Discussion paper

Including migration in hedonic valuation: earthquakes

On the willingness to pay for earthquake reduction in and around the province of Groningen (2012-2018)

In collaboration with Tilburg University

Matei van der Meer

December 2021

Abstract

This study goes beyond the traditional wage-hedonic models by attempting to incorporate the psychological costs of migration into a discrete choice model to value the disamenity of future earthquake risk for residents in the northern part of the Netherlands, as perceived in 2012-2018. Using individual-specific data provided by Statistics Netherlands, this research estimates house amenity valuations, income parameters and measurements for migration costs. These are used as inputs to calculate the effect of earthquake risk on the utility valuation of a region. This study evaluates whether this innovative model can contribute to the research. It is concluded that this model of discrete choice needs some major adjustments before it can be properly applied on to the housing market in the Netherlands. This study shows that, in order to make the model of residential choice work in this situation, a better specified income specification needs to be formulated, which allows individuals to differ based on region of residence.

1. Introduction

The natural gas extraction in the northern parts of the Netherlands has been a 'hot topic' since 2012. The situation in the province of Groningen has received an enormous amount of (Dutch) media attention, lawsuits, and controversies. The negative externalities caused by this extraction are often ignored in cost-benefit analyses (Koster and van Ommeren, 2015), which has brought about a severe market imperfection: the earthquakes occurring in the wake of the extraction have been damaging the homes of the inhabitants, and the victims did not get any compensation, initially. Compensation programs have been set up of late, following multiple lawsuits. One burning question that has kept the researchers busy ever since has been: compensate with how much? Compensation is one thing, but how can one specifically evaluate the exact harm (including psychological costs, next to physical damage) done to the inhabitants by the earthquakes?

A lot has already been written in recent years about the (economic) consequences of the earthquakes in the northern parts of the Netherlands. An important part of the research has been mainly descriptive and so there is still a lot to be estimated through scientific research. It would be very interesting if one could estimate a homeowner's marginal willingness to pay for a reduction in earthquake risk; this would measure the exact value of the disamenity[1] caused by this risk. If this is

[1] Where a disamenity is the exact opposite of an amenity, which is a desirable feature of a building/location

known, one can predict how beneficial a certain policy of reducing gas extraction to reduce earthquake risk would be for homeowners. Also, to be considered, compensating only physical damage may not be enough: psychological costs should be accounted for as well.
In this research I attempt to develop such an estimation of the disamenity caused by earthquake risk, including an often-overlooked variable as control variable: migration costs. These costs can be described as the 'costs of moving': almost always, there are psychological costs which derive from moving to another place. Perhaps one was born in the old house or maybe one's friends live close to the old house, or one could have chauvinistic feelings for the region where the old home is located, et cetera. These are all factors that give more value to the current place of living, that are not accounted for in traditional economic models like the hedonic price method (Roback, 1982). These traditional models typically deal only with observed amenities, like the size of a house, year of construction and the type of the house. However, estimating homeowners' marginal willingness to pay for earthquake risk reduction without controlling for migration costs could severely bias the results. Because of the psychological value that someone living in the province Groningen might attribute to his house, the economic values of the total amount of observed amenities will be lower than the actual value a person has in mind for his house. If this is the case, the value of an amenity like 'lack of earthquake risk' might be undervalued in the estimation, as people are more hesitant to move to another house compared to what a 'normal' hedonic price model would predict. So, taking into consideration the fact that the value of a house is typically estimated by calculating the value of the total collection of the amenities present in this house, I am asking myself what value can be given to the disamenity called 'earthquake risk', controlling for the costs that come with migration. Note that this disamenity is about the expected risk of future earthquakes, not the actual damage already recorded.
By calculating this value, the question I want to answer in this research can be formulated in the following way: "What is the value of the disamenity of future earthquake risk for residents in and around the affected regions in the northern part of the Netherlands, as perceived in 2012-2018?"

As far as I know, there are no research studies about the Groningen earthquakes yet, which use the specific estimation I use including migration costs. This estimation method is replicated from the research performed by Bayer, Keohane and Timmins (2009), in which the key part is covered by a discrete choice model that is estimated by maximum likelihood. I will evaluate the extent to which this method is applicable for the housing market in the Netherlands as well.
My empirical analysis will proceed in two stages. In the first stage, I start with inferring the utility associated with living in a certain region, which is derived from the fixed effect of a regression estimated on individual houses with regional dummies. After that, I use a discrete-choice-model dealing with the choice of Dutch inhabitants regarding the region where they decide to buy a house, which provides me with an estimate of the fixed effect of each region in the Netherlands. This represents the common characteristics of a region, keeping in mind that all houses within this region can differ. This controls for different compositions of the housing stock. In this choice model, I control for income by estimating the level of income an individual would have earned in every specific region. Secondly, I

control for migration costs by adding a dummy for the region of birth in the estimation. If an individual moves to another house within his region of birth, he will not experience psychological costs, whereas if this move brings him outside his birth region migration costs will apply. In the second stage, I regress the calculated estimated fixed effects on the earthquake severity in the location where the house stands. This recovers the marginal willingness to pay (MWTP) for a decrease in earthquake risk in a certain area. This stage is what would normally happen when using the traditional hedonic approach, which would regress housing prices on earthquake risk. The empirical analysis will receive more thorough attention in the method section.

Note that in order to perform this estimation, I need individual-specific data: I need the characteristics of each specific house (like size and year of construction) to calculate the regional fixed effects and the characteristics of each individual (like education, age and birthplace) to estimate income for each region and include migration costs in the choice model. In addition, I also need data on specific house locations, to determine the earthquake risk for each house. Therefore, I perform this research in cooperation with Statistics Netherlands (CBS), who provided me with both all the personal data needed during the research process, as well as a domain that grants me access to the needed capacity to perform all tasks. Statistics Netherlands is also responsible for extensive research on this topic, so in this way I can give an efficient contribution to the work already done.

My findings suggest that the model of discrete choice proposed by Bayer et al. (2009) needs some major adjustments before it can be used to answer the questions concerning the housing market in the province of Groningen. Using this method, I find a range of possible results that is way too broad (a willingness to pay ranging between -2 percent and -34 percent) and therefore not reliable. The major cause of this is the specific nature of the Dutch housing market, where people often do not live and work in the same region. This tendency biases the residential choice model as the variables determining income and house price are not only caused within a region but are affected across regions as well. A better specified income specification – that allows individuals to differ based on region (e.g. city versus village) of residence – is needed to make the model work in this situation. Secondly, the basic discrete-choice model from Bayer et al. (2009) is too much oriented on the macroeconomy to be applicable for the earthquake situation in the province of Groningen. Adjustments need to be made to make the model more oriented on the micro-scale, allowing better incorporation of earthquake risk into the model. This paper continues by providing a literature review, some descriptive statistics, further specifications on the method used, a results section, an evaluation of the use of the residential sorting model in my research and ends with a short conclusion.

2. Literature Review

The theories in the literature on which my study is based can be separated into two groups: literature on the consequences of the earthquakes in the northern parts of the Netherlands and literature on the estimation method I will use in my model. I have made this choice of literature because (from what I have seen) there are no studies about this earthquake problem in the Netherlands that use, as I do, the discrete-choice model to account for migration costs. Jansen et al. (2016) compare thirteen models with one another. These all have been used in studies concerning the earthquakes in and around the province of Groningen. This includes the study of Statistics Netherlands (2020) as well. Some of the models do cover the price hedonic model, on which half of my method is based, but none of them use the discrete-choice model. In the continuation of this literature overview, I will first give attention to the research that has been done on the Dutch earthquake problems up to now. Secondly, I will cover the method on how to determine the earthquake damage intensity for each house. Thirdly, I will describe the basics of the literature that has been written on the estimation method that I will use.

2.1 Literature on the earthquakes in the northern parts of the Netherlands

Because of its relevance, fairly much research has already been done on the (economic) impact of the earthquake problem in the northern part of the Netherlands. Examples of this are the studies by Koster and van Ommeren (2015) and Durán and Elhorst (2018).

Koster and van Ommeren (2015) study the (long-run) negative external effect of the many small earthquakes that occur because of gas extraction, also considering earthquakes that could happen in the far future on this account. They name three mechanisms on how earthquakes affect prices of houses. The first two are the past costs of actual earthquake damage (which has not been repaired yet) and the potential costs of future earthquakes, for which past earthquakes provide a signal of increased likelihood. The third mechanism is the one they focus on in their paper: previous earthquakes may signal additional non-monetary costs of future earthquakes. In theory, these psychological costs encompass all costs related to any discomfort one can think of, occurring because of earthquakes. They argue that, because of the compulsory compensation schemes provided by the NAM in the Netherlands (which also compensates for any potential future damage caused by earthquakes), this third mechanism is the most important one for researchers to look at. Using panel-data estimation techniques, Koster and van Ommeren (2015) show that homeowners experience substantial negative, non-monetary economic effects, even when their damage is fully compensated. Within their estimation, they use past earthquakes to determine the effect of perceived future risk on house prices (assuming households do this as well). They use temporal

variation in both property prices as well as in the occurrence of earthquakes, controlling for all time-invariant attributes of a region. They assume that the compensation schemes fully cover the monetary damage, which means that they can ascribe every significant fall in house prices to non-monetary costs. They estimate these non-monetary costs to be a total of 500 euros per household (these are all households, including the ones that are not affected at all) in the province of Groningen. A study by Van der Voort and Vanclay (2015) adds to this criticism of the compensation scheme by showing that households nonetheless experienced it as insufficient and time consuming. This could cause significant transaction costs that are not accounted for in the initial compensation. Following the research of Koster and van Ommeren (2015), I will look at psychological costs in my research as well. I am using a totally different approach though: they use a traditional hedonic price model (which typically focuses on only physical/monetary values), assuming the compensation scheme assures them that the found results cover all non-monetary costs, whilst I will include these psychological costs in the model itself (so I do not need to make this assumption). Also, they observe actual earthquakes to calculate earthquake risk, while I use the 'earthquake risk image' of a region to calculate this variable. This is because I believe that when it comes to earthquake risk, potential buyers of a house tend to look more to the image of a whole region instead of the factual earthquake history of a specific building.

Durán and Elhorst (2018) look at the impact of the earthquakes caused by gas extraction on house prices as well. Their study can be seen as a valuable addition to the work done by Koster and van Ommeren (2015). Like the study by Koster and van Ommeren they also look to the regions outside the province of Groningen, as people from regions close to Groningen also have reported earthquake damage. To add to the study of Koster and van Ommeren, they employ data over a longer period and control for more house and neighborhood characteristics (from 18 to 81). They can control for common factors by using microdata that was used by Koster and van Ommeren as well (provided by NVM, an association of real estate agents in the Netherlands), which consists of individual data of more than 200,000 transactions in the three most northern provinces of the Netherlands over the period 1993-2014. As this gives them the geographical location of each house, they can develop a seismological model that measures earthquake intensity/risk for each house as well, which will make their results more trustworthy (as they can include even the smallest (unfelt) earthquakes in their model). This covers the biggest difference between the two studies: Koster and van Ommeren compute a discrete variable counting the number of earthquakes that could be felt (implying that the value of a house that suffers lots of earthquakes could in theory get below zero), whereas Durán and Elhorst (2018) accumulate all earthquakes, big and small, to one certain value. Their approach consists of a spatio-temporal hedonic price model similar to Koster and van Ommeren (2015), accounting for local spatial dependence and both non-observable and observable common factors (they name the economic business cycle and population decline as examples of this). While they use a simultaneous approach, I will use a two-stage approach in my model, as this allows me to incorporate the costs of migration in the model itself. Using price hedonic modelling, they find house prices to keep decreasing through the 'earthquake years', from 3.0 percent in 2009 to 9.3 percent in 2014. Both Dutch studies use data collected by the Dutch meteorology institute (KNMI), which has

kept track of all earthquakes with a magnitude greater than 1 on the Richter scale since 1985. In my study, I will also base earthquake risk in regions on this data.

Statistics Netherlands (CBS) has already given broad attention to calculating the impact of the earthquakes on selling prices of homes. They bring out a new report each year in which they update their results, so I will use their most recent one (2020). In their research, they also use microdata from real estate agents (NVM). To determine the development of the prices they use hedonic regression when possible (whenever enough amenities of a house are known), and the Sales Price Appraisal Ratio (SPAR) method when needed (if only the WOZ[2] value of a house is observed). This SPAR method is less common. It compares actual selling prices in a certain period with estimated values in an earlier period (in the case of Statistics Netherlands, this is the WOZ appraisal value). Whenever a house is sold, the selling price is matched with the estimated WOZ value and included in the calculation. In this way, the SPAR method controls for changes in the composition of properties that have been sold in a certain period. The fact that one does not need much specific data of a home is a big advantage of this method. The disadvantage is obviously the dependence on the trustworthiness of an estimated value (like WOZ). These values are prone to biases like neglecting adjusting for positive price changes after a renovation or improvement of the building, or negative price changes like depreciation. This is why Statistics Netherlands prefers to report hedonic prices if possible. Next to changes in selling prices, Statistics Netherlands also analyses four other market indicators, including the time it takes for a house that is for sale to be sold. Observing the period of 2012-2019, they report a recovery after the first plummeting that started in 2012, but they observe this trend to be slower for the homes that are most affected by the earthquakes (CBS, 2020). They report a selling price increase of 13.3 percent of affected houses, but this is significantly lower compared to similar homes in none-earthquake-regions, where this value is 19.9 percent. Statistics Netherlands also observes that the duration before a house-for-sale is sold has decreased by around 30 percent. However, this is still significantly lower compared to similar houses in other regions where this selling duration has decreased by approximately 60 percent.

2.2 Side note on determining earthquake damage intensity per house

When it comes to determining the degree to which a certain house is affected by the earthquakes, as to how much damage it had to go through since 2012, Jansen et al. (2016) observe two approaches across the thirteen models they evaluate. The first (and most used) method distinguishes between a ‘risk-region’ and one or more ‘reference-regions’. The other approach takes every house separately and calculates the distance between the building and the epicenter of an earthquake, controlling for factors like soil composition. Statistics Netherlands (2020) uses the first approach, Koster and van Ommeren (2015) and Durán and Elhorst (2018) both use the second one. Jansen et al. (2016) name a disadvantage of the first

[2] The WOZ value is a common property valuation in The Netherlands.

approach: it assumes houses within the same region are all affected in the same way by the earthquakes, while in reality it is clear there can be huge differences in damage intensity between houses within the same region. Earthquakes do not care about regional borders. Another limitation one could think of is that both approaches use past events to predict future earthquake risk. One could wonder if this is the most accurate way to predict the future, as it might be better to use an expert's advice on each region to calculate the chances of heavy earthquakes occurring. However, when it comes to the future my research is about expected earthquake risk, not actual risk. If the people that determine the house price (house owners, consumers) expect a totally different risk compared to the actual earthquake risk, using the actual risk might even bias the results of the research. Jansen et al. (2016) notice this factor as well: where the first approach controls for 'image damage' of a certain region, the second approach does not. Consumers could avoid a whole region, just because it is known for being a 'risky area'. If this is true, the direct link between previous earthquakes and a certain building is of less importance. It might – in fact – bias the results if consumers are affected more by the image of a certain region than the earthquake history of a specific house, which is quite likely. Jansen et al. therefore remain ambiguous about which approach is best. Because I perform my study in cooperation with Statistics Netherlands (2020), I decided to use the same approach (thus, the first one based on regions) to measure risks as they do in their paper. This will make the results of my research more comparable to the findings of the research done by Statistics Netherlands and makes the differences caused by controlling for migration costs more apparent. To measure the future risk of earthquakes I also follow the method of Statistics Netherlands (2020) by looking at previous earthquake severity. Another way to do this is by looking at honored earthquake damage claims per region. However, in chapter five of their report, Statistics Netherlands (2020) shows that there is no substantial difference between using one method or the other.

A possible problem of the approach of Statistics Netherlands (2020) is the fact that the reference regions they take as a benchmark are all very close to the risky regions; often they border each other. This makes their research prone to spillover effects: inhabitants of risky regions could, because of the earthquakes, move to the reference regions, which are closest to their 'old' home. This could in theory increase demand and therefore prices in the reference areas because of the risks in risky areas. This inflates the price differences between the two regions and therefore overestimates the impact of the earthquakes. Jansen et al. (2016) call this the 'waterbed effect'. This potential problem is solved in my study by taking all regions in the Netherlands into account, not just the few in and around the risky area. It should be noted that in practice, Statistics Netherlands does not observe such a striking migration between these regions.

2.3 Literature on methods used to estimate house prices

Three important and relevant studies for the method I will use are the studies by Bayer et al. (2009), Roback (1982) and Rosen (1974). The most important one is the study by Bayer et al. (2009), as I will replicate the method they used in my paper. To find an answer to my research question, I will use their formulas in my

analysis, with the data from Statistics Netherlands concerning the needed variables from the regions in and around the province of Groningen.

The study of Bayer et al. is an empirical panel data study. Using hedonic price models and discrete-choice models, they estimate elasticity of willingness to pay with respect to air quality. With this willingness to pay, they evaluate the amenity of 'clean air', to see how important this specific amenity is for the housing location decision of people. They are the first to control for migration costs, to see if these costs influence the outcome. These migration costs can be seen as 'psychological' costs, i.e. the unwillingness of someone to move to another house because they administer psychological value to the place where they currently live. Bayer et al. find that the estimates from their newly developed model, in which they (try to) control for psychological costs, are three times greater than the marginal willingness to pay estimated by a conventional hedonic model. This finding implies that leaving out migration costs from the hedonic model could very well bias the results. In their situation, Bayer et al. find that one would underestimate the valuation of the amenity of clean air, when one does not control for migration costs. A major drawback to their report is its conciseness: the formulas are introduced very fast one after the other, with little explanation in between. This might not be a problem for initiates and attractive for journals, but it is problematic for people with a bit less professional knowledge (although some more elaboration on the mathematical part can be found in Bishop(2008). I therefore aim to make my paper more understandable for a broader public. The residential-sorting model proposed by Bayer et al. has nonetheless already been used in the last years. For example, Rodríguez-Sánchez (2014) replicates this model in order to estimate the willingness to pay for clean air enhancements in Mexico.

The papers of Roback (1982) and Rosen (1974) are fundamental studies within the field of hedonic price models. Bayer et al. (2009) builds upon their work as well. Roback studies the role of the two basic components of wage and rent in the housing location choice of a person. These locations are characterized by different quantities of amenities. She finds a positive effect of a productive amenity (like 'clean air' or 'lack of earthquake risk') on the rent gradient, while the gradient of wage is ambivalent. She shows that regional wage differences can be largely explained by differences in local amenities. This means that not only the value of the land on which a house is build is affected by regional amenities, but wage is as well.
So, in this paper the importance of amenity valuation within, for example, the housing market is clearly shown. Rosen (1974) creates a spatial equilibrium in which the set of hedonic prices guides the consumers and producers to choose a location with fitting characteristics. These characteristics (amenities) and product prices define this set of hedonic prices. In this way this research creates the hedonic pricing model that Bayer et al. (2009) and Roback (1982) also use. This hedonic pricing model is greatly applicable to the housing market. A house can be viewed as a bundle of characteristics, a list/vector of objectively (like number of rooms, year of construction) and subjectively (like the beauty of the surrounding area) measured amenities. House price is thought to depend on this bundle of characteristics that together form the house. Economic agents base their

locational decision on this price. Here it should be noted that as Bayer et al. (and thus me as well) propose an alternative method to estimate hedonic prices, the method used by Roback (1982) and Rosen (1974) covers only half of my analysis. It does cover the hedonic price model which I will use in the second stage of my method, but it does not include the discrete-choice model that I will use in the first stage of my research project.

3. Data description

I perform my research in cooperation with Statistics Netherlands, who provided me with individual specific (private) data on each Dutch individual. Statistics Netherlands also supported this research by providing a personal domain with enough computational power to run the relevant functions and by providing staff who had permission to view my private data and give feedback. The data can be roughly divided into two parts: data on houses and individuals and data on regions and earthquakes. After matching all datasets the dataset consists of 15,442.861 observations, one for each Dutch inhabitant.

3.1 Data on houses and individuals

The data from Statistics Netherlands covers two specific moments: the first of January in 2012 and the first of January in 2018. The big earthquake near Huizinge in March 2012 can be seen as the unofficial start of the gas extraction controversy, which lasts to this day. Therefore, January 2012 is the moment closest to the start of the controversy while it did not really start yet. Obviously, the earthquakes where already present for more than a decade, but it was the earthquake at Huizinge that made the general public aware of how sizeable the damage done in the earthquake prone regions actually was. I take the year 2018 as the other date, as this is the most recent moment at which reliable data on all the needed variables is available at Statistic Netherlands. It is important to take the most recent moment possible, because it typically takes time before people decide to move out of a place. It is therefore plausible that house prices will adjust with a considerable delay to the actual effects of the earthquakes on the value of housing. Also, it is to be expected that the buildings that are 'better off' in the earthquake prone regions will be sold faster than the 'losers', which also causes the effect of the problem to change over time.
I further specify the dataset by only looking to the younger people, who were born between the first of January of 1972 and December 31 of 1990 (that is: everyone between the age of 22 and 40, measured at 01-01-2012). This ensures that the people who live in a certain region, chose this region based on its current characteristics. People of age who moved to a region 50 years ago might have made their decision based on totally different characteristics than what is at hand in the present time. Also, this makes my results more comparable to the study of Bayer et al. (2009), which only accounts for individuals below the age of 35. Lastly, I filter a handful of people of whom region of birth and/or income is unknown out of the dataset. The same is done to dwellings that are not classified as a residence or have a WOZ-value that is under zero. After these specifications, 22.6 percent of the initial dataset is left over. Some descriptive statistics on this resulting data are displayed in Tables 1 (for households) and 2 (for individual-specific characteristics). Table 1 gives the median value (or average in percentages, when indicated as '%') and the value at a quarter and three-quarters of the total population for each variable of interest, in both years. The table shows how the same people in the dataset got richer in 2018 compared to 2012, with higher household incomes and

a shift to more expensive property (like a higher share of people living in a detached house, and a higher share who owns a house). This is to be expected, considering the economic growth in The Netherlands in this period and the fact that people tend to become richer as they get older.

Table 1: Summary statistics households

Variable	2012			2018		
	M	*25%*	*75%*	*M*	*25%*	*75%*
1. Household income (in euro)	64912	40276	92998	79666	51758	110805
2. WOZ-value (in euro)	193000	148000	253000	203000	152000	275000
3. Year of construction	1972	1953	1989	1975	1956	1995
4. Living area (m^2)	106	82	131	111	87	137
5. Number of residents	2.8			3		
	%			*%*		
6. Building types						
Apartment	31			27		
Attached – Terraced	35			37		
Attached – Terraced corner	13			15		
Semi-detached	8			9		
Detached	9			11		
Farm	0.7			0.4		
7. Owned	60			66		

Table 2: Summary statistics individuals

Variable of interest	Share of total dataset	
	%	*M*
Female	50.5	
Age (median)		31.0 (min 20.0, max 40.0)
Migration background[a]	24.0	
Higher education[b]	38.0	

[a] A dummy variable that turns to one if at least one parent or grandparent of an individual was not born in the Netherlands.

[b] A dummy variable that turns to one if an individual finished college education, that is: either HBO-level or university level.

3.2 Data on regions and earthquakes

The data from Statistics Netherlands allows me to distinguish between regions on several levels. Five ways of dividing The Netherlands into sub-regions are given in Table 3.

Table 3: The Netherlands in sub-regions

Region type[a]	Dutch translation	Number of residents (in thousands)	Resulting number of regions
Province	Provincie	400 - 3700	12
COROP[b]	COROP	100-1400	49[c]
Municipality	Gemeente	1 - 900	355
District	Wijk	~3	~4000
Neighborhood	Buurt	~1	~13000

[a] Note that every province consists of one or more COROP regions, which consist of one or more municipalities, and so on. See Appendix I (Figures A.1 and A.2) for a graphical illustration.

[b] COROP regions are classified based on a core city and its area of major influence.

[c] Note that the official amount of COROP regions in the Netherlands is 40. I however divide the COROP regions (3 in total) that form the province of Groningen into the municipalities of which they consist (12 in total). This allows me to still control for differences in earthquake risks between the municipalities of interest, in the province of Groningen.

Note that the values for districts and neighbourhoods are approximations, as the number of inhabitants can vary a lot based on the exact location. Some have virtually no inhabitants, some neighbourhoods still have more than 20,000 inhabitants. More information on region types can be recovered from the KWB publications from Statistics Netherlands (2020).

In my analysis, I run my estimations based on both municipalities as well as COROP regions, to see if this causes major differences. The obvious advantage of using municipalities is that smaller regions make the differences between houses caused by area-specific variables within a region smaller. The most evident disadvantage of using smaller areas is that there are less observations per area used in the estimation. The functions I use (that are further specified in the method section) allow for a limited number of observations. This means that estimating on municipalities rather than COROP regions decreases the amount of observations per area in the function by about seven times (355/49). This will reduce the accuracy of the estimates of the regions. For this reason, I also produce estimates based on COROP regions, to see if this results in big differences between the coefficients.

Two issues concerning the data description of my research remain: how to specify the earthquake risk of a certain region and additional comments on the measure for psychological costs: region of birth versus region of living. As argued in the literature review, I will look to the 'earthquake risk image' of a region rather than the actual physical damage caused by earthquake per house per region. This means I do not have to compute a specific measure of earthquake harm myself, but I can build upon the work already done by other researchers. In their report on the method used, Statistics Netherlands (2020) provides a nice overview of calculated earthquake risk per neighbourhood (see Appendix II, Figure A.4). They use data provided by the KNMI (the Dutch meteorological institute) on the prevalence of (relatively) strong earthquakes with their exact location in the northern parts of the Netherlands since the situation started. Figure A.4 displays four different risk regions – from high to no risk – and two types of regions that are left out of the estimation. This is either because these regions consist of

unique characteristics (such as neighbourhoods in the big city of Groningen) or if they are to0 far away from the region of interest. As I perform my estimation based on municipalities (this is also relevant for the province of Groningen when pooling municipalities to COROP regions, see Table 3c), I will need to adjust my earthquake risk classification alike. This classification, based on the research of statistics Netherlands, is shown in Table A.1, which is displayed under Appendix II. This Appendix also contains a map on the division of municipalities in the northern parts of the Netherlands (Figure A.3).

Lastly follow some additional comments on my measure for psychological costs: birth region versus the region of living. Bayer et al. (2009) argue that people display a strong preference for living in the region where they were born, based on their data. A similar analysis on all individuals in The Netherlands gives more evidence for this. The results are reported in Table 4. Here, the province of residence of an individual out of my dataset is placed on the y-axis and his/her province of birth on the x-axis. The numbers in bold indicate that people, indeed, show a strong preference for living in the region of birth. People from the south of the Netherlands are the most 'chauvinistic', with a share of 80 percent of the inhabitants who live in the place they were born. The share of the province of interest of this study (Groningen) is 60 percent. Going to a smaller level, Figure A.5 (Appendix III) displays this same balance of birth region versus region of residence in the ten largest municipalities of the Netherlands[3], for individuals born before 1996. The corresponding share for the municipality of Groningen is 30 percent. These statistics substantiate the claim that controlling for region of birth is a good first attempt to cover psychological costs, albeit far from perfect.

Table 4: Mobility between regions: the ratio birth province – province of residence (2018)

	Birthprov												
Residen-ceprov$_{2018}$	Gron	Frie	Dre	Over	Flev	Geld	Utre	Noho	Zuho	Zeel	Brab	Limb	Un-known
Gron	**.60**	.05	.08	.01	.02	.007	.009	.009	.008	.003	.003	.002	.02
Frie	.04	**.69**	.04	.01	.04	.006	.007	.01	.008	.005	.002	.003	.01
Dren	.13	.03	**.60**	.03	.03	.009	.007	.009	.01	.008	.003	.002	.01
Over	.03	.03	.08	**.70**	.07	.04	.02	.01	.01	.007	.006	.007	.05
Flev	.01	.02	.02	.02	**.56**	.01	.02	.06	.009	.006	.004	.003	.03
Geld	.04	.04	.06	.10	.08	**.71**	.11	.04	.04	.04	.04	.03	.08
Utre	.03	.03	.03	.03	.04	.06	**.60**	.05	.04	.03	.03	.02	.08
Noho	.05	.06	.04	.04	.09	.04	.08	**.70**	.06	.05	.03	.03	.22
Zuho	.04	.03	.04	.03	.04	.04	.08	.07	**.73**	.09	.04	.03	.31
Zeel	.002	.002	.004	.002	.008	.005	.008	.005	.01	**.68**	.01	.004	.02
Brab	.02	.02	.01	.02	.02	.05	.04	.03	.05	.08	**.80**	.07	.11
Limb	.004	.004	.007	.007	.008	.02	.009	.008	.009	.006	.03	**.80**	.05

[3] The municipality of Almere is filtered out of this, as the first homes in the region were only build from 1976 onwards.

4. Method

4.1 The basis of the model

Recall my research question: "What is the value of the disamenity of earthquake risk for inhabitants of the affected regions in the northern part of the Netherlands, as perceived in 2012-2018?" To answer this question, I replicate the structure of Bayer et al. (2009). Their model is suitable for my study as well, as their research question is similar. The amenity they are interested in ('the lack of clean air') is also a negative externality produced by firms: an exogenous shock to the housing market. My empirical analysis will proceed in two stages. The first stage is based on the location decision of individuals, the second on the valuation of the commodity of interest (in my case, earthquake risk). The location decision is determined by a utility maximization problem: using a discrete-choice model I calculate the fixed effects per housing region by maximum likelihood. These fixed effects are a measure of the utility level associated with living in a specific region in the Netherlands (i.e. the valuation of attributes of which a specific region consists). The intuition behind the discrete choice model in my situation can be grasped as follows: I combine every individual in the sample with each possible region (so, COROP or municipality size), creating a dataset in which every observation represents one individual-region combination. For each of these combinations, I observe migration costs and income that follow for this individual if he/she chooses to live in this specific region, as well as his/her final choice of where he/she decides to live. Using maximum likelihood, I calculate the coefficients for migration cost, income and the fixed effect of a region that best fit with all the final decisions from the individuals on where they decide to live. This answers the question: what valuation can be given to migration costs, income changes and every observed region itself, given the distribution of choices made by the people? In the second stage, I use these fixed effects for the different regions as dependent variable in a straightforward OLS regression with different amenities as independent variables (consisting of the amenity of interest: earthquake risk). In this way, I can calculate the effect of the earthquakes on the valuation of a certain region.
In the road to formulating the function for the above-named discrete choice model, the first equation of importance is a utility function for individual i living in location j. As in Roback's model (1982), utility U is determined by the consumption of a composite commodity 'C', housing expenditures as a constant fraction of income (elaboration given at (3)) 'H' and a quantity X of the amenity of interest, earthquake risk:

$$U_{i,j=} C_i^{\beta_C} H_i^{\beta_H} X_j^{\beta_X} e^{M_{ij+\xi_j+\eta_{ij}}}. \quad (1)$$

As follows from this Rosen-Roback function, the composite commodity is specific to the individual, whilst the value of earthquake risk is specified by location. In my study, location canbe the COROP-region or the municipality where a house is located, so I will assign different earthquake-risk values to each region of interest.

$M_{i,j}$ is the variable for the long-run disutility an individual experiences from moving to location j. To determine these migration costs, I observe the birth region of individual i, where a move to a place outside of this municipality/COROP-region causes a certain value of disutility, while a move from a place to another one within the same region does not cause disutility. As in Bayer et al., ξ_j captures unobserved attributes of the specific location. The individual-specific idiosyncratic component of utility of individual I in location j is represented by $\eta_{i,j}$. In this lies the definition that this component of utility is independent of the mobility costs and neighbourhood characteristics.
The next step is to define the budget constraint and incorporate this into utility function (1). This gives:

$$\max_{\{C,H,X_j\}} U(C,H;X_j,M_j) s.t. C + \rho_j H = I_j \quad (2)$$

$$H^*_{i,j} = \frac{\beta_H}{\beta_H+\beta_C}\frac{I_{i,j}}{\rho_j}. \quad (3)$$

Here, (2) is the maximization problem and (3) the result of incorporating this into function (1) and differentiating with respect to H_i, giving the optimal value for housing expenditures. $I_{i,j}$ is the variable for income and ρ_j is the price of housing in location j. From function (3), it follows that individuals spend a specific, constant fraction of their income on housing.
As in Bayer et al., I substitute for H* in utility function (1) and use (3) to get the indirect utility function (4). The function measures the amount of utility an individual can receive, given his/her location of residence j and income.

$$V_{i,j=} I_{i,j}^{\beta_I} e^{M_{ij}-\beta_H \ln \rho_j+\beta_X \ln X_{j+\xi_j+\eta_{ij}}}. \quad (4)$$

Here, the beta for income (β_I) consists of the sum of β_H and β_C. The marginal willingness to pay (MWTP) for the disamenity 'earthquake risk' (X_j) is the marginal rate of substitution between this amenity and income: $\text{MWTP}_i = (\beta_X/\beta_I)(I_{i,j}/X_j)$. As I assume β_X to be constant across individuals, the marginal willingness to pay depends on income. Now, there is obviously no data on what an individual would have earned in the regions that he did not choose. Therefore, I need to estimate what the income of each individual would have been in every separate neighbourhood. The data from Statistic Netherlands provides me with the income of an individual in his chosen region and I can observe other individuals in other regions with similar characteristics. So, I will estimate a series of location-specific regressions of incomes on income-affecting personal attributes. The regression model has the following specification[4]:

$$\ln I_{i,j,t} = \alpha_{0,j,t} + \alpha_{MIGRATION,j,t} MIGRATION_{i,t} + \alpha_{FEMALE,j,t} FEMALE_{i,t} + \alpha_{AGE,j,t} AGE_{i,t} + \alpha_{AGE^2_{j,t}} AGE^2_{i,t} + \alpha_{COLLEGE,j,t} College_{i,t} + \varepsilon^I_{i,j,t} \quad (5)$$

[4] A similar estimation can be found in the appendix from Bayer et al. (2009)

'College' displays whether one has attended and finished advanced education (in the Netherlands, this means either polytechnic or university level). Note that Bayer et al. (2009) also control for random sorting of individuals across locations. This measures the (observed) percentage of individuals with given education level ED, born in region J_E, that are found to be living in region J_D. However, this measure can only be implemented if the data contains movements from every region to every region, which is not the case in my situation. I call the calculated estimated income $\hat{I}_{i,j}$. This is composed of the predicted mean of income and an additional error term, i.e: $I_{i,j} = \hat{I}_{i,j} + \varepsilon^I_{i,j}$. Substituting this into the indirect utility function (4) and then taking (natural) logs yields:

$$\ln V_{i,j} = \beta_I \ln \hat{I}_{i,j} + M_{i,j} + \theta_j + \upsilon_{i,j}. \tag{6}$$

Here, θ_j captures all the attributes of a certain location that are constant across individuals. This parameter is given by:

$$\theta_j = -\beta_H \ln \rho_j + \beta_X \ln X_j + \xi_j. \tag{7}$$

The error term $\upsilon_{i,j}$ in (6) consists of all unobserved, idiosyncratic preferences of an individual for a certain location:

$$\upsilon_{i,j} = \beta_I \varepsilon^I_{i,j} + \eta_{i,j}. \tag{8}$$

Taking all this in mind, we come to the location decision of individuals. I assume the region an individual chooses to live in is the one that gives him/her the highest amount of utility. Using a discrete choice model of maximum likelihood, I assume $\upsilon_{i,j}$ is independently and identically distributed, so I can use a logit specification to calculate the share of the population choosing to live in a certain region j. It follows that the probability of individual i choosing to live in location j is calculated by:

$$P(\ln V_{i,j} \geq \ln V_{i,l}\ \forall \mathrm{l} \neq \mathrm{j}) = \frac{e^{\beta_I \ln \hat{I}_{i,j} + M_{i,j} + \theta_j}}{\sum_{q=1}^{J} e^{\beta_I \ln \hat{I}_{i,j} + M_{i,j} + \theta_j}}. \tag{9}$$

So, the probability of someone choosing to value a certain municipality/COROP-region higher than any other is dependent on estimated income, migration costs and the city-specific fixed effects (given by θ_j). Bayer et al. call this variable the "quality of life" that a region offers. This vector (which also comprises my amenity of interest: earthquake risk) is in this way made independent of income and migration costs. In the second stage, I will regress this estimated value on earthquake risk across regions, using (7).

Next, Bayer et al. mention two econometric issues that must be addressed when estimating the second stage, using (7). The first one is the problem that ρ_j (the price of housing services) varies with observable characteristics of a given region, while it is also likely to be correlated with ξ_j (the unobserved local characteristics). This problem is solved by moving -β_H * lnρ_j from (7) to the left-hand side of the equation. Taking (3) and rearranging, I have $\beta_H = \beta_I(\rho_j \mathrm{H}_i{}^*/\mathrm{I}_{i,j})$. Parameter β_I is estimated in the first stage, and thus the rest of the equation, $(\rho_j \mathrm{H}_i{}^*/\mathrm{I}_{i,j})$, can be

called 'the share of housing expenditure in income'. This can be set equal to the median value from the data. I will call this median value ω. Bayer et al. (2009) call the newly formed variable ($\theta_j + \beta_H \ln\rho_j$) the "housing-price-adjusted quality of life", or the "net value of living in location j after accounting for housing prices." The second econometric issue is that amenity levels could be correlated with local unobserved attributes. In Bayer et al. (2009), local economic activity could very well be positively correlated with air pollution as well as local rents and wages. A consequence of this would be that an unadjusted estimation of (7) with OLS might lead to biased results. To address this bias, two solutions are proposed: the use of the first differences approach and an instrumental variable correcting for unobserved attributes to cover air quality. However, in my case this threat of bias might be a lot less plausible, as earthquakes are not likely to be correlated with economic activity as air pollution is. Of course, the earthquakes in the province of Groningen are caused by economic activity, but I assume the share of inhabitants living in the area that participate in this specific kind of economic activity to be rather low. For this reason, I will only use the first differences modification, but I will not use any kind of IV-approach. This also makes my project more manageable, as it would be hard to come up with a proper instrument. In first differences, I will use panel data from 2012 and 2018. This all changes (7) into:

$$\Delta\theta_j + \beta_H \Delta \ln\rho_j = \beta_X \Delta(X_j) + \zeta_j \quad (10)$$

So, for example, here $\Delta\theta_j$ is formed by θ_{2018} - θ_{2012} for a particular region j.

4.2 Further specifications

To implement the basic model described above, several definitions are still needed. I start with estimating housing prices and add the estimated incomes in each COROP-region/municipality. Then, I will further define migration costs. To finish the first stage, I will use a logit model of location choice to estimate the fixed effects for each region. Lastly, in the second stage I will use these fixed effects by regressing them on the local amenities. In this analysis, i represents households, j municipalities or COROP-regions and t either the year 2012 or 2018.
To estimate housing prices per location, I use the following price-hedonic function, including regional dummies:

$$\ln P_{j,i,t} = \ln \rho_{j,t} + \lambda_{j,t}\Omega_{i,t} + h'_{j,t}\varphi_t + \varepsilon^H_{i,j,t}. \quad (11)$$

Here, the dependent variable is the value of a home owned by household i, in location j, in time t. This valuation is given by the WOZ-value (i.e., the 'appraisal') of a house. As argued in the literature section, an appraisal is not the perfect measure for the exact value of a house, but it is the most trustworthy variable that also is available for virtually all units in the Netherlands at a certain point in time. Ω is a dummy variable controlling for home ownership, with a value of '1' if the home is owned by the individual him/herself and 0 otherwise. $h'_{i,t}$ is a vector of the most important attributes of a house that affect price. This vector includes the year of construction, size, number of residents of the house, a dummy determining

the type[5] of the house and a dummy that displays whether the building got another function (like working, for a shop owner) next to living. Together with φ_t, this forms the index of housing services in both 2012 and 2018, which is defined as $H_{i,t} = e^{h'_{i,t}\varphi_t}$. $\rho_{j,t}$ is the effective price of these housing services. This is a fixed effect for every individual region (so one of the 355 (municipalities) or 49 (COROP-regions)) in the dataset, at a specific time (either 2012 or 2018).

From here on, I take a small random sample out of the population, of 5,000 individuals for municipalities and 12,000 for COROP-regions. Running a maximum-likelihood regression takes a large amount of disk storage, which is why these two values are the maximum sizes that do not cause computational problems within my personal domain provided by Statistics Netherlands. I omit the municipalities with the fewest inhabitants[6], leaving me with 338 municipalities (note that as COROP-regions have more inhabitants, I do not have to erase regions for the COROP-estimation). For the further specifications, I estimate for each individual the income he/she would earn in every separate region, and the migration costs he/she would experience if he/she moved to this specific municipality or COROP-region. This means I create a dataset with (5,000 * 338) 1,690,000 observations for municipalities and a dataset of (12,000 * 49) 588,000 for COROP-regions: for each possible individual-region combination one observation.

For income, I estimate a series of location specific regressions of incomes from a set of individual attributes, as described in the previous section. In this way I formulate income as estimated income plus an error term:

$$I_{i,j} = \hat{\mathrm{I}}_{i,j} + \varepsilon^{I}_{i,j}. \tag{12}$$

This gives every location a coefficient on the effect of age, gender, migration background and education on income in this specific region.

To calculate (psychological) costs of moving, I use the birth municipality of an individual. Aas previously mentioned, Bayer et al. (2009) show that people tend to settle close to where the household head was born, indicating a strong preference for living in the area of birth, and significant costs when moving out of this region. I calculate migration costs in the following way:

$$M_{i,j,t} = \mathrm{f}_M(d_{i,j,t}; \mu) = \mu_s d^{M}_{i,j,t} + \mu_s d^{P}_{i,j,t}. \tag{13}$$

Here, dummy variable $d^{M}_{i,j,t}$ gets a value of 1 if municipality (or COROP-region) j is outside of the birth municipality (or COROP-region) of individual i in year t, and turns to 0 if this is not the case. Dummy variable $d^{P}_{i,j,t}$ gets a value of 1 if municipality (or COROP-region) j is also outside of the birth province (which consists of 6 - 62 municipalities and between 1 and 7 COROP-regions) and 0 otherwise. I normalize the costs of migration to zero if individual i lives in his/her

[5] This distinguishes between apartments, farms, houses that are connected by multiple houses, houses that are connected to only one house and houses that are standing on their own.

[6] As these are so small that it is to be expected that they are not present in a random sample of 'just' 5,000 Dutch inhabitants.

birth region. So, after (12) and (13) are estimated, I have a dataset with all possible individual-municipality combinations in both years, with a unique income and accompanying migration cost dummy for each of them.
With all this in mind, everything is set and done to define the logit estimation of location choice. Assuming preferences to be constant over time, I estimate migration costs (μ_s), marginal utility of income (β_I) and neighbourhood-specific attributes (θ) in the years 2012 and 2018. As computational difficulties do not allow me to pool the data over the two years, I estimate separate regressions for 2012 and 2018. This translates into the following logit function of maximum likelihood:

$$L(\mu_s, \beta_I, \theta) = \Pi_i \Pi_{j=1}^{J} \left[\frac{e^{\beta_I \ln \hat{\imath}_{i,j,t} + \mu d_{i,j}^{M} + \mu d_{i,j}^{P} + \theta_j}}{\sum_{k=1}^{J} e^{\beta_I \ln \hat{\imath}_{i,j} + \mu d_{i,j}^{M} + \mu d_{i,j}^{P} + \theta_j}} \right]^{\chi_{i,j}} \tag{14}$$

Here, $\chi_{i,j}$ is a dummy function that turns one if household i chooses to live in location j. In all other cases, it turns to zero[7]. To implement this function into the calculations in Python, I use the PyLogit package, which is developed by Brathwaite and Walker (2018).
Now every region is attributed with a specific fixed effect based on location choice of individuals, it is time to use these fixed effects in the second stage. As described earlier, $\theta_{j,t}$ is a vector that consists of municipality-level (or: COROP-level) attributes for one of the two years. I denote the disamenity 'earthquake risk' in location j in 2018 by EQ_j,. Earthquake risk comes in three different degrees: high if municipality j lies in a high-risk area, low if low risk applies and zero if there is no earthquake risk in this region. As argued in the section on the data description, earthquake risk of an area is calculated in the same way as in CBS (2020). The model for the difference of the region-level attributes between the two years becomes:

$$\Delta\theta_j + \omega\Delta \ln \rho_j = \beta_{EQL}\Delta\left(\mathrm{EQ}low_j\right) + \beta_{EQH}\Delta\left(\mathrm{EQ}high_j\right) + \beta LnPop_j + \zeta_j. \tag{15}$$

LnPop is a parameter for the population of a certain region, as Bayer et al. (2009) show that this is the most significant macro-factor of influence. To keep my research manageable, I do not control for other regional amenities. Bayer et al. do include other regressors like the health care in a region, but find that population size is the most important one. As described earlier in this section on the method, ω covers 'the share of housing expenditure in income'. That is: the average share of yearly income that an individual spends on housing, each year.The mean housing expenditure[8] divided by the mean income in my dataset gives me a value of 0.24 for this parameter. This is a bit higher compared to the 0.20 of Bayer et al.

[7] Here I need to make an arbitrary normalization of one of the fixed values of a region ($\theta_{j,t}$). For the estimation I run on COROP-regions, I set 'Noord-Friesland' to zero. For the estimation on municipalities, 'Zeewolde' is set to zero.
[8] Based on a 10-year fixed mortgage rate of 3 percent, which is the average in the period 2012 - 2018

5. Results

The first outcomes of interest are the resulting values for the regression on housing services given in equation (11). They are displayed in Table A.2 (Appendix IV). The parameters have expected signs and have plausible values. The number of residents yields a higher valuation of housing services, similar to the living area of the house (both representing the size of the building). The year of construction (as a measure of quality) has a positive significant effect as well. Furthermore, all types of houses yield a significant higher housing service valuation, apart from apartments. All estimates are significant at the usual levels in statistics, meaning they are different compared to the null-hypothesis that these variables have no effect on the price of a house. Note that the intercept (as well as the value for the parameter on the premium for owned housing) corresponds to the fixed effect of the first region in the estimation (in this case the COROP-region of 'Noord-Friesland'), which is left out of the regression to identify the model. This fixed effect covers the baseline utility that is associated with living in a certain region, which is used in the final equation (15). When observing the price of housing of a chosen set of Dutch regions I do not see surprising results (that is: regions that are commonly known to be rich have high values and for poor regions this is the other way around).

Tables 5 and 6 report the estimates for the parameters from the residential choice model (14). Note that I run this regression eight times: four for each region-level (municipalities and COROP). Next to this distinction, I make three other distinctions (which is why I must run the regression 2*2*2 times): one based on year (2012 or 2018), one based on income (one allowing income to differ across regions as described under equation (5) and one where I fix income based on the current income of an individual[9]).

Table 5: Results discrete choice model based on COROP-regions (2012, 2018)

	COROP 2012				COROP 2018			
	Fixed income		Flexible income		Fixed income		Flexible income	
	Coefficient	*SD*	*Coefficient*	*SD*	*Coefficient*	*SD*	*Coefficient*	*SD*
$Migcost_{Cor}$	-2.68	.03***	-2.68	.03***	-2.58	.03***	-2.58	.03***
$Migcost_{prov}$	-1.64	.02***	-1.64	.03***	-1.66	.03***	-1.66	.03***
LnIncome			-0.61	.09***			-.63	.21***
R^2	.45		.46		.44		.44	
Number of Observations	588000 (49* 12000)							

***p < .001

[9] That is: fixing income means that I do not allow individuals anymore to earn a different income across regions, but assume that his/her income will not differ from the income currently earned, to whatever region he/she might move.

Table 6: Results discrete choice model based on municipalities (2012, 2018)

	Mun 2012				Mun 2018			
	Fixed income		Flexible income		Fixed income		Flexible income	
	Coefficient	*SD*	*Coefficient*	*SD*	*Coefficient*	*SD*	*Coefficient*	*SD*
$Migcost_{gem}$	-3.11	.04***	-3.46	.05***	-2.92	.04***	-2.93	.04***
$Migcost_{prov}$	-2.48	.04***	-1.86	.04***	-2.44	.04***	-2.44	.04***
Lnincome			-2.44	.08***			.28	.04***
R^2	.33		.31		.31		.31	
Number of observations	1690000 (338*5000)							

***p < .001

For the migration cost parameters, the results are in line with expectations. For both COROP-regions as well as municipalities, I observe a highly statistically significant utility cost that comes from moving out of one's birth region. These costs increase when moving out of the macro-region (that is: the province of birth) as well, but at a declining rate: the β coefficient of 'prov' is negative but smaller than the negative β coefficient of 'municipality' or 'corop'. Migration costs remain almost constant over time, which is conform intuition. Note that Bayer et al. (2009) pool the data over both years and thus produce one constant migration cost parameter for both years, but the difference between my coefficients is so small that I do not consider this as a bad thing[10]. One remarkable observation concerning the migration costs is the fact that they are higher in the estimation based on municipalities compared to the COROP-estimations. This suggests that people experience more disutility from moving out of a municipality than they would experience if they moved out of a COROP-region, which is practically impossible. This shows that the recovered utilities cannot be compared across different region sizes. The higher migration cost parameters of municipalities as well as the strange differences within the municipality-sized regressions in 2012 between the results with fixed income and flexible income indicate, however, a bigger problem: the sample size concerning municipality-sized regions might be too small. More on this will follow in the next section, in which I evaluate using the model of residential sorting in this research.

While the migration cost parameters still seem ain line with expectations (especially when using COROP-regions), Tables 5 and 6 display some striking and rather concerning results for the income parameters. These parameters are constant across time, but the negative signs for three out of the four corresponding values are totally counter-intuitive. These three values suggest that,

[10] In fact, this might represent reality even better, as within the age group of 22 to 40 a jump of six years further in time might make people a little less affiliated to the region they grew up as a child, thus slightly decreasing migration costs.

on average, earning a higher income in a region makes this place less attractive for an individual. In this estimation, earning a higher income corresponds with a statistically significant utility decrease. Whilst a proper number of philosophers would agree with this conclusion, the classical economic theory that I use does not. Apart from the counter-intuitive signs, a second problem concerning the income parameter is the huge difference between its value for municipalities (Table 6) in 2012 versus its value in 2018. This makes the only positive reported marginal utility of income (0.28) unreliable as well and fuels the concern that the sample size for the estimations based on municipalities is too small. Although the most important parameters of the model (the regional fixed effects) are intuitive, the strange values for the control variables make these results entirely unreliable. Before delving deeper into these problems in the next section, I still use the (probably biased) estimators recovered from the residential sorting model to calculate the earthquake risk parameters from equation (15). These results are displayed in Table 7. Here, the dependent variable is formed by the regional-specific, fixed utilities. When I use the distinction between low and high earthquake risk, I cannot reject the null hypothesis that the earthquakes in the province of Groningen do not cause utility costs for the affected residents. Estimations using different region sizes (corop vs municipalities) and/or different ways to calculate income (flexible vs fixed) do not give significant parameters for the earthquake risk variables, which means that I do not find evidence for a causal relation between the earthquakes and utility. Table 8 shows the coefficients for the dummy variable covering high earthquake risk, when the low-risk variable is omitted from the regression. It is shown that this is beneficial for the level of which the independent variable covers all variation of the dependent variable (the R-squared values slightly increase) and the significance level of the coefficient (the p-value of one of the parameters falls below the threshold of five percent).

Table 7: Results from Earthquake risk regressions

	COROP				Municipality			
	Flexible Income		Fixed Income		Flexible Income		Fixed Income	
	Coefficient	*SD*	*Coefficient*	*SD*	*Coefficient*	*SD*	*Coefficient*	*SD*
Intercept	-.21	.373	.04	.40	3.12	.29***	-3.71	.42***
$Risk_{low}$	-.03	.09	-.07	.09	-.16	.16	.37	.23
$Risk_{High}$	-.18	.10	-.16	.10	-.23	.17	-.01	.26
Lnpop	.014	.03***	-.005	.03	.02	.03	.34	.04***
R^2	.15		.05		.001		.17	
Number of observations	49		49		338		338	

***p < .001

Table 8: High Earthquake risk parameters

		Coefficient	*SD*	*p*	R^2
COROP	Flexible Income	-.16	.07	.04	.17
	Fixed Income	-.11	.07	.13	.06
Municipality	Flexible Income	-.23	.17	.17	.001
	Fixed Income	-.02	.26	.94	.17

Lastly, Table 9 reports the values for the willingness to pay for a reduction of earthquake risk that are recovered from all the different estimations. Willingness to pay is calculated by dividing the coefficient of the corresponding variable Δ(EQhighj) by the marginal utility of income (βI) that was recovered from the residential sorting model. I distinguish between three different values for marginal utility of income: -0.62 (the average of the resulting values from the two estimations on COROP-regions), 0.67 (the value recovered by Bayer et al., 2009)[11], and 1 (the value for a fixed income across regions). Note furthermore that a reduction of earthquake risk means a jump from a high-risk area to an area with no earthquakes. Ignoring the results for the first named marginal utility of income (-0.62), the WTP-values vary between a range of -2 percent to -34 percent. A concrete example might help to grasp the intuition behind these estimates. Consider two COROP-regions with different earthquake risk parameters, for example the municipality (recall that for the regions in the province of Groningen I use municipalities instead of COROP-regions) of Loppersum and the COROP-region of Zuidoost-Drenthe. An observed move from the earthquake prone region of Loppersum to the 'safe' region of Zuidoost-Drenthe in 2018 would correspond to an increase in willingness to pay ranging between 2 percent and 34 percent of the per-capita income of Loppersum (which is around 23,000 euros). This implies the benefits of safety recovered from moving to this region were worth between 460 to 7,820 euros in consumption foregone. However, all earlier named concerns do not allow me to interpret these results causally. Therefore, I will evaluate the use of the discrete choice model in my situation in the next section, to find out what could explain the difference between my outcome and the one from Bayer et al. (2009).

Table 9: Willingness to pay for earthquake risk across the estimations

	β_I = -.62	β_I = .67	β_I = 1
	%	*%*	*%*
$COROP_{flex}$	26	-24	-16
$Municipality_{flex}$	37	-16	-23
$COROP_{fix}$	18	-34	-11
$Municipality_{fix}$	3	-3	-2

[11] Hereby making the additional assumption that the marginal utility of income for a Dutchman is the same as the value for an American.

6. The residential choice model: limitations and future research.

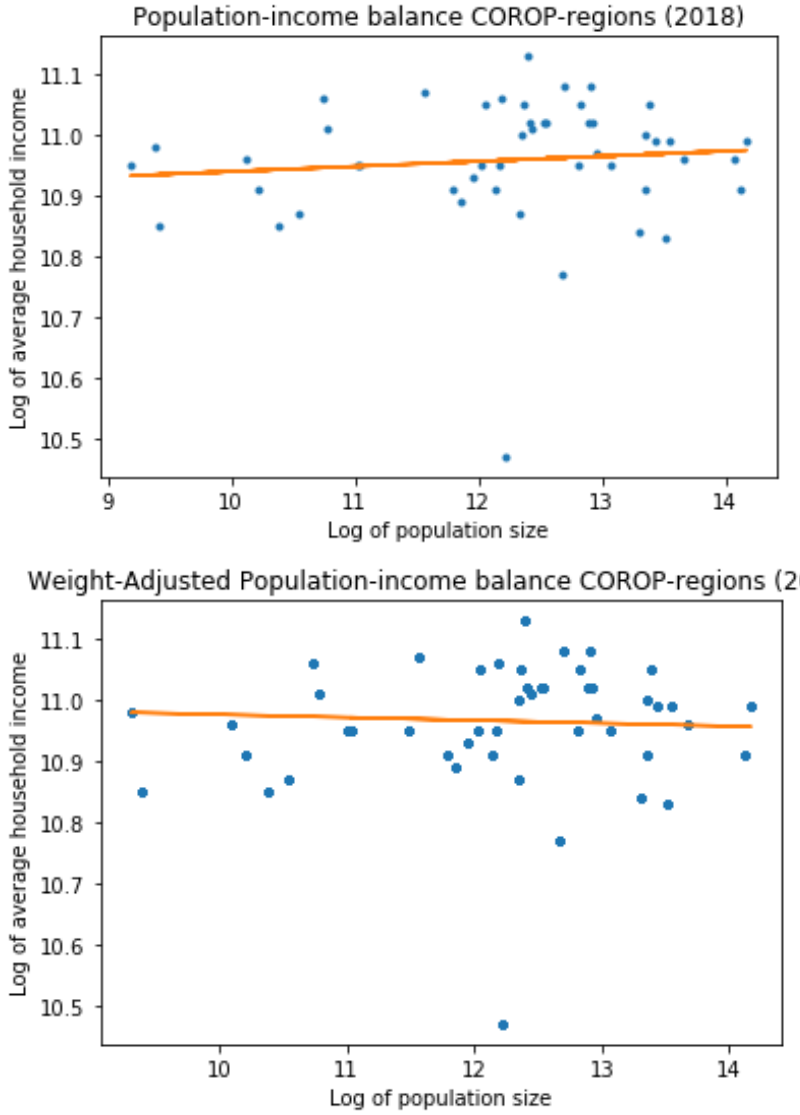
The possible caveats that come with the residential choice model as developed by Bayer et al. can be divided into two groups: problems concerning the method of estimating income (resulting in an unreliable estiate of the utility per region) and secondly problems with the discrete choice model itself (including sample size and possibly the packages used). I argue that this first group is the real troublemaker in my situation. Briefly summarized, this is caused by a negative correlation between the income and the popularity of the regions in the Netherlands: people in villages earn, on average, more than city people. The discrete choice model falsely interprets this correlation as causal, and therefore produces a marginal utility of income that is negative. I elaborate on this in the following paragraph.

6.1 Problems concerning the estimation of income

When using the model of discrete choice, the population of a region is an important determinant of the resulting fixed effect for this region. After all, the more people live in an area, the more often this area occurs in the sample (and thus the more often it is chosen)[12]. This does not bias my results, as I estimate the differences within a region, not across regions. However, it does affect the parameter for the marginal utility of income. This is because in the Netherlands, there is a negative correlation between the number of inhabitants and the average income across regions. Because the country is small and the infrastructure is good, people tend to prefer living in a different region than where their work is. This is especially the case around the big cities, which are bordered by many small-sized cities or villages in which the wealthy workers of this city take their residence. Here they enjoy both the benefits of tranquil surroundings and the comfort of living at a reasonable distance from their workplace. This problem is partly solved when moving from municipality-sized regions to COROP-regions (which are typically formed around a main city), but the negative correlation is still visible. This slightly negative correlation is shown in Figure 1. The difference between these two graphs is the fact that in the second graph, the COROP-regions are weighted according to their population size. This adjustment causes a region with a million residents to weigh ten times more than a region with a hundred thousand residents. This is what also happens in the discrete choice model: the first-named region occurs on average ten times more often compared to the latter. This causes the regression line to flip and display a (albeit small) negative correlation.

[12] This correlation between a region's popularity (fixed effect) and number of residents is shown in Table A.3, which can be found in Appendix V.

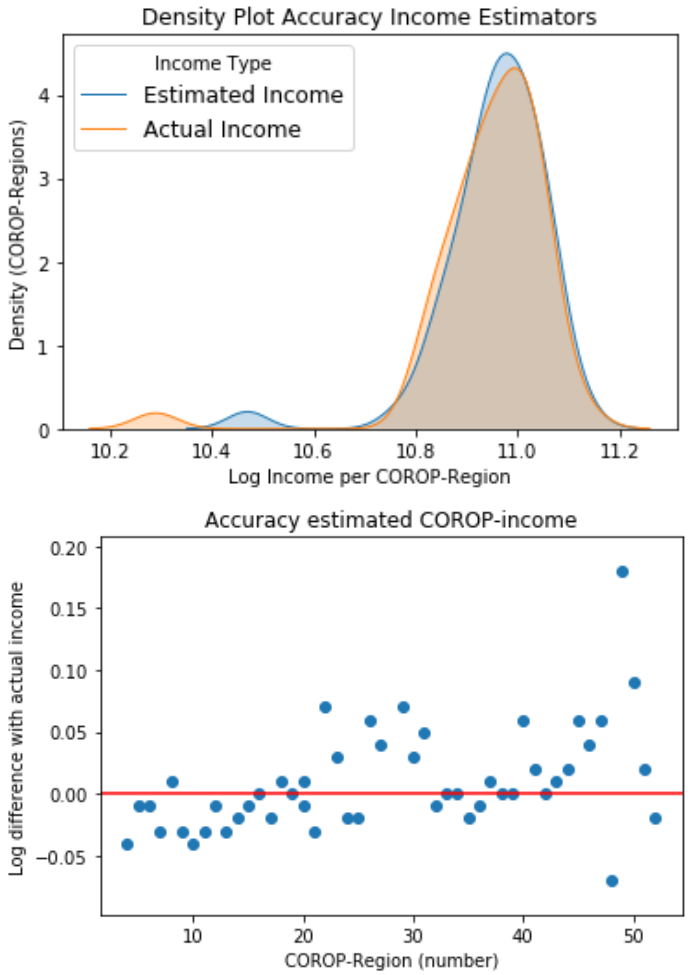
Figure 1: Population-income correlation across COROP-regions



The second question concerning income that needs to be answered is whether my estimated values are accurate enough. My income specification is relatively simple: it only covers some basic human characteristics and misses potentially important variables like the total years of followed education (as I only control for higher education with a dummy variable). Figure 2 provides two robustness checks to show that my estimators do not differ much from the actual values. For both graphs I compare the estimated values with the actual values based on the average household income for every COROP-region in 2018. To recover my estimated average household income in a COROP-region I sum all estimated individual incomes for a specific region and divide by the sample size (which is 12,000). The first graph shows how the income distribution of COROP-regions based on the estimated values is very similar to the actual data. Both measurements describe how most COROP-regions have an average log of household income between 10.9 and 11.1 (which corresponds to a range of 54,000 to 66,000 euros). The second graph shows the log difference between my estimated average income and the actual average value for each specific COROP-region. These COROP-regions are represented by numbers, which match with the

official number of Statistics Netherlands. The numbers after 40 are the municipalities within the province of Groningen. These last 12 numbers show a slightly higher variance, which might be caused by the smaller sample size of these regions. Note that at the average of 10.95, a difference of the log of income of, for example, 0.05 corresponds to a value of around 3,000 euros. All in all, these differences are not too concerning. Most important of all: they do not display a certain bias towards one side, and these differences are averaged out in total (as the first graph shows as well).

Figure 2: Accuracy Estimated income relative to actual income (COROP-Regions, 2018)



These results suggest that the counter-intuitive marginal utility of income that is recovered by the discrete choice model is not caused by a biased income specification. The most plausible conclusion therefore is that the specification as given by Bayer et al. should not be applied in the same way in the housing market of the Netherlands. The previously mentioned observation of the 'rich villager versus the poor citizen' causes the Dutch housing market to differ markedly from the American housing market. Whilst the function used to estimate income does not produce a bias on a general/regional level, it does often not properly estimate income for an individual. The first solution that comes to mind is to produce a better specified income specification. This specification would have to take into account that a non-citizen moving to the city earns a higher income than

estimated by traditional models. This can be supported by arguing that this specification looks to one specific individual, not to the general case. For example, consider differences in job types. The general case does not make a distinction in this and by this it claims that all job types are equally distributed across the country. However, it could very well be that a city contains relatively more low-paid jobs than a village. If this is true, a random selected villager moving to the city has a lower chance of performing a low-paid job, and thus a higher chance of earning a high salary, compared to a citizen moving to another city. The new specification would need to attribute a certain weight to these variables, that would (for example) make a distinction between citizens and villagers. Thereby it adjusts the general case more towards the specific situation in the Netherlands. Also, a solid basis in economic literature needs to be assembled to support this idea of allowing individuals to differ based on region type. The new specification to be developed would need to be built upon this basis.

6.2 Problems concerning the discrete choice model.

Some potential concerns about the discrete choice model need to be addressed as well. These pertain to the reliability of the results of the choice model, considering the relatively small sample size it allows. A first problem with using this model in the situation of the Groningen earthquakes is the macro-level on which this model is based, whilst the Groningen situation is oriented on a significantly smaller level. On the one hand, the division based on COROP-regions provides regions that are still too small for the discrete-choice model to be of good use (as there is too much working-living overlap between the regions). On the other hand, a division based on municipalities provides regions that are still too big to really incorporate all the different ways in which the earthquakes in Groningen affect its residents (see Figure A.4 in Appendix II). Further research on the possibilities concerning specifying the income estimation method could provide a solution in which the model can be applied at a much smaller (say: neighbourhood) level, but for now these two things cannot be combined.

Another, more practical, question is whether the PyLogit package provided by Brathwaite and Walker (2018) is the best fit for my situation (as it limits the possible sample size). This might be a valid concern, as other relevant packages do exist. However, I estimated my functions using other Python packages as well and the most important problem, the negative marginal utility of income, occurred in all of them. As these packages bring their own characteristic drawbacks, it is of higher importance to first find a solution for the major problem of income before focusing on minor problems like the most appropriate Python package to be used. Another solution for the computational obstacles is provided in the study of Berry (1994) (also referred to by Bayer et al. (2009)). This study describes an approach in which the values for the fixed effects are imputed indirectly, considerably lowering the needed computational power.

Another drawback of my research can be found in the use of region of birth to control for the psychological costs coming with migration. There are obviously great differences between the psychological valuations of people for their birth region. One might be glad to leave the parents' house, whilst another is much more homebound.

Lastly, the the lack of other macro-variables in the last equation on earthquake risk (15) should be named. Adding regional variables like the quality of healthcare, education and other factors that are named by Bayer et al. (2009) will definitely increase the reliability of the final parameter on earthquake risk.

7. Conclusion

With this study, I value the disamenity of future earthquake risk for inhabitants of the affected regions in the northern part of the Netherlands, as perceived in 2012-2018. Going beyond the traditional wage-hedonic models I control for migration costs to calculate the willingness to pay for a reduction of earthquake risk. Using individual-specific data provided by Statistics Netherlands I incorporate house amenity valuations, income parameters and measurements for migration/psychological costs into a model of discrete choice. Estimating by maximum likelihood yields the utility level associated with living in every municipality in the Netherlands. Regressing earthquake risk on these utility evaluations gives me results that are both surprising and concerning. That is, I find a negative marginal effect of income on the utility associated with a region and a broad range of unreliable values for the willingness to pay (ranging between -2 percent and -34 percent).
This suggests that the model of discrete choice proposed by Bayer et al. (2009) needs some major adjustments in order to apply it to the housing market in the Netherlands. I conclude that the tendency of Dutch people to live and work in separate regions biases the choice model, as by this behaviour the variables determining income and house price are affected across regions. The macro-economic orientation of the discrete choice model is another major problem that needs to be addressed before this method can be used for the (micro-economic sized) earthquake situation in the province of Groningen. Further research will have to formulate a better specified income specification, which allows individuals to differ based on region of residence and is adjusted in such a way that earthquake risk can be incorporated properly into the model.

Acknowledgement

I would like to thank Florian Sniekers, Barteld Braaksma and Jan de Haan for the major role they each played in the realization process of this Master Thesis. Florian Sniekers, my supervisor from Tilburg University, provided me with valuable feedback on earlier versions of this paper. Barteld Braaksma and Jan de Haan were indispensable for introducing me to the way of working at Statistics Netherlands and provided me with very useful feedback, support and supervision along the way. Furthermore, I would like to thank dr. Khulan Altangerel for her work as coordinator of the MSc thesis economics & business economics, and lastly all fellow-workers from Statistics Netherlands who contributed in some way – from building the dataset to answering my programming-related questions – to the realization of this document.

8. References

Bayer, P., Keohane, N., & Timmins, C. (2009). Migration and hedonic valuation: The case of air quality. Journal of Environmental Economics and Management, 58(1), 1-14. doi:10.1016/j.jeem.2008.08.004

Berry, S. (1994). Estimating discrete choice models of product differentiation. The RAND Journal of Economics, 25(2), 242-262. doi:10.2307/2555829

Bishop, K. C. (2007). A dynamic model of location choice and hedonic valuation (Doctoral dissertation, Duke University). Retrieved from https://are.berkeley.edu /~ligon/ARESeminar/Papers/bishop07.pdf

Brathwaite, T., & Walker, J. L. (2018). Asymmetric, closed-form, finite-parameter models of multinomial choice. Journal of Choice Modelling, 29, 78-112. doi:10.10 16 /j.jocm.2018.01.002

CBS. (2020). Kerncijfers wijken en buurten 2020. Retrieved from https://www. cbs.nl/nlnl/maatwerk/2020/29/kerncijfers-wijken-en-buurten-2020

CBS (2020). Methoderapport Woningmarktontwikkelingen rondom het Groningen veld. Retrieved from https://dashboards.cbs.nl/v2/woningmarktontwikkelingen _groningenveld/

Duran, N., & Elhorst, J. P. (2018). A Spatio-temporal-similarity and Common Factor Approach of Individual Housing Prices: The Impact of Many Small Earthquakes in the North of the Netherlands. SOM Research Reports, 2018007-EEF. Retrieved from http://hdl.handle.net/11370/ac688cbe-a965-4b19-b06f-5c705611a527

Koster, H. R., van Ommeren, J. (2015). A shaky business: natural gas extraction, earthquakes and house prices. European Economic Review, 80, 120-139. doi:10.1016/j.euroecorev.2015.08.011

Jansen, S., Boelhouwer, P. J., Boumeester, H. J. F. M., Coolen, H. C. C. H., de Haan, J., & Lamain, C.

J. M.(2016). Beoordeling woningmarktmodellen aardbevingsgebied Groningen. Retrieved from Delft University of Technology, OTB Research for the Built Environment website: http://pure.tudelft.nl/ws/files/21636634 /2016_J_ Beoordeling_woningmarktmodellen_aardbevingsgebie_Groningen.pdf

Roback, J. (1982). Wages, rents, and the quality of life. Journal of Political Economy, 90(6), 1257-1278. doi:10.1086/261120

Rodríguez-Sánchez, J. I. (2014). Do Mexicans care about air pollution? Latin American Economic Review, 23(1), 1-24. doi:10.1007/s40503-014-0009-z

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy, 82(1), 34-55. doi:/10.1086/260169

Van der Voort, N., & Vanclay, F. (2015). Social impacts of earthquakes caused by gas extraction in the Province of Groningen, The Netherlands. Environmental Impact Assessment Review, 50, 1-15. doi:10.1016/j.eiar.2014.08.008

Appendices

8.1 Appendix I: Figures A.1 & A.2

Figure A.1. [Division of the Netherlands based on COROP regions]. Reprinted from Ministry of the Interior and Kingdom Relations website, by RegioAtlas, 2021, retrieved from https://www.regioatlas.nl/kaarten#_coropregios

Figure A.2. [Division of the Netherlands based on municipalities]. Reprinted from CBS website, by CBS, 2020, retrieved from https://www.cbs.nl/nl-nl/achtergrond/2020/13/kaarten-regionale-indelingen-2020 Copyright 2021 by CBS.

8.2 Appendix II: Figure A.3, Table A.1 & Figure A.4

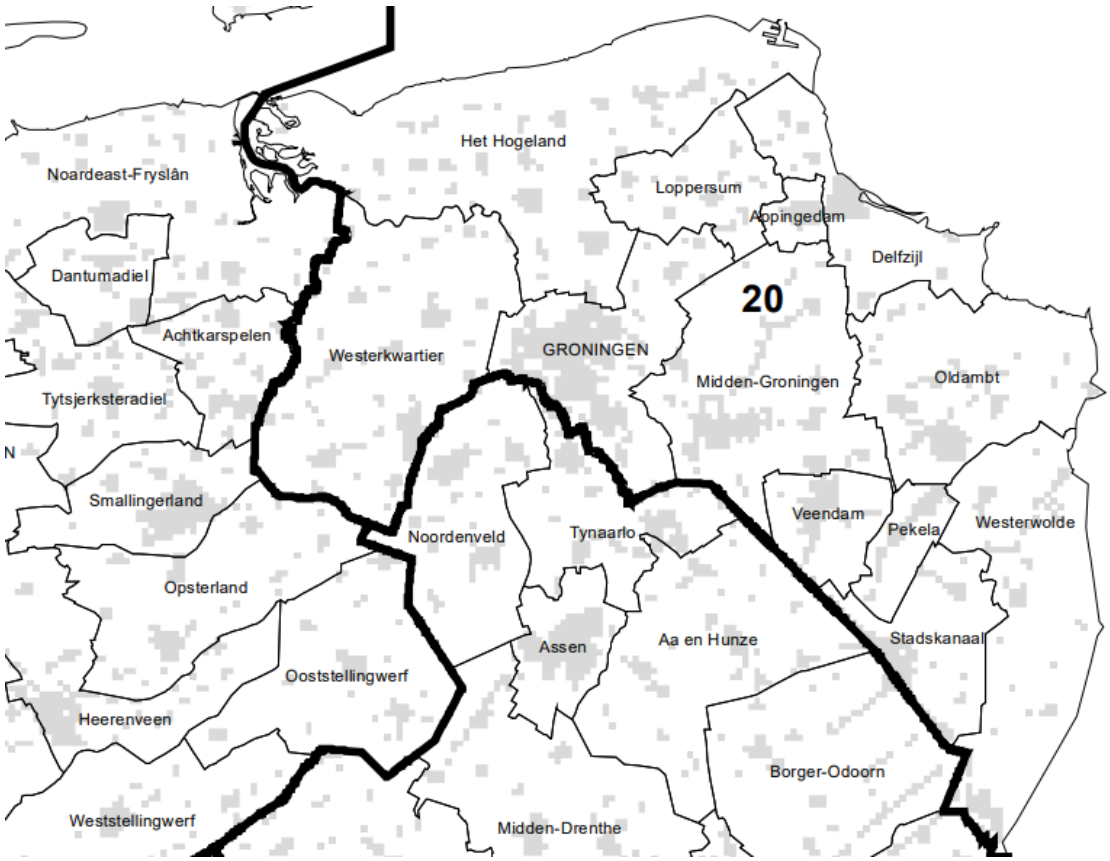


Figure A.3. [The municipalities in the northern parts of the Netherlands]. Reprinted from CBS website, by CBS, 2020, retrieved from https://www.cbs.nl/nl-nl/achtergrond/2020/13/kaarten-regionale-indelingen-2020 Copyright 2021 by CBS.

Table A.1: Earthquake risk classification

Classification	Municipality name
High risk	Het Hogeland, Loppersum[a], Appingedam[a], Delfzijl[a], Midden-Groningen
Low risk	Oldambt, Veendam, Pekela, Westerwolde, Stadskanaal, Westerkwartier
Exceptional regions (which I treat as a reference region)	Groningen
Reference regions	Rest of the Netherlands

[a] These three regions were pooled in the municipal reorganization of 2021 and categorized as one municipality: Eemsdelta. However, I will still use the old layout of three municipalities as this enables me to make a distinction between the more urban area of Delfzijl and the more rural area of Appingedam and Loppersum.

Figure A.4: Earthquake risk across the province of Groningen

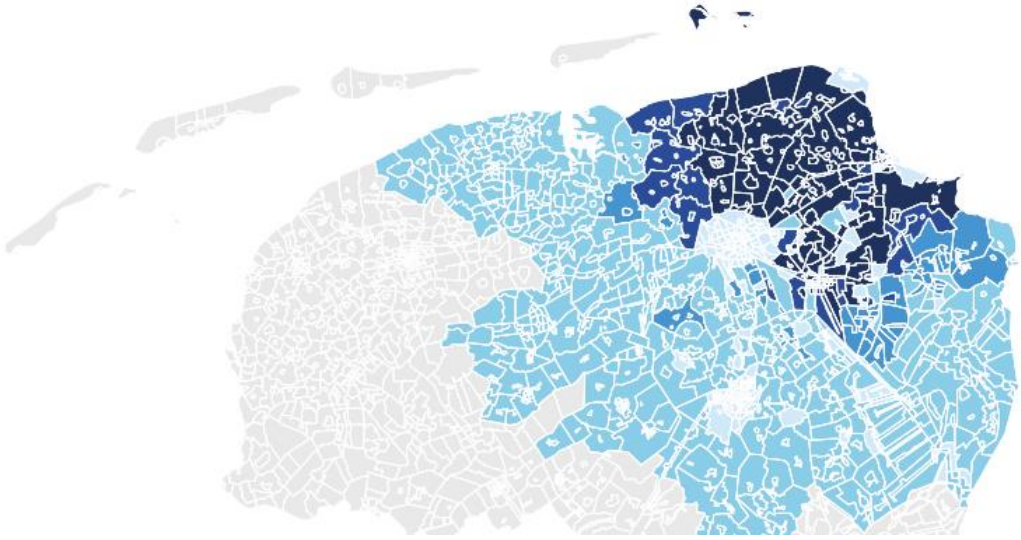
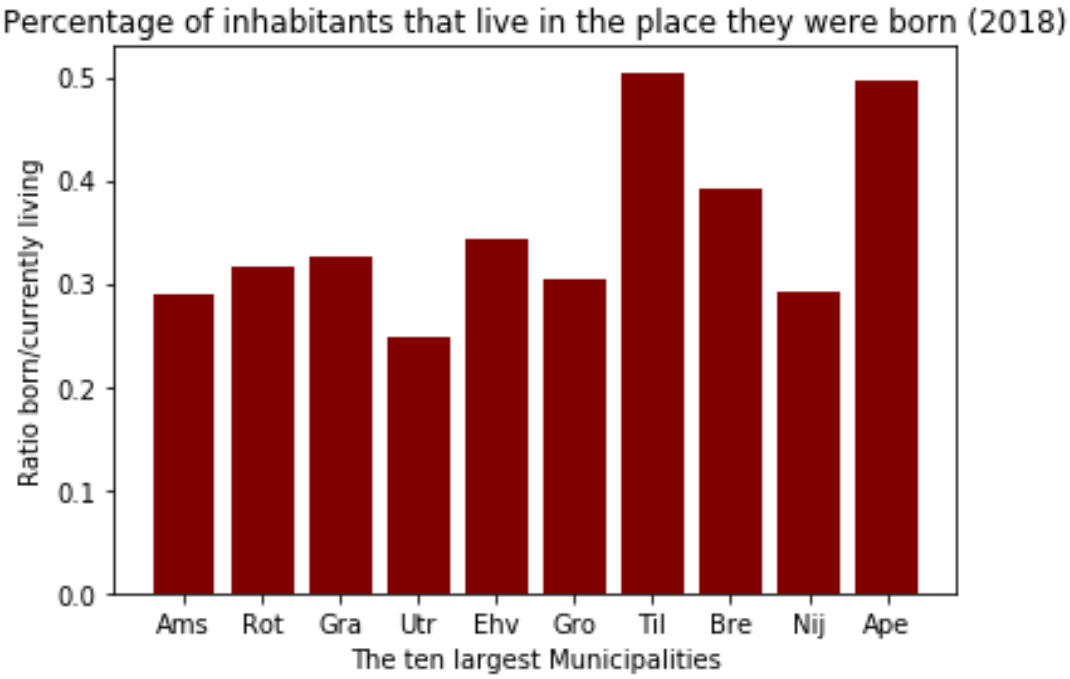


Figure A.4. [Earthquake risk in the neighborhoods of the northern parts of the Netherlands]. Reprinted from Statistics Netherlands (CBS) website, by CBS, 2020, retrieved from https://dashboards.cbs.nl/v2 /woningmarktontwikkelingen_groningenveld/ Copyright 2020 by CBS.

Note: This figure distinguishes between six regions: High risk (dark blue), medium risk (blue), low risk (light blue), reference regions (even lighter blue), exceptional regions that are left out (almost white) and regions too far away that are left out (grey).

8.3 Appendix III

Figure A.5: Birth region preference in Dutch municipalities



8.4 Appendix IV[13]

Table A.2: Parameters for the regression on house values – COROP-regions Number of observations: 100.000 randomly selected Dutch houses.

	Log house amenities 2012		Log house amenities 2018	
	Coefficient	*SD*	*Coefficient*	*SD*
Intercept	8.89	.06***	8.51	.08***
Owned	.10	.01***	.14	.02***
Number of residents	.02	.001***	.03	.001***
Year of construction	.0013	.00003***	.0014	.00003***
Living area (m_2)	.0017	.00001***	.0018	.00001***
Apartment	-.14	.007***	-.14	.05***
Attached – Terraced	.08	.007***	.06	.05
Attached – Terraced corner	.14	.007***	.12	.05**
Semi-detached	.30	.008***	.27	.05***
Detached	.57	.008***	.50	.05***
Farm	.0002	.012	-.02	.015
R^2	.59		.58	

**p < .01

***p < .001

8.5 Appendix V

Table A.3: COROP-regions ranking

Region name	Fixed effect rank	Residents rank	Region Name	Fixed effect rank	Residents rank
Utrecht	1	3	Zuidoost-Noord-Brabant	6	5
Groot-Amsterdam	2	2	Veluwe	7	7
Groot-Rijnmond	3	1	Arnhem/Nijmegen	8	6
Agglomeratie 's-Gravenhage	4	4	West-Noord-Brabant	9	9
Flevoland	5	14	Midden-Noord-Brabant	10	12

[13] The dependent variable is the WOZ appraisal value. Regional dummies are included in the regression, to recover the fixed effect of living in a certain region.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information