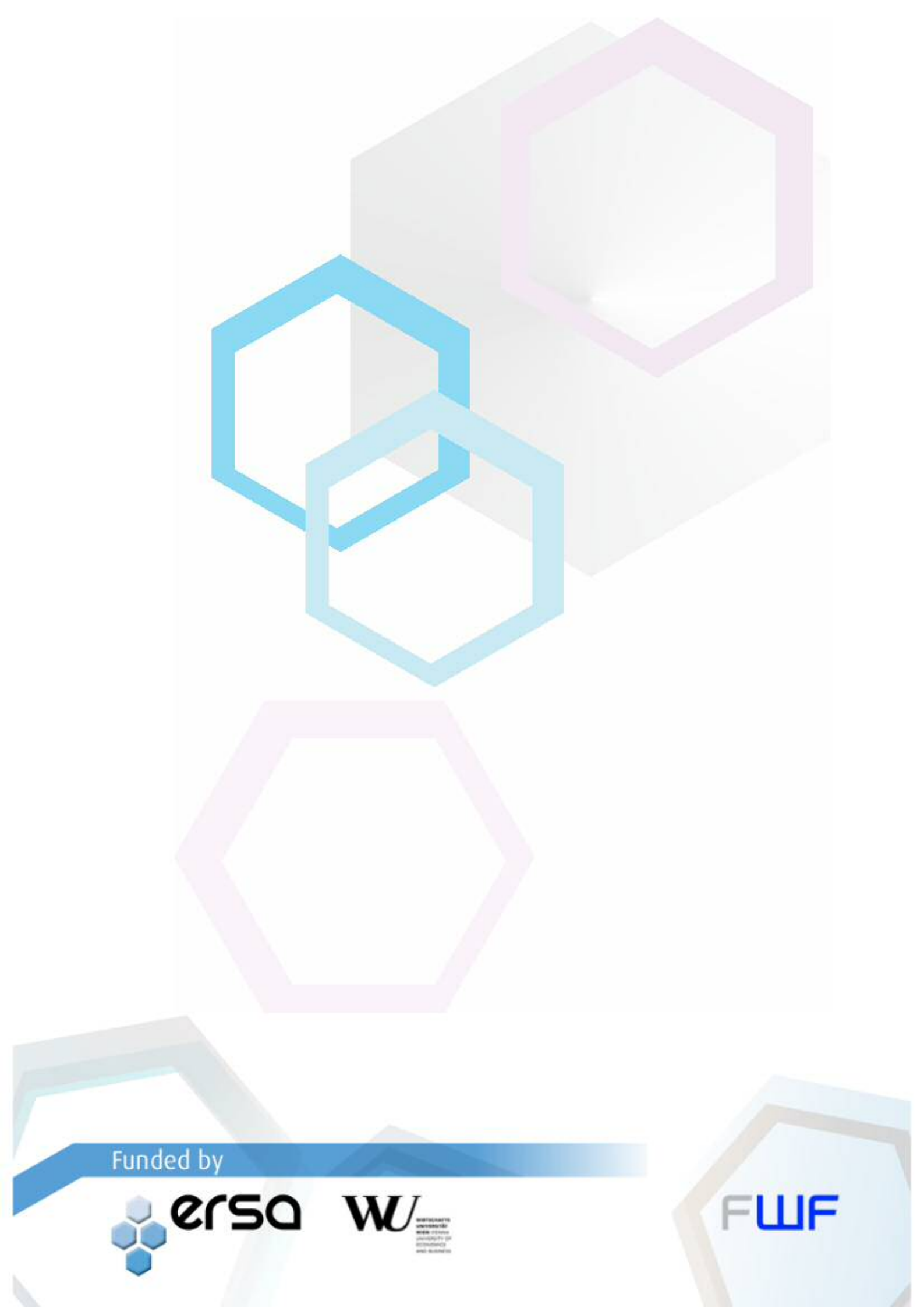
Xiao Y, Orford S, Webster CJ (2016) Urban configuration, accessibility, and property prices: a case study of Cardiff, Wales. *Environment and Planning B: Planning and Design* 43: 108–129. CrossRef.

A Appendix: Results of spatial error model estimations

Table A.1: Estimated results for the hedonic house price models based on the spatial error model formulation as specified in equation (9)

Variable Name	M0	M1	M2	M3	M4	M5
Constant	13.060	12.506	13.031	12.180	12.710	11.596
	(120.88)	(75.10)	(119.33)	(66.87)	(48.96)	(37.09)
Lotsize	-0.187	-0.204	-0.183	-0.202	-0.189	-0.214
	(-6.11)	(-6.61)	(-5.94))	(-6.56)	(-6.15)	(-6.93)
LotSize^2	0.024	0.026	0.024	0.026	0.024	0.026
	(9.55)	(9.98)	(9.39)	(9.99)	(9.58)	(10.30)
RurLotsize	-0.025	-0.021	-0.026	-0.023	-0.025	-0.019
	(-8.76)	(-7.26)	(-8.86)	(-7.86)	(-8.73)	(6.47)
Age	-0.1809	-0.178	-0.183	-0.182	-0.182	-0.180
	(-10.71)	(-10.51)	(-10.78)	(-10.82)	(-10.78)	(-10.67)
Age^2	0.013	0.012	0.013	0.013	0.013	0.013
	(4.64)	(4.43)	(4.75)	(4.77)	(4.75)	(4.60)
Garage	0.039	0.040	0.038	0.039	0.039	0.039
	(6.71)	(6.82)	(6.59)	(6.70)	(6.68)	(6.67)
Size of house	0.493	0.491	0.493	0.490	0.493	0.491
	(52.31)	(52.19)	(52.31)	(52.27)	(52.47)	(52.36)
YearDum04	0.0955	0.096	0.096	0.096	0.096	0.096
	(9.87)	(9.91)	(9.87)	(9.93)	(9.88)	(9.90)
YearDum05	0.199	0.199	0.199	0.199	0.199	0.199
	(20.41)	(20.45)	(20.39)	(20.46)	(20.41)	(20.43)
YearDum06	0.370	0.370	0.370	0.370	0.370	0.370
	(38.33)	(38.34)	(38.34)	(38.36)	(38.32)	(38.35)
YearDum07	0.5573	0.557	0.557	0.557	0.557	0.556
	(58.91)	(58.88)	(58.92)	(58.93)	(58.93)	(58.82)
DistCBD	-0.0574	-0.101	-0.040	-0.105	-0.074	-0.163
	(-2.58)	(-4.14)	(-1.58)	(-3.93)	(-2.96)	(-5.71)
DistCBD^2	-0.047	-0.029	-0.053	-0.028	-0.045	-0.018
	(-11.06)	(-5.07)	(-9.17)	(-3.96)	(-9.48)	(-2.82)
ACCESS		0.059		0.088		0.109
		(4.36)		(5.81)		(6.07)
IN-COM			-0.033	-0.082		
			(-1.78)	(-4.39)		
OUT-COM			-0.006	0.034		
			(-0.22)	(1.65)		
MCT					0.315	0.306
					(1.48)	(1.47)
MCT^2					-0.063	-0.035
					(-1.44)	(-0.80)
λ	0.374	0.364	0.373	0.359	0.374	0.358
	(23.67)	(22.85)	(23.60)	(22.30)	(23.60)	(22.24)
Log-likelihood	1420.26	1429.58	1422.56	1440.23	1422.69	1440.57

Note: Z-values in parentheses.



Funded by