I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10694

Incorporating Neighbourhood Choice in a Model of Neighbourhood Effects on Income

Maarten van Ham
Sanne Boschman
Matt Vogel

APRIL 2017

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10694

Incorporating Neighbourhood Choice in a Model of Neighbourhood Effects on Income

Maarten van Ham
Delft University of Technology and IZA

Sanne Boschman
Utrecht University

Matt Vogel
University of Missouri

APRIL 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Incorporating Neighbourhood Choice in a Model of Neighbourhood Effects on Income

Studies of neighbourhood effects often attempt to identify causal effects of neighbourhood characteristics on individual outcomes, such as income, education, employment, and health. However, selection looms large in this line of research and it has been repeatedly argued that estimates of neighbourhood effects are biased as people non-randomly select into neighbourhoods based on their preferences, income, and the availability of alternative housing. We propose a two-step framework to help disentangle selection processes in the relationship between neighbourhood deprivation and earnings. We first model neighbourhood selection using a discrete choice framework and derive correction components to adjust parameter estimates in a subsequent neighbourhood effects model for the unequal probability that an individual 'chooses' to live in a particular area. Applying this technique to administrative data from the Netherlands, we find significant interactions between personal and neighbourhood characteristics in the selection model. This confirms individual differences in neighbourhood preferences; individuals non-randomly select into neighbourhoods. The baseline neighbourhood effects model reveals a significant effect of average neighbourhood income on individual income. When we include correction components for the differential sorting of individuals into specific neighbourhoods, the effect of neighbourhood income diminishes, but remains significant. These results suggest that researchers need to be attuned to the role of selection bias when assessing the role of neighbourhood effects on individual outcomes. Perhaps more importantly, the strong, persistent effect of neighbourhood deprivation on subsequent earnings suggests that neighbourhood effects reflect more than the shared characteristics of neighbourhood residents; place of residence partially determines economic well-being.

JEL Classification: I30, J60, R23

Keywords: neighbourhood effects, neighbourhood sorting, neighbourhood choice, income, selection bias, discrete choice, longitudinal data

Corresponding author:
Maarten van Ham
OTB – Research for the Built Environment
Faculty of Architecture and the Built Environment
Delft University of Technology
P.O. Box 5043
2600 GA Delft
The Netherlands

E-mail: m.vanham@tudelft.nl

1. Introduction

The neighbourhood effects literature concerns itself with identifying causal effects of living in (deprived) neighbourhoods on a range of individual level outcomes such as income, education, employment, and health. The literature on neighbourhood effects is far from conclusive and there is a major debate on the size and significance of neighbourhood effects, and whether the effects found are causal or not. Several studies suggest that selection and not causality is behind most of the current neighbourhood effects 'evidence' (e.g. Oreopoulos 2003; Bolster et al. 2007; van Ham and Manley 2010; van Ham et al. 2012a). According to this perspective, many existing studies fail to convincingly show real causal neighbourhood effects, because they ignore, or fail to adequately address, neighbourhood selection (Durlauf 2004; van Ham and Manley 2010). This leaves the impression that neighbourhood effects are important, while in reality these studies might just show correlations between individual and neighbourhood characteristics (Cheshire 2007). From this vantage point, it is likely that studies claiming to have found that poor neighbourhoods make people poor(er) only show that poor people live in poor neighbourhoods because they cannot afford to live elsewhere (Cheshire 2007).

The problem with estimating neighbourhood effects on, for example, individual income, is that people are non-randomly allocated to neighbourhoods; people select into neighbourhoods based on their personal preferences and resources, in combination with housing availability. That is, people tend to move to neighbourhoods where affordable dwellings are available, match their tenure preferences, and where landlords will not discriminate against them. As a result of this selection process, it is likely that parameter estimates of neighbourhood effects are inflated as the characteristics that drive households into certain areas are highly correlated with the outcomes of interest to most researchers. Several econometric techniques have been proposed to correct for selection effects, for example by using instrumental variables, or through fixed effects models that hold constant time-invariant factors that presumably vary across households. Although these techniques can reduce selection bias, there is no perfect fix to completely rule out threats posed by endogeneity (Harding, 2003; Boschman, 2015). Perhaps more importantly, controlling for neighbourhood selection using such approaches is suboptimal as the processes that funnel certain households into particular neighbourhoods are theoretically meaningful and should be modelled explicitly (Hedman and van Ham 2012). Instead of treating neighbourhood selection as a nuisance which needs to be controlled away, we present an empirical framework which directly incorporates neighbourhood selection in models of neighbourhood effects (see also van Ham and Manley 2012).

There are only a few studies which have attempted to model neighbourhood choice to correct for selection bias in models of neighbourhood effects (see Hedman and Galster 2011; Ioannides and Zabel 2008; Sari 2011). Following Ioannides and Zabel (2008) we model neighbourhood choice using a discrete choice model and subsequently incorporate correction components into a neighbourhood effects model of individual income from work. This approach allows us to adjust our neighbourhood effects model for selection processes driven by various neighbourhood characteristics that are assessed simultaneously and in combination. Our approach diverges from prior work using discrete choice models (Ioannides and Zabel, 2008; Hedman et al. 2011) in that we use the full choice set of available neighbourhoods in the regional housing market, instead of a random choice set. We argue that the full choice set is necessary to control for non-random selection of neighbourhoods. We estimate our models on longitudinal population data from the Netherland's Social Statistical Database (SSD), a population registry composed of geocoded individual level data covering the entire population of the Netherlands from 1999 through the present.

2. Background

The body of literature on the so-called 'neighbourhood effects' – defined here as the independent influence of the residential environment on individual outcomes – has grown considerably over the last two decades (see for review Ellen and Turner 1997; Galster 2002; Sampson et al. 2002; Durlauf 2004; van Ham et al. 2012a; 2012b; van Ham and Manley 2012; Sharkey and Faber 2014; Nieuwenhuis and Hooimeijer 2016; Nieuwenhuis 2016) and many studies have reported neighbourhood effects on outcomes such as school dropout, childhood achievement, transition rates from welfare to work, deviant behaviour, social exclusion, social mobility, and income.

Since the seminal work by Wilson (1987), theoretical explanations of neighbourhood effects have been expanded to include role model effects and peer group influences, social and physical disconnection from job-finding networks, a culture of poverty leading to dysfunctional values, discrimination by employers and other gatekeepers, access to public services, and exposure to criminal behaviour (see Galster 2012 for an excellent overview of potential causal mechanisms). The neighbourhood effects literature suggests that living in a low income neighbourhood, or a poverty concentration neighbourhood, can have a negative effect on the incomes of individuals. Various causal mechanisms could lead to such negative contextual effects on individual incomes (Galster 2012). For example, those living in poverty concentration neighbourhoods could have difficulties accessing good employment opportunities due to the spatial distribution of jobs and the lack of transportation. Also, people living in poverty concentration neighbourhoods might lack job finding networks that could help them to find (better) paid positions. Or the lack of positive role models in the residential neighbourhood might lead to negative attitudes towards paid employment. People living in poverty concentration neighbourhoods can also face discrimination from employers which reduce the probability of finding a job, or increasing earnings.

The concept of neighbourhood effects is academically intriguing, and has been embraced by policy makers to justify area-based policies (van Ham and Manley 2012). Despite the popularity of the concept, and the ever growing body of literature, there remains considerable debate on the importance of neighbourhood effects above and beyond the shared characteristics of neighbourhood residents. And although there is increasing evidence that neighbourhoods are relevant for the social and economic well-being of their residents, many studies struggle with the identification of causal neighbourhood effects as they ignore, or fail to adequately address the forces that differentially funnel certain people into particular areas (Durlauf 2004; van Ham and Manley 2010). The main problem is that people do not choose where they live at random; the neighbourhood sorting process is highly structured and often the outcome of interest (for example, income) may also be responsible for people selecting into deprived neighbourhoods in the first place (van Ham and Manley 2012). In other words, impoverished neighbourhoods may not make residents poor(er); rather, low income households tend to live in particular types of places – for instance, where rent is low, landlords are less discriminating, and most importantly, housing is available (Desmond 2016). Disentangling the shared characteristics of neighbourhood residents is paramount for understanding whether and how characteristics of residential places influence the health, safety, and economic well-being of individuals.

A growing body of literature underscores the importance of neighbourhood choice in determining the spatial distribution of households across metropolitan areas. Most studies model the probability that a household moves to a type of neighbourhood based on only one or two neighbourhood characteristics (typically the level of deprivation and/or the level of concentration of ethnic minorities) (Logan and Alba 1993; Brama 2006; Clark and Ledwith 2007). Hedman and colleagues (2011; see also Sermons 2000; Boschman and Van Ham 2015) took a different approach. Following Ioannides and Zabel (2008) and Quillian and Bruch (2010), they applied a conditional logit model (McFadden 1974) which allowed for multiple characteristics of destination neighbourhoods that are assessed simultaneously and in combination. The conditional logit model estimates the probability that a household chooses a certain neighbourhood from a set of alternative neighbourhoods, based on interaction effects between household characteristics and a range of neighbourhood characteristics. Using administrative data from Sweden, Hedman and colleagues report that neighbourhood sorting is a highly structured process. Households were more likely to choose neighbourhoods where the population composition matched their own social and demographic backgrounds. Income was the most important driver of the sorting process; higher income households were most likely to sort into high income neighbourhoods, and low income households into low income neighbourhoods. But also other socio-economic and demographic characteristics were important; ethnic minorities moved to neighbourhoods with higher shares of ethnic minorities and families with children to neighbourhoods with many families with children. As a result of the neighbourhood choices made by moving households, neighbourhood characteristics were reproduced over time. Hedman and colleagues (2011, p1395) are careful to point out that

> *the concept of choice needs to be used with caution. Households make choices within a restricted choice set. Choices are restricted by household preferences, resources, and restrictions, but also by constraints imposed by the structure of the housing market. It is very likely that poor households do not `choose' to move to poverty neighbourhoods, but move there because they cannot afford to live anywhere else.*

Consistent with this observation, Van Ham and Manley (2012) argued that one of the most pressing challenges for research on neighbourhood effects is to explicitly incorporate neighbourhood selection in models of neighbourhood effects. Controlling for selection effects through econometric modelling alone may not be sufficient as selection is at the very heart of understanding neighbourhood effects. They further advocate for the necessity of a theory of selection bias to help explicate the “unmeasured characteristics which cause people to move to certain neighbourhoods, and also cause people to have a certain income, health or other outcome” (van Ham and Manley 2012 p2791). There are only a few studies which have tried to explicitly model neighbourhood choice itself, and use the outcomes to correct for bias in models of neighbourhood effects (see for example, Hedman and Galster 2011; Ioannides and Zabel 2008; Sari 2011). Although there are several papers attempting to deal with the sorting problem, we will briefly discuss three different approaches.

Hedman and Galster (2011) specified a structural equation model where both neighbourhood income mix (neighbourhood sorting) and individual income (neighbourhood effects) were modelled as mutually reinforcing. This approach was designed to avoid both selection on unobservables and endogeneity due to non-random neighbourhood selection. Their results suggest that models which fail to control for endogeneity underestimate the true neighbourhood effect. In other words, the parameter estimates for neighbourhood effects were smaller in the models that did not correct for selection bias. This seems somewhat counterintuitive as one would expect that controlling for selection should reduce the effect of neighbourhood characteristics on individual outcomes.

Sari (2012) uses a different approach to address the endogeneity problem which results from the fact that residential location may be jointly determined with employment status as a result of non-random sorting. Two different models were estimated. First, a bivariate probit estimated the probability of living in a deprived neighbourhood and the probability of being employed. Second, a probit model was estimated on a sub-sample of households living in public housing, assuming that the location choice was exogenous in this sample. The results of this approach show that individual unemployment depends not only on experience and skills, but is also related to residential location (Sari 2012).

Finally, Ioannides and Zabel (2008) developed a two-step model of housing structure demand which controlled for the non-random sorting into neighbourhoods. The first step used a conditional logit model to model choice for a specific neighbourhood from a set of alternative neighbourhoods. The choice set was determined by the chosen neighbourhood in which the household lived plus a sample of 10 alternative census tracts, randomly selected from all census tracts comprising the metropolitan area. This resulted in a choice set of 11 tracts (of which 10 were random). The conditional logit model included interactions between individual characteristics and tract-level characteristics and, similar to Hedman and colleagues (2011), confirmed that individuals select into tracts with neighbours like themselves. Ioannides and Zabel (2008) subsequently modelled housing structure demand and included eleven bias correction terms, one for probability of choosing each of the alternative neighbourhoods in the choice set. Like Hedman and Galster (2011), the results from this two-stage model demonstrated that neighbourhood effects were strengthened when neighbourhood choice was controlled for (Ioannides and Zabel 2008).

The current study builds upon and moves beyond prior research incorporating selection in neighbourhood effects research. Following Ioannides and Zabel (2008) we employ a two-stage model to address (1) neighbourhood selection and (2) neighbourhood effects on individual income. We depart from prior research as we utilize a full, closed choice set of all alternative neighbourhood options within a large urban housing market. As we describe in greater detail below, this approach presents an improvement over prior research as we are able to capture all possible neighbourhood options in a large-urban housing market and thereby produce precise estimates of the selection processes that differently sort households into specific neighbourhoods. To the best of our knowledge, this is the first study to do so.

3. Data and methods

3.1 Data and research population

Our empirical analyses draw on longitudinal population data from the Netherland's Social Statistical Database (SSD), a population registry composed of geocoded individual-level data covering the entire population of the Netherlands from 1999 through the present. We append these data to neighbourhood-level information, including ethnic, household, dwelling, and income composition, compiled by Netherlands Statistics (Kerncijfers Wijken en Buurten). We focus on heads of household who moved within the Utrecht urban region during 2009. We first estimate a selection model in which household heads select their neighbourhood based on neighbourhood characteristics prior to the move (2008). We then model the effects of neighbourhood characteristics after the move (January 1, 2010), on subsequent income from work in 2013.

Our decision to focus on the Utrecht urban region is twofold. First, the neighbourhood selection models necessitate a study area that functions as a single housing market to ensure that, at least in theory, all neighbourhoods within this area are part of the choice set of moving households. Second, we wanted an area with a large variation in neighbourhood types. The Utrecht urban region, which consists of the city of Utrecht and the surrounding suburban municipalities, meets these criteria. In the Netherlands, more than 70% of moves are within urban regions (Vliegen, 2005). Within the Utrecht urban region the social housing sector uses a choice based letting system which allows applicants to bid on dwellings all over the urban region. The region is characterised by large variation in terms of ethnic composition, dwelling prices, housing tenure, and accessibility of facilities between neighbourhoods. Consistent with prior research in the Netherlands, we use administrative neighbourhoods (buurten) to reflect residential neighbourhood boundaries. These neighbourhoods are relatively small scale, administratively determined geographic areas. In urban areas, these neighbourhoods are analogous to the more familiar census-tract from US-based research, often consisting of relatively homogenous populations and comprising, on average, one-half square-kilometres in land area. There are 256 neighbourhoods in Utrecht which comprise our initial sample.

Based on the administrative data, we identified 25,643 household heads who lived in the Utrecht urban region on the first of January 2010 and who moved there from within the urban region after the first of January 2009, thus meeting our selection criteria. Households who moved to the Utrecht urban region from elsewhere were excluded from the analytic sample because we cannot assume that they only included neighbourhoods within the Utrecht urban region in their choice set. Of the 256 neighbourhoods in Utrecht, we excluded 53 because of missing data on neighbourhood average income and average dwelling values. Average incomes are only provided for neighbourhoods with at least 200 inhabitants, and average dwelling values are only provided for neighbourhoods with at least 5 dwellings. Excluding the 53 neighbourhoods resulted in the exclusion of 848 heads of household who moved to these neighbourhoods. Our modelling strategy necessitates information on the income of the household, we therefore had to exclude another 601 household heads for which there was no data available on income. This left us with an analytic sample of 24,014 individuals who lived in 203 neighbourhoods.

3.2 Modelling strategy

Our modelling strategy unfolds in two steps. We first estimate a discrete choice model in which all 24,014 household heads select one neighbourhood from a choice set of 203 neighbourhoods within the Utrecht urban region (the selection model). We assume that all households have all neighbourhoods in the region in their choice set. The model is based on interactions between personal characteristics and the characteristics of the neighbourhoods in the choice set. Following Ioannides and Zabel (2008), our selection model provides correction terms analogous to those proposed by Heckman (1979). Although Heckman correction terms are traditionally derived from Probit models estimated on dichotomous outcomes, we follow Ioannides and Zabel (2008) to estimate correction terms from a conditional logit model. These corrections terms represent the likelihood that a specific individual selects a specific neighbourhood. Because individuals can select among 203 neighbourhoods, the model yields 203 correction terms per individual. The conditional logit model has clear advantages over alternative modelling strategies as it allows us to address selection effects associated with multiple individual characteristics as well as multiple neighbourhood characteristics.

In the second step we estimate a neighbourhood effects model in which we predict individual income from work in 2013 as a function of the characteristics of the residential neighbourhood on January first, 2010. In other words, we examine the effect of neighbourhood characteristics on subsequent earnings among heads of household who moved within the Utrecht region in 2009. Our neighbourhood effects model includes the correction components derived from the neighbourhood selection model. We restrict this model to heads of household who were employed in 2013 (thus excluding students, entrepreneurs, or people on welfare benefits) as the causal mechanisms that produce neighbourhood effects on income will be different for employees than for other groups. Of the 24,014 household heads in the selection model, 13,430 were employed in 2013 and therefore included in the neighbourhood effects model.

3.3 The selection model

We use a conditional logit model to model neighbourhood selection. In this model, a household i selects the neighbourhood j with the highest utility from a choice set of J neighbourhoods. The utility of a neighbourhood depends on the neighbourhood's characteristics and the value of these characteristics to households and is therefore calculated as neighbourhood characteristics times parameters plus an error term (Hoffman and Duncan 1988; McFadden 1974). If we assume that the error term is identically and independently extreme value distributed across neighbourhoods, the probability that household *i* chooses neighbourhood *j* – thus that the utility of neighbourhood *j* to household *i* is higher than the utility of all other neighbourhoods – can be estimated. Thus, let P_{ij} denote the probability that household *i* will choose neighbourhood *j*, based on the characteristics of the of the *j*th neighbourhood (N_j), and the characteristics of the other neighbourhoods in the choice set (N_k). Following Hoffman and Duncan (1988), the conditional logit model is written:

$$P_{ij} = \frac{\exp(\beta N_j)}{\sum_{k=1}^{J} \exp(\beta N_k)} \qquad (1)$$

The utility of a neighbourhood to a specific household depends on the match between individual and neighbourhood characteristics, thus on the value of the neighbourhood's characteristics to the specific household. The selection of a neighbourhood is modelled *within* a household; therefore the household characteristics do not vary between neighbourhood options. In order to include household characteristics in the model, they must be interacted with neighbourhood characteristics. This can be included in equation 1 by letting Xi denote the characteristics of the *i*th household.

$$P_{ij} = \frac{\exp(\beta N_j X_i)}{\sum_{k=1}^{J} \exp(\beta N_k X_i)} \qquad (2)$$

All households in our model moved during the 2009 calendar year and thus selected a new neighbourhood; the selected neighbourhood is the neighbourhood where the household lived on January 1, 2010. When possible we used neighbourhood characteristics from 2008 in the selection models as presumably households select their neighbourhood based on the characteristics of the neighbourhood before they move. We model neighbourhood selection based on the following neighbourhood characteristics: household composition, housing characteristics (tenure composition, share of dwellings built after 2000), accessibility, dwelling values and the share of non-western minorities (see Table 1). It is important to measure neighbourhood characteristics before the move to avoid endogeneity problems (Manski, 1993), in other words, conflating the characteristics of the in-migrants with the later composition of the neighbourhood. The data on neighbourhood housing characteristics are, however, only available in 2009, therefore we use this information as a proxy for the housing characteristics in 2008, before the move. Characteristics of moving households might affect the neighbourhood ethnic and household composition, but cannot affect the building period or tenure composition of the neighbourhood.

These neighbourhood characteristics are interacted with personal characteristics to estimate differences between households in neighbourhood selection. We use household characteristics of the new household, after the move (thus measured on January first 2010). If households change during a move, for instance when two people start living together, or when an individual leaves the parental home, the characteristics of the new household, rather than the old household, determine residential preferences and therefore neighbourhood selection. We make the assumption that households do not experience any unexpected changes between the move (somewhere in 2009) and January first 2010. It is, however, possible that a couple that selected a new neighbourhood based on their shared residential preferences and opportunities is separated on January first 2010.

3.4 Neighbourhood effects models incorporating neighbourhood selection

The neighbourhood effect models estimate the effect of neighbourhood income, the share of non-western minorities and the share of social housing on individual income from work in 2013. We model the income for all employed persons in 2013 based on neighbourhood characteristics in 2010. We compare three different neighbourhood effects models; a model without controls, a model controlling for personal characteristics, and a model with correction components derived from the selection model in step 1 (described in greater detail in section 4.3). Both the personal characteristics and the correction components are measured at the individual level, therefore we use clustered standard errors to account for the non-random distribution of individuals across neighbourhoods.

4. Results

4.1 Descriptive statistics

Tables 1 and 2 present the descriptive statistics for the neighbourhood-level and individual variables included in the selection models, respectively. The average housing value in 2008 was 291 thousand euros. The average neighbourhood was 3.9 kilometres from a train station, had 76.2 restaurants within walking distance, and was 2 km from a highway. In the average neighbourhood, 30 percent of homes were social housing, 14 percent of homes were built in the past 10 years, 41 percent of residents were single and 12.5 percent of residents were non-western minorities. The majority of individuals in the analytic sample were native Dutch, just over half were single, roughly 35 percent were younger than 25, and the average household income was 47 thousand euros in 2010. Tables 3 and 4 present the descriptive statistics for the variables included in the neighbourhood effect models. These models are only estimated on people who work on January first 2013. Their average monthly income from work is 3,258 Euro. People were on average 31 year old, 20% were living with a partner and children and 27% with a partner, 9% were western minorities and 14% were non-western minorities.

***Table 1 about Here ***

*** Table 2 about Here***

*** Table 3 about Here***

*** Table 4 about Here***

4.2 Modelling neighbourhood selection

Table 5 presents the results from the conditional logit model in which individual and neighbourhood characteristics have been interacted to predict neighbourhood choice. Most of the parameter estimates from the resulting 11 sets of interactions are significant, demonstrating pronounced differences between ethnic groups, household types, age groups and income groups in the effects of neighbourhood characteristics on neighbourhood choice. For example, non-western minorities were the most likely to select neighbourhoods with a high percentage of minorities. Similarly, families and those over 65 were less likely to select a neighbourhood with a high percentage of non-western ethnic minorities than single people and those under 65. Based on these individual and neighbourhood characteristics we can only partly explain which neighbourhood people select. Many neighbourhoods will be similar in dwelling values, housing market composition, accessibility, household composition and ethnic composition. Whether people select one neighbourhood over a similar neighbourhood will partly be based on coincidence or on other unmeasured neighbourhood characteristics.

Table 5 about here

4.3 Calculating correction terms from the selection model

Based on the neighbourhood selection model we can predict the conditional probability that an individual will select a specific neighbourhood over all other alternative neighbourhoods. As all individuals select a neighbourhood from a choice set of 203 neighbourhoods, the selection model yields 203 predicted probabilities per individual. These probabilities reflect the likelihood that an individual will decide to live in a given neighbourhood based on his or her own sociodemographic background and the characteristics of the neighbourhood in question. Similar to Ioannides and Zabel (2008), we use these predicted probabilities to generate correction terms analogous to the more familiar Inverse Mills Ratio's (IMRs) popularized by Heckman's two-stage regression framework (Heckman 1979). These correction terms are subsequently incorporated in the model of neighbourhood effects to control for non-random selection into neighbourhoods.

As noted above, our selection model builds upon prior work by Ioannides and Zabel (2008). However, there is one very important difference in that we do not use a random choice set of neighbourhoods like Ioannides and Zabel, but we use the full choice set (and, consequently, the full range of 203 predicted probabilities). The main reason is that we believe that the correction terms to be included in the neighbourhood effects model can only be based on the full choice set, which we will illustrate below. Ioannides and Zabel used a choice set which was determined by the chosen neighbourhood in which the household lived plus a sample of 10 alternative census tracts, randomly selected from all census tracts comprising the metropolitan area. This resulted in a choice set of 11 tracts (of which 10 were random). Prior research suggests that this approach provides an effective means of estimating neighbourhood selection (see Hedman, van Ham and Manley 2011). For the selection model it does not matter whether a random or the full choice set of neighbourhoods are used; the outcomes of the selection model are identical. To support this argument, we estimated two similar neighbourhood selection models based on both a full choice set (FC) and on a random choice set (RC). For comparability, we included all 203 neighbourhoods in the RC model; however, the order of the neighbourhoods was randomized within individuals, similar to the approach used by Ioannides and Zabel (2008) and Hedman and colleagues (2011). As might be expected, the outcomes of the two selection models were identical.

However, the correction terms resulting from the random (RC) and the full choice set (FC) selection models are very different. First, if the selection model is estimated on the full choice set, the first correction term represents for every individual the likelihood of selecting the first neighbourhood. Based on the characteristics of this first neighbourhood (average dwelling values, accessibility, ethnic composition, etc.), the likelihood of selecting this particular neighbourhood will be high for certain people and low for others. Therefore it is possible to control for neighbourhood selection by including these correction terms in the neighbourhood effects model. However, if the selection model is estimated on a random choice set, the first correction term represents for every individual the likelihood of selecting the first random neighbourhood. For every individual this will be a different, randomly selected, neighbourhood, with different neighbourhood characteristics. Random neighbourhood 1 might be attractive to one household because of the relative low dwelling values and to another household because of the relatively high dwelling values. We argue that therefore predicted probabilities based on a random choice set are not effective to control for neighbourhood selection.

The 203 correction terms which are based on the FC selection model are highly intercorrelated. This makes sense as the correction terms reflect the probability that certain types of people will select certain types of neighbourhoods. For instance, ethnic minorities demonstrate a preference to live with other ethnic minorities and young families prefer to live among other young families. Some households may strongly prefer a handful of neighbourhoods and demonstrate an aversion to living in other types of areas. These preferences are strongly allocated along sociodemographic lines. Thus, in the second stage model, the correction terms display high-levels of collinearity, prohibiting the estimation of the neighbourhood effects regression models with all correction terms entered simultaneously. Given the randomized nature of the RC model, the corresponding correction terms did not display the same degree of correlation, and as a consequence, all 203 terms can be included in the neighbourhood effects model without collinearity problems (as the random nature of the sorting assures that no two correction terms are collinear with any other individuals correction terms in subsequent models).

In an effort to address collinearity issues with the correction terms based on the FC model, we performed a Principal Components Analysis (PCA) to reduce the number of variables necessary to capture all variance in the correction terms (and remedy the high degree of correlation). The model produced 8 Principal Components with Eigenvalues greater than 1.0 that collectively captured 98.7 percent of the total variance. These 8 (orthogonal) Principal Components were subsequently included as correction components in a weighted factor regression score to generate 8 correction terms to be included in the second-stage neighbourhood effects model. These correction components can be interpreted as the likelihood of a household head selecting a *certain type of neighbourhood*, instead of the likelihood of selecting *a specific neighbourhood*. For comparability reasons, we also used PCA to calculate correction components based on RC selection model. While PCA on the correction terms from the FC model yields 8 PCs, PCA on the correction terms from the RC models yields 99 PCs with Eigenvalues greater than 1.

Including the 8 correction components based on the FC selection model (using the full choice set) in the model of neighbourhood effects, significantly improves the model fit (R2 = .3478; F = 788.06; p <0.001). Inclusion of the 99 correction components based on the RC selection model in the model of neighbourhood effects does not lead to a significant improvement of the model (R2=.0518; F=1.45). While including the correction components of the selection model based on the full choice set leads to much smaller neighbourhoods effects on individual income (see next section 4.3), the 99 correction components or the 203 correction terms of the RC selection model do barely change the size of the neighbourhood effects. We therefore argue that the likelihood of selecting into a random neighbourhood is not an effective control for neighbourhood selection, and therefore correction components based on the full choice set should be used.

4.4 Estimating neighbourhood effects on income with correction for neighbourhood selection

Table 6 presents the parameter estimates from the regression model predicting log transformed individual earnings as a function of neighbourhood income, percent of social housing, and percentage of non-western minorities. The first model presents the baseline effect of neighbourhood characteristics on individual earnings. This model reveals a small, albeit statistically significant relationship between average neighbourhood income and individual income – one thousand dollar increase in average neighbourhood income is associated with an expected 2 percent increase in the annual salary of neighbourhood residents. Neither the percentage of social housing nor the share of non-western minorities emerged as significant predictors of earnings.

The second model (Table 6, Model 2) introduces the individual level covariates. The model shows that ethnic minorities have significantly lower incomes than natives. The household composition dummies show that household heads in couples and couples with children have significantly higher incomes than singles. The parameter estimates for the age-variables show that income first increases and then decreases with age. Controlling for individual characteristics reduces the parameter estimate for the average neighbourhood income on personal income by 31.8% [(.022-.015)/.022], however this effect remains significant. This suggests that the association between neighbourhood income and individual earnings can be partially explained by the personal characteristics of households most likely to live in areas with a certain income composition. Interestingly, the inclusion of the personal characteristics reveals a suppression effect – the parameter estimate for share of social housing in the neighbourhood becomes significant; suggesting that household heads that moved to a neighbourhood with high shares of social housing in 2009 have a lower income in 2013. The suppression effect is likely driven by the inclusion of ethnicity in the model. It appears that native-born Dutch tend to have lower incomes when they live in areas with a high concentration of social housing. Conversely, social housing concentration has a protective function for minorities, perhaps due to the presence of informal networks that aid in securing employment, thus increasing earnings over time. The effect of the percentage of non-western ethnic minorities remains insignificant. The inclusion of the individual-level characteristics provides a significantly better fit to the model than the baseline model with the neighbourhood characteristics alone (R^2 = .2028; F = 322.4, p <.001).

Model 3 substitutes the individual-level characteristics for the 8 correction components derived from the neighbourhood selection model. Assuming that selection processes are at play, the parameter estimates for correction components should emerge as statistically significant and their inclusion in the model should reduce the magnitude of the coefficients for the neighbourhood-level variables. Indeed, six out of the eight correction components emerge as statistically significant predictors of income, further supporting the contention that people select into neighbourhoods at least partially based on shared characteristics that will ultimately bear on their later earnings. In other words, residential preferences are strongly correlated with income. Perhaps more importantly, the inclusion of the correction components attenuates the effects of both social housing concentration and average neighbourhood income on individual earnings. The inclusion of the correction components reduces the effect of average neighbourhood income, decreasing the magnitude of the parameter estimate by 68.2 % [(.022 - .007)/.022]; however, the coefficient retains statistical significance. This indicates that while much of the relationship between neighbourhood income composition and individual earning can be attributed to the differential sorting of low income household to low income areas, neighbourhood income still has a residual effect on individual earnings. Put more simply, poor people indeed move to poor neighbourhoods, but moving to impoverished neighbourhoods further dampens future earnings potential. The inclusion of the correction components provides a better fit to the data than the baseline model ($R^2 = .3478$; F = 788.06; p <0.001).

Table 6 about here

5. Conclusions

One of the most significant challenges confronting neighbourhood effects scholars are the assorted issues with neighbourhood selection. Household are not randomly distributed across urban areas rather, individuals choose neighbourhoods based on their preferences and their income. Such non-random allocation to neighbourhoods makes it difficult to establish causal relationships between neighbourhood characteristics and individual outcomes. Where most of the literature sets out to control for selection effects, either through covariate controls or counterfactual models, we argue that processes through which certain households decide to move to certain neighbourhoods should be examined and explicitly incorporated in models of neighbourhood effects (see also Hedman and Galster 2011; Ioannides and Zabel 2008; Sari 2011; Hedman and van Ham 2012).

This paper presents an empirical framework to help disentangle selection processes in empirical models of neighbourhood effects. We build upon prior research by modelling neighbourhood choice using a discrete choice model and subsequently incorporating correction components into a neighbourhood effects model of individual income from work. In the first step we modelled neighbourhood selection for all movers and generated the conditional probability that each head of household would select a certain neighbourhood from a choice set of 203 neighbourhoods in the Utrecht urban region. Here we found, in line with previous research, that the neighbourhood selection process is highly structured and that households are likely to prefer neighbourhoods where the population composition matches their own social and demographic background.

In the second step we modelled the effect of three neighbourhood characteristics on individual income from work, where we included correction components for neighbourhood selection in our model. This approach crucially diverges from Ioannides and

Zabel (2008) in that we use the full choice set of available neighbourhoods in the regional housing market, instead of a random choice set. We showed that using this full choice set is necessary to control for the non-random selection of neighbourhoods. We found that the effect of the average neighbourhood income on individual income is reduced when controlling for the neighbourhood selection mechanism. In addition we found that the model with correction terms explains the variation in the data much better than the standard models.

The conclusion from our models is that controlling for neighbourhood selection leads to less biased neighbourhood effects. But most importantly, even after controlling for neighbourhood selection we still found a significant negative relationship between living in a deprived neighbourhood and individual income. This is an important finding as our results suggest that neighbourhood effects reflect more than the shared characteristics of neighbourhood residents; place of residence partially determines economic well-being.

Acknowledgements

We are very grateful for the help of both Yannis Ioannides and Jeff Zabel for helping to understand the detailed mechanics of their approach used in the 2008 paper. We would also like to thank both David Manley and Lina Hedman for their contributions to developing the neighbourhood choice model. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 615159 (ERC Consolidator Grant DEPRIVEDHOODS, Socio-spatial inequality, deprived neighbourhoods, and neighbourhood effects) and from the Marie Curie programme under the European Union's Seventh Framework Programme (FP/2007-2013) / Career Integration Grant n. PCIG10-GA-2011-303728 (CIG Grant NBHCHOICE, Neighbourhood choice, neighbourhood sorting, and neighbourhood effects).

References

Bolster, A., Burgess, S., Johnston, R., Jones, K., Propper, C. & Sarker, R. (2007). Neighbourhoods, households and income dynamics: a semi-parametric investigation of neighbourhood effects. *Journal of Economic Geography*, 7(1), 1–38.

Boschman, S. (2015). *Selective mobility, segregation and neighbourhood effects*. Delft: A + BE Architecture and the Built Environment.

Boschman, S. & van Ham, M. (2015). Neighbourhood selection of Non-Western ethnic minorities: testing the own-group effects hypothesis using a conditional logit model. *Environment and Planning A*, 47(5), 1155-1174.

Brama, A. (2006). “White flight? The production and reproduction of concentration areas in Swedish cities 1990–2000”. *Urban Studies,* 43, 1127–1143.

Cheshire, P. (2007). *Segregated Neighbourhoods and Mixed Communities*. York: Joseph Rowntree Foundation.

Clark, W. & Ledwith, V. (2007). How much does income matter in neighbourhood choice? *Population Research and Policy Review,* 26, 145–161.

Desmond, M. (2016). *Evicted: Poverty and profit in the American city*. Broadway Books.

Durlauf, S. (2004). Neighborhood effects. In J.V. Henderson and J.F. Thisse (eds) *Handbook of Regional and Urban Economics, V4, Cities & Geography*, 2173–2242. Elsevier: Amsterdam.

Ellen, I.G. & Turner, M.A. (1997). Does neighbourhood matter? Assessing recent evidence. *Housing Policy Debate,* 8, 833–866.

Galster, G. (2002). An economic efficiency analysis of deconcentrating poverty populations. *Journal of Housing Economics*, 11(4), 303-329.

Galster, G. (2012). The Mechanism(s) of Neighbourhood Effects: Theory, Evidence, and Policy Implications. In van Ham Manley Bailey Simpson Maclennan. *Neighbourhood Effects Research: New Perspectives*. Dordrecht: Springer.

Harding, D.J. (2003). Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology*, 109, 676-719.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.

Hedman, L. & Galster, G. (2011). Neighbourhood income sorting and the effects of neighbourhood income mix on income: A holistic empirical exploration. *Urban Studies*, 50(1), 107-127.

Hedman, L. & van Ham, M. (2012). Understanding neighbourhood effects: selection bias and residential mobility. In van Ham Manley Bailey Simpson Maclennan. *Neighbourhood Effects Research: New Perspectives*. Dordrecht: Springer.

Hedman, L., van Ham, M. & Manley, D. (2011). Neighbourhood choice and neighbourhood reproduction. *Environment and Planning A*, 43, 1381-1399.

Hoffman, S.D. & Duncan, G.J. (1988). Multinomial and conditional logit discrete-choice models in demography. *Demography*, 25, 415-427.

Ioannides, Y.M. & Zabel, J.E. (2008). Interactions, neighborhood selection and housing demand. *Journal of Urban Economics*, 63, 229-252.

Logan, J. & Alba, R. (1993). Locational returns to human capital; minority access to suburban community resources. *Demography*, 30, 243–268.

Manski, C. (1993). Identification of endogenous social effects: the reflection problem. *Review of Economic Studies*, 60, 531-542.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In *Frontiers in Econometrics* Ed. P Zarembka (Academic Press, New York) pp 105-142.

Nieuwenhuis, J. (2016). Publication bias in the neighbourhood effects literature. *Geoforum*,70, 89-92.

Nieuwenhuis, J. & Hooimeijer P. (2016). The association between neighbourhoods and educational achievement, a systematic review and meta-analysis. *Journal of Housing and the Built Environment*, 31, 321-347.

Oreopoulos, P. (2003). The long-run consequences of living in a poor neighborhood. *Quartarly Journal of Economics*, 118, 1533–1575.

Quillian, L. (2003). How long are exposures to poor neighbourhoods? The long-term dynamics of entry and exit from poor neighbourhoods. *Population Research and Policy Review*, 22(3), 221-249.

Sampson, R.J., Morenoff, J.D. & Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology*, 28, 443-478.

Sari, F. (2012) Analysis of Neighbourhood Effects and Work Behaviour: Evidence from Paris. *Housing Studies*, 27 (1), 45-76.

Sermons, M. (2000) Influence of race on household residential utility. *Geographical Analysis*, 32, 225–246

Sharkey, P. & Faber, J.W. (2014). Where, when, why, and for whom do residential contexts matter? Moving away from the dichotomous understanding of neighborhood effects. *Annual Review of Sociology*, 40, 559-579.

van Ham, M. & Manley, D. (2010). The effect of neighbourhood housing tenure mix on labour market outcomes: a longitudinal investigation of neighbourhood effects. *Journal of Economic Geography*, 10(2), 257-282.

van Ham, M. & Manley, D. (2012). Neighbourhood effects research at a crossroads. Ten challenges for future research. *Environment and Planning A*, 44, 2787-2793.

van Ham, M., Manley, D., Bailey, N., Simpson, L. & Maclennan, D. (2012a). New Perspectives. In van Ham, Manley, Bailey, Simpson L, Maclennan D (Eds) *Neighbourhood Effects Research: New Perspectives*. Chapter 1. Springer: Dordrecht.

van Ham, M., Manley, D., Bailey, N., Simpson, L. & Maclennan, D. (Eds) (2012b). *Neighbourhood Effects Research: New Perspectives*. Dordrecht: Springer.

Vliegen, M. (2005). *Grootstedelijke agglomeraties en stadsgewesten afgebakend*. Den Haag/Heerlen: Centraal Bureau voor de Statistiek.

Wilson, W.J. (1987). *The Truly Disadvantaged*. Chicago: University of Chicago Press.

Table 1: Descriptive statistics neighbourhood characteristics (N=203) (selection model)

	Mean	Std. Dev.	Min	Max
Average dwelling values (x1000) (2008)	291.3	138.5	138	1098
Restaurants within 3km (2008)	76.2	92.3	0	268.3
Distance to train station (2008)	3.9	3.4	0,3	12.2
Distance to highway access lane (2008)	1.9	0.9	0,1	6,4
Share of dwellings built >2000 (2009)	14.0	26.2	0	100
Share of social housing (2009)	30.5	24.2	0	100
Share of private rental (2009)	14.0	11.8	0	92
Share of Singles (2008)	41.3	18.5	10	97
Share of Couples (2008)	26.8	6.7	3	46
Share of Non-western minorities (2008)	12.5	12.2	0	79

Table 2: Descriptive statistics personal characteristics (N=24,014) (1-1-2010) (selection model)

Ethnicity	N	%
Native Dutch	17,283	72
Non-western minority	4,258	18
Western minority	2,473	10
Household type		
Couple with children	4,301	18
Couple	5,572	23
Single or other	14,141	59
Age		
<25	8,574	36
25-65	13,725	57
>65	1,715	7
	Mean	Std. Dev.
Gross household income (x1000)	47.1	45.8

Table 3: Descriptive statistics individual characteristics (effects model) (N=13,430)

	Mean	Std. Dev	Min	Max
Dependent variable				
Ln (income from work) 2013	7.88	.650		
Personal characteristics				
Moroccan	.039		0	1
Turkish	.023		0	1
Surinamese	.022		0	1
Antillean	.010		0	1
Other non-western	.046		0	1
Western	.093		0	1
Couple	.271		0	1
Couple with children	.200		0	1
Age	30.82	9.00	15	78

Table 4: Descriptive statistics neighbourhood characteristics (effects model) (N=200)

	Mean	Std. Dev	Min	Max
Average income	23.69	5.52	7.5	46.7
Share of social housing	.31	.24	0	1
Share of non-western minorities	.13	.12	0	.79

Table 5: Neighbourhood selection model based on interactions between personal characteristics and neighbourhood characteristics (N=24,014)

	B	p
Interactions with average dwelling values		
Non-western minority	-0,0048	0,000
Western minority	-0,0015	0,000
Couple	-0,0043	0,000
Couple with children	-0,0026	0,000
Young (<25)	-0,0023	0,000
Old (>65)	-0,0011	0,001
Household income	0,0000	0,013
Interactions with # restaurants <3km		
Non-western minority	-0,0012	0,000
Western minority	0,0002	0,508
Couple	0,0000	0,903
Couple with children	-0,0030	0,000
Young (<25)	0,0007	0,000
Old (>65)	-0,0066	0,000
Household income	0,0000	0,000
Interactions with distance to train station		
Non-western minority	-0,0443	0,000
Western minority	-0,0482	0,000
Couple	-0,0442	0,000
Couple with children	-0,0399	0,000
Young (<25)	-0,0836	0,000
Old (>65)	-0,0873	0,000
Household income	-0,0010	0,000
Interactions with distance to highway access lane		
Non-western minority	0,0571	0,029
Western minority	-0,0345	0,260
Couple	0,0564	0,013
Couple with children	0,0642	0,008
Young (<25)	-0,1307	0,000
Old (>65)	-0,0311	0,314
Household income	-0,0010	0,000
Interactions with share of building built after 2000		
Non-western minority	0,0034	0,000
Western minority	0,0002	0,868
Couple	0,0010	0,146
Couple with children	0,0010	0,128
Young (<25)	0,0037	0,000
Old (>65)	0,0041	0,000
Household income	0,0001	0,000
Interactions with share of non-western minorities		
Non-western minority	3,6129	0,000
Western minority	1,3285	0,000
Couple	0,4052	0,014
Couple with children	-0,1690	0,365
Young (<25)	0,1190	0,304
Old (>65)	-0,6407	0,012

Household income	0,0113	0,000
Interactions with share of western minorities		
Non-western minority	2,8926	0,000
Western minority	1,9160	0,003
Couple	1,7145	0,004
Couple with children	4,8522	0,000
Young (<25)	-3,8600	0,000
Old (>65)	-1,4014	0,190
Household income	0,0205	0,000
Interactions with share of social rented dwellings		
Non-western minority	0,0037	0,003
Western minority	-0,0067	0,000
Couple	0,0007	0,554
Couple with children	0,0120	0,000
Young (<25)	-0,0110	0,000
Old (>65)	0,0078	0,000
Household income	-0,0001	0,000
Interactions with share of private rental dwellings		
Non-western minority	0,0111	0,000
Western minority	-0,0054	0,073
Couple	0,0050	0,029
Couple with children	0,0143	0,000
Young (<25)	-0,0140	0,000
Old (>65)	0,0202	0,000
Household income	0,0000	0,094
Interactions with share of singles		
Non-western minority	-0,0004	0,868
Western minority	0,0131	0,000
Couple	-0,0182	0,000
Couple with children	-0,0371	0,000
Young (<25)	0,0326	0,000
Old (>65)	0,0137	0,000
Household income	-0,0001	0,002
Interactions with share of couples		
Non-western minority	0,0250	0,000
Western minority	-0,0010	0,875
Couple	0,0212	0,000
Couple with children	0,0115	0,023
Young (<25)	-0,0504	0,000
Old (>65)	0,0358	0,000
Household income	0,0002	0,000
Log Likelihood		-119218
Log likelihood 0-model		-127591
LR (chi-square test statistic)		16747
Pseudo - R2		0,0656

Table 6: neighbourhood effects on individual income (N=14,340)

	Model 1		Model 2		Model 3	
	B	p	B	p	B	p
Neighbourhood characteristic						
average income (x1000)	0.022	0.000	0.015	0.000	0.007	0.003
share social housing	-0.189	0.069	-0.179	0.005	-0.058	0.232
share non-western minorities	0.151	0.352	0.164	0.135	0.021	0.765
Personal characteristics						
Ethnicity (reference native)						
Moroccan			-0.201	0.000		
Turkish			-0.160	0.000		
Surinamese			-0.230	0.000		
Antillean			-0.232	0.000		
Other non-western			-0.221	0.000		
Western			-0.084	0.000		
Household (reference single)						
Couple			0.180	0.000		
Couple with children			0.083	0.000		
age			0.133	0.000		
age2			-0.002	0.000		
Instruments						
Component 1					0.037	0.000
Component 2					0.052	0.000
Component 3					-0.040	0.000
Component 4					0.000	0.908
Component 5					-0.001	0.856
Component 6					0.008	0.026
Component 7					0.036	0.000
Component 8					-0.025	0.000
intercept	7.423	0.000	5.110	0.000	7.743	0.000
R2		0.0416		0.2028		0.3478
F		24.10		96.25		438.22