Commuter flow predictions in POA: Evaluation study

Maastricht University

ROA

Commuter flow predictions in POA: Evaluation study

Emiel Verkade
Jessie Bakens

ROA Technical Report

ROA-TR-2020/5

Researchcentrum voor Onderwijs en Arbeidsmarkt | ROA
Research Centre for Education and the Labour Market | ROA

Commuter flow predictions in POA: Evaluation study

Emiel Verkade
Jessie Bakens

ROA-TR-2020/5
July 2020

This research is part of Education and the Labour Market Project (POA; https://www.roa.nl/research/research-projects/project-onderwijs-arbeidsmarkt-poa), funded by Nationaal Regieorgaan Onderwijsonderzoek (NRO; file number 405-17-900), the UWV Werkbedrijf, the Cooperation Organisation for Vocational Education, Training and the Labour Market (SBB), and employment agency Randstad Netherlands. Four Dutch ministries contribute to the NRO funding: the Ministry of Education, Culture and Science (OCW), the Ministry of Interior and Kingdom Relations (BZK), the Ministry of Social Affairs and Employment (SZW), and the Ministry of Agriculture, Nature and Food Quality (LNV).

Research Centre for Education and the Labour Market
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
T +31 43 3883647 F +31 43 3884914

secretary-roa-sbe@maastrichtuniversity.nl
www.roa.nl

ISSN: 2666-884X

Content

1 Introduction

The Labour Market Forecasts (POA) by ROA are produced every two years. The goal of these forecasts is to offer insights into the medium-term (6 years ahead) prospective of the labour market, and for this in turn to aid policy-makers, students and other stake-holders in their decision-making. A key indicator resulting from POA is the indicator of future labour market position (ITA). The ITA indicates labour market prospects by field of study as a ratio of the supply of labour over the demand for labour. The ITA hereby indicates the expected development in labour market tightness by field of study.

The POA project first estimates labour market forecasts for the entirety of the Netherlands. In a subsequent step, it is disaggregated for each of the 35 labour market regions in the Netherlands and ITAs are calculated for each region. The calculated regional ITAs are then adjusted in reaction to the behaviour of commuters, and this is called the (commuter)-adjusted ITA (see Bakens et al., 2018, for more information). Imagine that the labour prospects in a certain labour market region are more favourable than in your current region. You might expect a fraction of the labour force to start working in this more favourable region, which would result in fewer unemployed workers in your region, and more employed workers in the favourable region. This results in a change in the labour market prospects for both your region and the more favourable region. Most notably, the ITA of your region may decrease (improve), whereas the ITA of the more favourable region would increase (deteriorate).

In the current methodology, the calculation of the adjusted ITA is made based on specific assumptions about how commuting affects regional imbalances on the labour market. Some of the key mechanisms in the ITA adjustment is that commuting is corrected via the inflow of workers into a region from a different region, and that the size of this adjustment depends on the estimated strength of push and pull factors between regions.

The goal of the research in this report is to evaluate the key assumptions currently made in the calculation of the commuter-adjusted ITA and test whether the right approach and parameters are used to adjust the ITA. This report is thus a validation study of the current methods used for the ITA adjustment in the regional labour market forecasts. To this end, we look towards academic models to estimate commuter flows between regions and dig deeper into the gravity models (Anderson, 2011). Gravity models exploit the relationship found in physics to model the gravitational forces of push and pull acting on two (celestial) bodies, namely that the forces acting on the bodies are related to the mass of each body, as well as the distance between them. Gravity models are often employed to represent the flow of trade between economies, and so with slight adaptations these models are used to represent the flow of human capital (be it migrants or commuters) across regions. We use the working population of each labour market region in the Netherlands as a proxy for their size as well as using the distance between them. We test different specifications and functional forms of the gravity model and select the best specification for the adjusted ITA. We also use different variables to attempt to test which variables

best explain the relationship between two labour market regions.

We find that the current way of calculating the ITA adjustment and the size of the coefficients in the calculation that we currently use for this, are in line with what we find in this study based on the literature, and in line with different model specifications and coefficients that are found in other research for the Netherlands. However the ITA adjustment could be improved if these are implemented by educational level instead of homogeneous for all commuter flows.

This paper is structured as follows; first, we explain how the ITA correction is currently calculated and which parts of this calculation are evaluated in this research. Next, we explain the best fitted model to calculate commuter flows and briefly summarise the papers that introduced them. Then, we outline how the data was gathered that is needed as inputs for the model. Next, we compare the results and their findings. Lastly, we conclude how our findings fit into the bigger picture of the POA report and the implicatons for the current calculation of the adjusted-ITA.

2 Commuter-Adjusted ITA

In this section we briefly explain how the adjusted regional ITA is calculated, and which components of this model are tested in this research. For a fully detailed description of the methods of the regional forecasts, we refer to Bakens et al. (2018).

2.1 Conceptual description

The ITA is the ratio of labour market supply and demand, and calculated as follows:

$$ITA = \frac{100 + IN\% + WLH\%}{100 + max\{0, UV\%\} + VV\% + SV\%}, \tag{1}$$

with the supply in the nominator consisting of IN, the inflow of graduates in the labour market, and WLH, the short-term unemployment. The denominator is the demand consisting of UV, the expansion demand, VV, the replacement demand, and SV, the substitution demand. Because labour is mobile, and workers can commute between labour market regions to work at the regional level, the ITA is adjusted for the mobility of workers. This is conducted by re-estimating the ITA after the first initial regional estimation of the ITA. More specifically, the re-estimation solely focuses on adjusting the regional inflow of graduates, IN, in the labour market.

In the initial fase, the first round, the inflow of graduates in regional labour market x, is calculated as:

$$IN_x = IN_{x,x} + IN_{x,y}, \tag{2}$$

with IN_x the inflow of graduates in region x, which is the region of residence, $IN_{x,x}$, the inflow of graduates residing and working in region x, and $IN_{x,y}$, the inflow of graduates residing in region x but working in region y. $IN_{x,y}$ is the commuter outflow of region x to y.

As is stated above, workers are mobile across labour market regions if perspectives are better (or worse) in one region than in another. We therefore want to correct the initially calculated ITA for this adjustment mechanism. The corrected ITA is re-estimated by adjusting the regional inflow of graduates in the labour market. This is done as follows:

$$IN_x^{'} = IN_x + dIN_{y,x}, \tag{3}$$

with $IN_x^{'}$ the corrected inflow of graduates in region x, IN_x the inflow of graduates in region x, and $dIN_{y,x}$ the adjusted inflow of graduates residing in region y and working in region x. It is this mobility adjustment, $dIN_{y,x}$, that needs

to be modeled and estimated based on empirical evidence. In this research paper, we estimate and test different empirical models that are used as input in estimating the mobility adjustment.

2.2 Technical Derivations of Mobility Adjustment in the Regional ITA

In the technical report of the regional forecasts (Bakens et al., 2018) all steps in this derivation are given in more detail, but here we describe the most important ones and the underlying intuition behind the steps. Mobility adjustments are based on theoretical models on mobility flows, and adjustments in these flows (see also the next section for a full discussion on these theoretical models).

We start with equation 4.3 from the technical report, at the start of section 4.2 of the technical report, which is as follows:

$$IN_{y,x} = A^{\gamma_1} W_y^{\gamma_2} W_x^{\gamma_3}, \tag{4}$$

where $IN_{y,x}$ represents the graduates living in region y and working in region x. Equation 4 then states that the graduates living in region y and working in region x must be a function of A, which denotes the distance between region x and y, multiplied by W_y, the size of the working population in region y, and by W_x, the size of the working population in region x. This is a simple gravity model, where A is the distance between two objects, and W_x and W_y represent the masses of the two bodies x and y respectively. We then take the logarithm of this equation to arrive at the following relationship:

$$lnIN_{y,x} = \gamma_1 lnA + \gamma_2 lnW_y + \gamma_3 lnW_x, \tag{5}$$

which now yields a linear relationship between the graduates and these gravitational factors. To see how these change over time, we take the derivative of the equation, yielding:

$$\frac{dIN_{y,x}}{IN_{y,x}} = \gamma_1 \frac{dA}{A} + \gamma_2 \frac{dW_y}{W_y} + \gamma_3 \frac{dW_x}{W_x}, \tag{6}$$

with $dIN_{y,x}$ our variable of interest, dA the change in the distance between regions x and y, dW_y the change in the working population of region y, and dW_x the change in the working population of region x. We make the assumption that the distance between regions will not change, and therefore we assume that $dA = 0$, meaning that we have no interest in or use for γ_1. We then isolate the variable of interest, yielding:

$$dIN_{y,x} = IN_{y,x}(\gamma_2 \frac{dW_y}{W_y} + \gamma_3 \frac{dW_x}{W_x}). \tag{7}$$

Since dW represents the change in the labour force of a region, we could use the established UV, which is the predicted growth of employment. However, an assumption we make is that the growth of the labour force is not solely driven by UV, but instead driven by demand in the regional labour markets, and therefore driven by the ITA's, which are also outlined in the report. Therefore, we define the relationship between the demand in a labour market region and its ITA as follows:

$$D = ITA^{-1} - 1. \tag{8}$$

Using this identity and simple manipulation as outlined in the appendix of Bakens et al. (2018), we find the following two relations:

$$\frac{dW_y}{W_y} = -D_y(1 + \frac{IN_y + WLH_y}{W_y}), \frac{dW_x}{W_x} = D_x(1 + \frac{IN_x + WLH_x}{W_x}), \tag{9}$$

where WLH denotes the short-term unemployment in the region. Substituting these relations into equation 7 yields our final equation (equation 4.9 in Bakens et al. (2018)):

$$dIN_{y,x} = IN_{y,x}[-\gamma_2 D_y(1 + \frac{IN_y + WLH_y}{W_y}) + \gamma_3 D_x(1 + \frac{IN_x + WLH_x}{W_x})]. \tag{10}$$

From equation 10 we can therefore see that the change in the working population in region x stemming from region y is a result of graduates working in region x and living in region y multiplied by the push of region y and the pull of region x. The push of region y is captured in $-\gamma_2 D_y(1 + \frac{IN_y + WLH_y}{W_y})$, where we can see that the higher γ_2, the stronger the push outwards of region y. The pull of region x is captured by $\gamma_3 D_x(1 + \frac{IN_x + WLH_x}{W_x})$, and we also see that the higher γ_3, the stronger the pull towards region x.

We can further explain this relation through two examples. In the first example, suppose region y has an oversupply of graduates ($ITA > 1$), and region x has an undersupply of graduates ($ITA < 1$). More specifically, suppose $ITA_y = 1.25$ and $ITA_x = 0.8$. Then, we have that $D_y = -0.2$ and $D_x = 0.25$. For higher gammas we see that more workers will be pushed out of region y, as indicated by the negative coefficient of D_y as well as the negative sign in front of γ_2. Furthermore, the positive sign in front of γ_3 and the positive coefficient of D_x demonstrates that region x also attracts more workers. The lower the ITA, the more favourable, and the stronger the pull of region x.

In a second example, suppose the ITA's are reversed between region x and y. So now we have $D_y = 0.25$ and $D_x = -0.2$. Then we actually see that the net coefficient of the first term in the summation of equation 10 is negative, which is also true for the second term. In other words, in this case we find

that the change in workers living in region y and working in x is negative. This is to be expected, as the labour market prospects are much better in region x than region y, so there should not be a reason for workers in region y to seek employment in region x, but they should instead seek employment in the more favourable labour market.

We need empirical models to estimate the coefficients γ_2 and γ_3 used in equation 10. The inflow of graduates that is depicted in equation 5 resembles the previously described gravity models in regional economics, and we will use these models to estimate the size of coefficients γ_2 and γ_3 and see whether the estimations of the coefficients in this study are in line with what is currently used in the regional forecasts.

3 Model Selection

A large amount of research has been done on predicting flows of various kinds between regions. There has also been a development in the way these flows can best be estimated in terms of functional forms. We discuss here some of the much used models with their advantages and disadvantages.

The research on regional flows of migrants, trade or commuters is based on gravity models. Gravity models emulate the gravitational model found in nature. The gravitational model outlines the forces pushing and pulling two large bodies to each other. The force exerted between two bodies is given by

$$F = G\frac{m_1 m_2}{r^2} \tag{11}$$

where G is the gravitational constant, r is the distance between the two objects, m is the mass of an object, and finally F is the magnitude of the force between the two objects. The idea of the traditional gravitational model has been taken and adapted to trade models for a long time. Therefore to transform this model to suit our needs, we use proxies for the mass of the two regions, as well as the distance between them. The exact nature of these specifications is outlined in section 4, where we specify our method of data collection as well as summary statistics.

It has long been the case that the gravitational model given by equation 11 was log-linearized for estimation. This is a very practical implementation of the model and easy to interpret. However, this functional form has many econometric and practical disadvantages. Of these practical disadvantages, the most pressing is that flows between entities that are equal to 0 cannot be easily included as the log of zero does not exist. This means that these flows are excluded, which results in underestimation of the importance of distance between two entities for the size of flows. Another work-around is to make the zero flows very small to log-linearize the number, but for obvious reasons, this is a second best solution in the most optimistic scenario. Silva and Tenreyro demonstrate that performing OLS on the log-linearized gravity model did not lead to proper estimates or conclusions of the underlying relationship.[1]

In this section we discuss a number of different specifications of the gravity equation that can be estimated. However, based on the literature and our results, the Poisson Pseudo Maximum-Likelihood model (PPML) as proposed by (Silva and Tenreyro, 2006) is the most preferred specification. We will briefly discuss all the model specifications that we have estimated in this paper, but only discuss the PPML model in more detail. The other model specifications can be found in more detail in the Appendix.

A way of estimation the gravity model is using Count Data Models, for example a Negative Binomial model (NEGBIN) or a Zero-inflated Poisson (ZIP) or Zero-inflated Negative Binomial (ZINB) model, as the observed flows are

[1] Because of the number of zero flows in our data and the conclusions based on Silva and Tenreyro (2006), we do not include the OLS estimation for the log-linearized model in this paper.

counts of occurrences (Greene, 2012; Winkelmann, 2008). The NEGBIN works in a similar way to a normal multiple regression, except for the fact that the dependent variable is a discrete count following a negative binomial distribution. The ZIP or ZINB model is employed when the data contains a large number of zeroes, but has more stringent assumptions on the underlying distribution of the data, which makes it more difficult to assume that the results are valid (see the Appendix for a discussion on this).

Another group of models often used originates in Spatial Econometrics (LeSage and Pace, 2008). The study of spatial econometrics is an active field of research, however we will mainly be using the findings of LeSage and Thomas-Agnan (2015) to construct a spatial econometric model. Here, we attempt to model the spatial dependence directly through adding spatial-connectivity matrices, and manipulating them in a particular manner. This is in contrast to what we look at before, where we simply looked at the spatial dependency through. The construction of the spatial-connectivity matrix and choosing the right way of operationalising relationships between regions is then very important.

Finally, Silva and Tenreyro propose to use a Poisson Psuedo Maximum-Likelihood (PPML) estimation technique to estimate gravity equations. The PPML functional form accounts for the fact that there are many zeroes in the data, is robust to heteroskedasticity in the data, and is relatively simple to estimate. The conditional expectation is assumed to be proportional to the conditional variance, and an objective function to be maximised is proposed, although no closed form solution exists. The regression is based upon the traditional gravity equation, given by:

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}, \tag{12}$$

where T_{ij} represents the trade flows between regions i and j, Y_i andY_j denote the GDP of region i and j respectively, and D_{ij} denotes the distance between i and j. The α's are the coefficients of interest. In the form of a Poisson-Pseudo Maximum Likelihood (PPML) regression, the following equation is estimated:

$$y_i = exp(\alpha + \beta \boldsymbol{x_i}) \tag{13}$$

where $\boldsymbol{x}$ consists of the logged values of the working population of the origin, the destination, and the distance between the two labour market regions. In the analysis of this model specification against other specifications we find that this specification gives the best results for our purpose as well. We will therefore put most emphasis on discussing this specification and its results.

4 Data

Figure 1: Labour market regions in the Netherlands

Source: `https://www.samenvoordeklant.nl/sites/default/files/arbeidsmarktregios_2018_september.pdf`

All regional data is collected at the level of labour market regions (AMR) as defined by the Dutch Employee Insurance Agency (UWV). Figure 1 gives an overview of this regional classification. The size of the commuting flows between AMRs is based on micro data from Statistics Netherlands (CBS) and gives the number of people commuting from their residential AMR to the AMR of work, specified by education level. The population data per labour market region was retrieved from the open data source from Statistics Netherlands, CBS Statline

(CBS (2018))[2]. This then allows us to get the total working population, defined as all inhabitants aged between 15 and 75 in each of the regions. The data allows to differentiate the population based on education in three categories: low, middle and high. For the regional ITA adjustments, we only need the middle and high educated.[3]

Next, we obtain our distance matrix D through looking at the shortest distance by road between all municipalities of the Netherlands. We do this via Google Maps, looking at every origin-destination combination of the Dutch municipalities. One assumption we make to reduce the computational time is $D_{i,j} = D_{j,i} \forall i, j \in M$, where M is the set of all municipalities. So we assume the distance between A and B is as long as the distance between B and A.[4] This reduces the amount of combinations to be checked by half. We store this in a matrix, and combine this with data about the working population of every municipality. We consider two ways to generate distances traveled between labour market regions. Firstly, we consider a simple weighted average. We look at the distance between each municipality of the origin region to each municipality of the destination region, and then divide by the number of municipalities in both regions. This gives us an equally-weighted average distance from each municipality in one region to the other, and we use this as a proxy for the distance between two labour market regions. We refer to this as Distance (Equally-weighted).

We also consider a second, more complex weighted average to get a more representative measure for the average distance traveled between regions. We use this to construct a weighted average of the distance between regions, and we store this information in the matrix G. We calculate the weighted average (WA) according to the following formula:

$$WA = \sum_{o \in R(region(o))} \frac{pop_o}{\sum_{i \in region(o)} pop_i} \times \sum_{d \in R(region(d))} D_{o,d} \times \frac{pop_d}{\sum_{i \in region(d)} pop_i} \tag{14}$$

where $o \in R(region(o))$ represents all origin municipalities in the corresponding region, and $d \in R(region(d))$ represents all destination municipalities in the matching region. $D_{o,d}$ is the distance between the municipalities o and d, and $\frac{pop_o}{\sum_{i \in region(o)} pop_i}$ represents the proportion of the working population living in municipality o against the total working population in the matching region. We refer to this measure of distance as Distance (Weighted-average).

The data, aggregated over education levels, is summarised in Table 1. We

[2] "Arbeid en sociale zekerheid" → "Arbeid en arbeidsmarkt" → "Beroepsbevolking" → "Arbeidsdeelregionen" → "Arbeidsdeelregionen; regio 2017". We then choose "Beroeps- en niet-beroepsbevolking" under the category of "Onderwerpen", and "Onderwijsniveau: middelbaar onderwijs" and "Onderwijsniveau: hoog onderwijs" under the "Persoonskenmerken" heading. Under the "Regio's" header, we select "Regiototalen" → "Arbeidsmarktregio's" and we select them all, and under "Perioden" we pick "2017"

[3] In the regional forecasts we do not report the ITA's for the low educated groups.

[4] An alternative way of measuring distance is not via kilometers traveled, but travel time. However, from Groot et al. (2012) it can be concluded that it does not matter for the results if distance or time is used in a gravity estimation.

Table 1: Summary statistics for Aggregated Data

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting flows	Persons	1225	4615	20880	0	343779
Working Population (Origin)	Persons	1225	257057.10	147645	72000	712000
Working Population (Destination)	Persons	1225	257057	147645	72000	712000
Distance (Equally-weighted)	km	1225	122.45	61.26	7.96	345.01
Distance (Weighted-average)	km	1225	121.41	62.11	6.01	345.27

Source: Own calculations based on CBS, Google Maps

have 1.225 observations (35 times 35 AMRs) in total of which 35 obervations (the within AMR commute) drop out in most specifications, with an average commuting flow of 4,615 persons between any two regions. As is shown, we also have zero flows between some regions and therefore need to incorporate this into the model as to not underestimate the effect of distance on commuter flows. Finally, the average commute in the data is over a distance of about 120 kilometers. However, as the minimum and maximum value shows, there is a rather high right-end tail in the commuting distance. The minimum distance values show that we calculated within-region commute as well based on the two types of calculating distances. Table 1 gives the summary statistics for all commuting flows, but looking at different educational levels is more informative. Tables 2 to 7 give the descriptive statistics by educational level. In the Netherlands we have 3 levels (2,3, and 4) of vocational education, and mostly levels 2 and 3 are considered together, and the bachelor is generally considered different from the master level of university education (also because there is a differentiation between universities of applied sciences and academic universities). Therefore we present three tables, one with the descriptive statistics of bachelor degrees (HBO), one with master degrees (WO), and one combined. The working population is only observed at the level of high education or middle education, and we assume these hold for all education levels within the high and middle education. We present all the different education levels for which we estimate the gravity equation in this paper.

Table 2: Summary statistics for MBO 2/3

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting flows	Persons	1225	1056	5218	0	80125
Working Population (Origin)	Persons	1225	146257	75829	46000	412000
Working Population (Destination)	Persons	1225	146257	75829	46000	412000
Distances (Equally-Weighted)	km	1225	122.449	61.261	7.961	345.01
Distances (Weighted-Average)	km	1225	121.409	62.109	6.009	345.27

Source: Own calculations based on CBS, Google Maps

Table 3: Summary statistics for MBO 4

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting flows	Persons	1225	1107	5187	0	75776
Working Population (Origin)	Persons	1225	146257	75829	46000	412000
Working Population (Destination)	Persons	1225	146257	75829	46000	412000
Distances (Equally-Weighted)	km	1225	122.449	61.261	7.961	345.01
Distances (Weighted-Average)	km	1225	121.409	62.109	6.009	345.27

Source: Own calculations based on CBS, Google Maps

Table 4: Summary statistics for HBO

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting flows	Persons	1225	1504	6789	0	122888
Working Population (Origin)	Persons	1225	110800	77242	26000	407000
Working Population (Destination)	Persons	1225	110800	77242	26000	407000
Distances (Equally-Weighted)	km	1225	122.449	61.261	7.961	345.01
Distances (Weighted-Average)	km	1225	121.409	62.109	6.009	345.27

Source: Own calculations based on CBS, Google Maps

Table 5: Summary statistics for WO

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting flows	Persons	1225	948	4673	0	121333
Working Population (Origin)	Persons	1225	110800	77242	26000	407000
Working Population (Destination)	Persons	1225	110800	77242	26000	407000
Distances (Equally-Weighted)	km	1225	122.449	61.261	7.961	345.01
Distances (Weighted-Average)	km	1225	121.409	62.109	6.009	345.27

Source: Own calculations based on CBS, Google Maps

Table 6: Summary statistics for Total Middle Education (MBO 2/3 and MBO 4)

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting Flows	Persons	1225	2163	10388	0	155903
Working Population (Origin)	Persons	1225	146257	75829	46000	412000
Working Population (Destination)	Persons	1225	146257	75829	46000	412000
Distances (Equally-Weighted)	km	1225	122.449	61.261	7.961	345.01
Distances (Weighted-Average)	km	1225	121.409	62.109	6.009	345.27

Source: Own calculations based on CBS, Google Maps

Table 7: Summary statistics for Total Higher Education (HBO and WO)

Variable	Units	Obs	Mean	Std. Dev.	Min	Max
Commuting Flows	Persons	1225	2452	11207	0	244221
Working Population (Origin)	Persons	1225	110800	77242	26000	407000
Working Population (Destination)	Persons	1225	110800	77242	26000	407000
Distances (Equally-Weighted)	km	1225	122.449	61.261	7.961	345.01
Distances (Weighted-Average)	km	1225	121.409	62.109	6.009	345.27

Source: Own calculations based on CBS, Google Maps

5 Analysis

Now that we have underlined the various models we will be using and we have clarified how our data was gathered, we can start our analysis of the various models. We will report the results for the aggregated data (aggregated over education levels) for all models, and will look into detail at the results by education level for the PPML model. All the other results are given in the Appendix. The results of these models gives us an indication of how large the γ_2 and γ_3 in equation 10 need to be to estimate the adjusted ITA in our regional labour market forecasts.

5.1 Results for aggregated data

Table 8: Zero-Inflated Negative Binomial Regression for all graduates (Aggregated)

ZINB	Commuting flows	Inflate
Working Population	2.79e-06***	-1.91e-05***
(Origin)	(2.68e-07)	(3.31e-06)
Working Population	4.19e-06***	-1.34e-05***
(Destination)	(2.43e-07)	(2.55e-06)
Distance	-0.0197***	0.0226***
(Equally-weighted with zeroes)	(0.000330)	(0.00334)
Constant	7.62***	-0.373
	(0.0929)	(0.676)
Observations	1,225	1,225

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 8 gives the results of the gravity analysis for the ZINB.[5] The way in which the ZINB regression is carried out reveals some complications for our analysis. The ZINB regression is carried out using maximum likelihood estimation (MLE). Initially, the log likelihood is maximised for the logistic model which predicts whether a worker will commute or not. This is an iterative procedure, starting with a model with no predictors. As iterations are made, variables are included until the difference in the log-likelihood between iterations is below a certain threshold, after which the second stage of the regression is carried out.

[5] An α coefficient is also produced when carrying out this regression which represents the amount of overdispersion. In general, if the α is significantly different from zero, we can be more confident in our choice of using a ZINB over a ZIP model, and we find an α significantly different from zero in all cases. Therefore, we omit the reporting of the alphas in our zero-inflated negative binomial regression result tables.

In the second stage, the full model is attempted to be maximised according to its log-likelihood function once more.

The first step is the one that proves troublesome, however, as the purpose of this investigation is to obtain elasticities of the population of the origin and destinations. Therefore, to obtain such information we would need to take the logarithm of our dependent variable, which would then result in all zeroes being omitted from the model. This would therefore disregard a crucial fraction of our data, and a ZINB model would then no longer be justified. Therefore, we cannot extract elasticities but we can still explore the interaction between our independent variables and the dependent variable.

To this end, we need to first transform all coefficients found into interpretable coefficients, and we do this by taking e^{coef}. In this manner, we re-transform these coefficients from their logged values to absolute values. As an example, for the logarithm of the working population of the origin in the total model, we have a coefficient of 2.79×10^{-6}. To transform this, we simply take $e^{2.79 \times 10^{-6}} = 1.000$. We summarise the transformations in the following table:

Table 9: Transformed Coefficients of ZINB regression

	Zero-Inflation Model	Count Model
Constant	0.689	2,039
Pop (Origin)	1.000	1.000
Pop (Destination)	1.000	1.000
Distance	1.023	0.980

From here, we can interpret theses results as follows. For the "Zero-Inflation Model" column of table 9 represents modeling zero-flows, and yields the probability of whether or not commuting flows exist between regions. The constant represents that the base chance of a flow between regions being 0 is 68.9%. If a region's working population would increase by one person, the odds that flows between regions would be zero would multiplicatively increase by 1.0000. In other words, the higher the working population in an origin region, the chance that flows between this origin region and other regions would be zero does not change. The same is true for the population of a destination region. If the population of a destination region increases by 1, the odds that flows to that region would be zero is increased multiplicatively by 1.0000. Lastly, if the distance between two regions increases by one kilometre, we have that the odds of zero flows would decrease by 1.023, or in other words 2.3%, which does not make intuitive sense. This model seems to suggest that an increase in the distance between two regions seems to increase the probability of non-zero flows.

In the second column, we are given information about the actual count of the flows between regions. We start once more with the constant, which now tells us the base commuting flow between regions is 2,039. We then interpret the coefficient in front of the population of the origin as if the origin region were to see an increase in its working population by one person, the expected flows between that region and all others would decrease by 1, ceteris paribus.

This means that the higher the working population of a region, the less flows there would be. We also have that an increase in the working population of the destination region by one would result in the expected flows into that region to decrease by a factor of 1, ceteris paribus. Lastly, we have that an increase in the distance between two regions would increase the expected commuting flow between them by 0.980. These results seem to not make intuitive sense with what we we would expect, and we also cannot yield elasticities from these models so we disregard this model for our further analysis.

Next we look at the results for the spatial econometric specification (LeSage and Pace, 2008)[6] which are given in Tabel 10. Once again, we are looking to extract information about elasticities, and so we look only at logged independent and dependent variables. We decide to use the model most closely resembling the gravity model, therefore including only the distance and the size of each region as independent variables, while also including the network effects as detailed by LeSage and Pace (2008). We note that since we took the logarithm of distances between regions, and we set the distance from region A to region A to 0, these 35 observations drop out. Therefore, as a result of including the logarithm of distance we expect to have $35^2 - 35 = 1,190$ observations. Furthermore, we take the logarithm of our dependent variable y was well, and so we have that even more observations drop out, as $log(y)$ is undefined wherever we might have had zero flows. Therefore, we are left with 1,141 observations.

We find a negative coefficient for distance, which seems to make intuitive sense as this suggests that an increase in the distance between two regions decreases the commuting flows between the regions. More specifically, we can say that a 1% increase in the distance between two regions results in a decrease in the flows by 1.818%. We can also conclude that according to the functional form of LeSage, an increase in the working population of the origin by 1% will lead to an increase in the commuting flows out of the origin of 1.016%. Similarly, we can say that an increase in the working population in the destination region by 1% would result in an increase in commuting flows into that region by 1.193%.[7]

A major advantage of this functional form is that we can extract elasticities of key components of the gravitational model and their effects on the commuting flows between regions. However, one major drawback of LeSage's OLS is that we lose the observations where commuting flows between regions are zero, since we take the logarithm of our dependent variable in order to extract elasticities. Therefore, we once again disregard this model, as it does not sufficiently model the complete relation between the regressors chosen and the commuting flows between regions.

[6] We estimate the following equation:

$$y = \rho_d \boldsymbol{W}_d y + \rho_o \boldsymbol{W}_o y + \rho_w \boldsymbol{W}_w y + \alpha \iota_N + \boldsymbol{X}_d \beta_d + \boldsymbol{X}_o \beta_o + \gamma g + \varepsilon \tag{15}$$

[7] In the Appendix the results by education level are given to see key differences between graduates of varying levels of education.

Table 10: LeSage's OLS Regression for all graduates (aggregated)

LeSage's OLS	Log of Commuting flows
W_o	1.42e-05***
	(2.07e-06)
W_d	8.90e-06***
	(2.14e-06)
W_w	5.09e-06*
	(2.89e-06)
Log of Working Population (Origin)	1.016***
	(0.0473)
Log of Working Population (Destination)	1.193***
	(0.0478)
Log of Distance (Equally-weighted with zeroes)	-1.818***
	(0.0611)
Constant	-12.580***
	(0.826)
Observations	1,141
R^2	0.755

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 11: PPML Regression for all graduates (Aggregated)

Log of Gravity	Commuting flows
Working Population (Origin)	0.703***
	(0.0513)
Working Population (Destination)	1.185***
	(0.0581)
Log of Distance	1.976***
	(0.0584)
Constant	-7.319***
	(1.037)
Observations	1,190
R^2	0.728

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Finally, we turn to the PPML results in Table 11. As a result of using the logged values of the independent variables, we once again have that 35 observations drop out, leaving us with a total of $1,190$ observations. Once again, we find the same signs of the coefficients and all variables are statistically significant. With the log of gravity model, we can immediately interpret the coefficients as the elasticity of the commuting flows in response to a change in each of the factors, and so we see for example that as distance increases, the commuting flows decrease. More specifically, according to this functional form if the distance between two regions increases by 1%, we have that the commuting flows between them should decrease by 1.976%. We also have that a 1% increase in the working population in an origin region will increase the flows between itself and other regions by 0.703%. An increase in the working population in a destination region of 1% will increase the flows between the region and other regions by 1.185%.

With the functional form proposed by Silva and Tenreyro, we can include all zero-flow observations of the dependent variable, and we can extract information about the elasticities of the independent variables. Therefore, this fulfills all the criteria for what we wanted to include in our model and what we wanted to extract from it, and therefore we settle on this model for discussing the ITA adjustment.

5.2 Education Level Comparison

From the previous section it can be concluded that distance has a negative relationship with the size of commuter flows between regions, and that both the working population in the origin and destination region have a positive relationship with the size of commuter flows between regions. In this section we look at the differences between the results by education level and give some explanations for the found results based on the literature. We distinguish 4 levels of education, vocational education (MBO) levels 2/3 and 4, Bachelor from universities of applied sciences (HBO), and Master from research universities (WO). Because the PPML model satisfies the properties of our model best, we only focus on the results of this estimation. The Appendix shows the results for the other model specifications. The results for the PPML regression by education level is given in Table 12.

Remember that despite the fact that we do not take the logarithm of the dependent variable, we can still interpret these coefficients as elasticities. We have 1,190 observations, missing only the intraregional flows. All variables are statistically significant at the 1% significance level and have the expected sign. If we first focus on the effect of distance, we see that for MBO2/3 (vocational education level 2/3) graduates, an increase in distance between two regions by 1% decreases flows into that region by 2.163%. The coefficients on distance decrease with the increase of education level. This means that the higher educated seem to be less sensitive for distance than low educated. This is in line with research on this topic in the Netherlands, for example, by de Groot (2015) and Groot et al. (2012). Higher educated workers commute much longer than

Table 12: Summary Results PPML Regression split by education

PPML	MBO 2/3	MBO 4	HBO	WO
Log of Working Population (Origin)	0.606***	0.619***	0.525***	0.927***
	(0.0742)	(0.0753)	(0.0495)	(0.0334)
Log of Working Population (Destination)	0.954***	1.153***	0.989***	1.360***
	(0.0800)	(0.0862)	(0.0570)	(0.0414)
Log of Distance (Equally-weighted with zeroes)	-2.163***	-2.078***	-1.932***	-1.677***
	(0.0720)	(0.0731)	(0.0639)	(0.0513)
Constant	-3.310**	-6.006***	-2.622**	-13.031***
	(1.1387)	(1.478)	(0.910)	(0.690)
Observations	1,190	1,190	1,190	1,190
R^2	0.558	0.542	0.676	0.868

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

lower (or middle) educated workers. Explanations for this pattern is that higher educated workers are more specialised and they are then willing to travel further away for a job that matched this profile. Another argument can be that higher educated workers earn a higher wage in a well matched job, which makes it worth while to travel further to a job. However, Groot et al. (2012) show that even when controlling for wages the results hold, i.e., given a certain wage, the higher educated are more likely to commute longer than middle or lower educated.

The commuting results depict a relationship between region of work and region of residence, as both the choice of where to work and where to live impacts commuting distance. The found coefficients for working population in the origin and destination also show that there are differences between higher and lower educated workers in the location of work and residence. The size of the working population in the origin has a stronger push-factor and the size of the working population in the destination region has a stronger pull-factor for higher educated workers. This can be explained by the sorting of workers of different education levels over residential areas and working areas, as is pointed out by Groot et al. (2012). If lower educated works commute between regions, they tend to commute from a residential region (suburbs) to the closest centre of employment, i.e., the closest city. Higher educated workers tend to commute to dense employment areas, i.e., not the closest but the largest cities. This shows in our results, as the size of the commuter flow of WO-educated workers increases by 1.360% if the size of the working population in the destination region increases by 1%. This effect is larger than for the other education levels. The same pattern is observed for the working population in the origin region, although the coefficients are not as large. One explanation for the pattern for the origin region is that higher educated workers tend to live in larger cities more often than lower and middle educated workers (if we do not considered

social housing).

Table 13: Summary Results PPML Regression split by categories

PPML	Middle Education	Higher Education	Aggregated	Stacked
Log of Working Population	0.613***	0.704***	0.703***	0.652***
(Origin)	(0.0727)	(0.0403)	(0.0513)	(0.0667)
Log of Working Population	1.063***	1.151***	1.185***	1.134***
(Destination)	(0.0816)	(0.0468)	(0.0581)	(0.0665)
Log of Distance	-2.117***	-1.820***	-1.976***	-2.236***
(Equally-weighted with zeroes)	(0.0707)	(0.0555)	(0.0584)	(0.0835)
Education × log of Working Population				0.001
(Origin)				(0.0196)
Education × log of Working Population				0.007
(Destination)				(0.0181)
Education × log of Distance				0.105***
(Equally-weighted with zeroes)				(0.0290)
Constant	-4.0875***	-6.511***	-7.319***	-6.022***
	(1.410)	(0.776)	(1.037)	(0.605)
Observations	1,190	1,190	1,190	4,760
R^2	0.561	0.791	0.728	0.658

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

In table 13 we give an overview of the results for the PPML estimation, but for different aggregated categories of education. We aggregate the educational levels over middle education, i.e., vocational education, and higher education, i.e., bachelor en master. The final two columns in the table show the results for all educational levels aggregated and pooled (stacked). We broadly get the same results as for the educational levels separately, however the coefficients are more averaged out. This shows thus immediately, that with aggregation of the commuting flows by education, the results underestimate the effect of distance for lower educated workers and overestimates the effect of distance for higher educated workers. In addition, this also holds for the working population in the origin and destination. An advantage of these aggregated categories, is that the data on the working population is available for exactly these categories and we don't have to make assumptions on the working population divided by WO and HBO, for example.[8]

[8]Which we assumed to be the same in the previous estimations

6 Conclusion and Implications for the ROA regional forecasts

Having explored different functional forms in this research, we have come to the conclusion that the correct way to estimate the coefficients γ_2 and γ_3 from equation 5 is by using the 'log of gravity' functional form from Silva and Tenreyro. As described in the previous section, the findings from this functional form are consistent with what we expect, and what we have seen in other studies (e.g. de Groot (2015)). As in the previous estimations of the regional ITA the size of the coefficients in equation 5 were based on these other studies, they are fairly consistent with what we have found in this study. Using the 'log of gravity' also allows to directly obtain the coefficients as elasticities, which is what we need to easily estimate the re-adjustment of the ITA. While this is also achieved with the OLS functional form, with OLS we lose substantially more observations and disregard the fact that there are a high number of zeros. This makes this approach highly unfit for our purpose.

Our findings from using the 'log of gravity' functional form are that there is a decrease in distance deterrence as the level of education increases. We also find that the working population in the destination region seems to have a higher influence as the level of education rises, while the working population in the region of origin seems to have a lower influence. More specifically, we find the following gammas from Tables 12 and 13:

Table 14: Gamma findings from 'log of gravity' for POA report

	MBO 2/3	MBO 4	HBO	WO
γ_2	0.606	0.619	0.525	0.927
γ_3	0.954	1.153	0.989	1.360

Table 15: Gamma findings from 'log of gravity' for POA report (continued)

	Middle Education	Higher Education	Aggregated	Stacked
γ_2	0.613	0.704	0.703	0.652
γ_3	1.063	1.151	1.185	1.134

The initial gammas used in the latest POA report from 2017 were $\gamma_2 = 0.8$ and $\gamma_3 = 1.2$ and no difference in the gamma was made between education levels. These used values are in line with what we find in this research, however, the values are rather high for workers with a vocational education and a bit too low for workers with an university education. This would mean that we over-adjust the ITA for lower educated workers and under-adjust the ITA for workers with a university education. From this research we can conclude that implementing a heterogeneous estimation for γ_2 and γ_3 will more accurately reflect the differences in commuter behaviour across educational levels, and improve the commuter-adjusted ITA. Within the current estimation model for the

regional forecasts, this adjustment could be fairly easily made. As the PPML model is estimated for past commuter flows, the estimation of the parameter values should be repeated with new data to check whether commuting patterns have not changed. Although commuting patterns are rather stable over time, changes might occur depending on economic circumstances but also depending on the introduction of faster modes of transportation or faster routes are opened (for example, a faster rail way connection). This suggests that the coefficients need to be evaluated about every 5 years or so.

We conclude this research with a final remark on the commuter-adjusted ITA. With the estimated parameters, significant adjustments to the inflow as calculated in equation 10 (or equation 4.9 in the technical report) are made, and we do see that a substantial amount of workers could potentially commute between two regions if the ITA between these two regions differs. In this sense, the regression results of the gravity equation provide us with a good estimation of the adjustments in commuting flows needed and the current correction that is applied succeeds in 'correcting' ITA's for commuting possibilities.

However, the inflow of equation 10 is only a small component of the supply side of the overall ITA, as is shown in equation 1. Even if a substantial adjustment in the inflow in equation 10 is calculated, the overall adjustment of the ITA is still relatively small most of the times. This means that if we take changes in commuting into account if ITA's between regions differ substantially, the commuter-adjusted ITA will at most differ only one ITA-qualification from the non-adjusted ITA. For the ITA to switch qualifications from one level to a next (up or down) fairly large changes in demand or supply components need to be calculated, which is harder to achieve if the commuter-corrected ITA only focuses on correcting the inflow component, as is done in the current model (both based on the model and the data available). Given the data availability (for example, as we only observe the residential location in the labour force survey, the regional labour market forecasts are estimated using the region of residence of workers) and model-setup, this is the only logic and correct way of calculating the adjusted-ITA.

References

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics*, 3(1):133–160.

Bakens, J., Cörvers, F., Dijksman, S., Fouarge, D., and Poulissen, D. (2018). Methodiek regionale arbeidsmarktprognoses 2017-2022. *Maastricht University Research Outputs*.

CBS (2018). Cbs statline. `https://opendata.cbs.nl/statline/#/CBS/nl/navigatieScherm/thema`.

de Groot, H. (2015). *Arbeids- en Woningmarktdynamiek*. Platform31.

Desmarais, B. A., Harden, J. J., et al. (2013). Testing for zero inflation in count models: Bias correction for the vuong test. *The Stata Journal*, 13(4):810–835.

Greene, W. H. (2012). *Econometric analysis (International edition)*. Pearson US Imports & PHIPEs.

Groot, S. P., de Groot, H. L., and Veneri, P. (2012). The educational bias in commuting patterns: Micro-evidence for the netherlands. *Tinbergen Institute Discussion Paper*, 12(080/3).

LeSage, J. P. and Pace, R. K. (2008). Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, 48(5):941–967.

LeSage, J. P. and Thomas-Agnan, C. (2015). Interpreting spatial econometric origin-destination flow models. *Journal of Regional Science*, 55(2):188–208.

Silva, J. M. S., Tenreyro, S., and Windmeijer, F. (2015). Testing competing models for non-negative data with many zeros. *Journal of Econometric Methods*, 4(1):29–46.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4):641–658.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333.

Wilson, P. (2015). The misuse of the vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127:51–53.

Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.

Appendix

In addition to the Log of Gravity models, we have estimated count data models and a spatial econometric model to test the most suitable specification of the commuter flows based on gravity models. Here we briefly discuss these models and present the results of these alternative models.

Description of Alternative models

Count Data Models

If we regard our model as counting the number of people commuting from one region to another, we see that we fall into a discrete class of models, and we use Winkelmann (2008) and Greene (2012) to aide our model selection.

In Greene, we are first introduced to the Poisson model as a count model, with the assumption that the conditional expectation of the dependent variable is equal to its conditional variance (equidispersion). If our data would not have this quality, we would look at alternatives such as the Negative Binomial model (NEGBIN). The NEGBIN works in a similar way to a normal multiple regression, except for the fact that the dependent variable is a discrete count following a negative binomial distribution. Greene (2012) provides a concise description and analysis and arrives at the following general NEGBIN P (NBP) class of models, where we can vary P:

$$Prob[Y = y_i|\boldsymbol{x}_i] = \frac{\Gamma(\theta\lambda_i^{2-P} + y_i)}{\Gamma(y_i + 1)\Gamma(\theta\lambda_i^{2-P})} \left(\frac{\lambda_i}{\theta\lambda_i^{2-P} + \lambda_i}\right)^{y_i} \left(\frac{\theta\lambda_i^{2-P}}{\theta\lambda_i^{2-P} + \lambda_i}\right)^{\theta\lambda_i^{2-P}}, \tag{16}$$

with $i = 0, 1, ..., N$, and where λ represents the distribution parameter of a Poisson population, so the mean rate of y per unit of time, θ represents the scale parameter with respect to the Poisson population. The CEF yields λ_i, while the conditional variance is given by:

$$Var[y_i|\boldsymbol{x}_i] = \lambda_i[1 + (1/\theta)\lambda_i^{P-1}].$$

The most common cases of the NBP model are when $P = 1$ or $P = 2$, where P is simply a classification of the model which stipulates the relationship between the variance and the mean. As an example, the NB1 model assumes a linear relationship between the variance and the mean, whereas the NB2 model assumes a quadratic relationship, as outlined in Greene (2012). In order to choose between the NB1 and NB2 models, Greene suggests using the Vuong test, Vuong (1989), for non-nested models to see which model performs statistically better.

One last class of models we consider are the Zero-Inflated models, for which we look exclusively at the Zero-Inflated Poisson (ZIP) model and the Zero-Inflated Negative Binomial (ZINB) model. These models are employed when the data contains a large number of zeroes, which is often the case when looking at flows between different geographic regions. Winkelmann (2008) states that

the Zero-Inflated Poisson (ZIP) model does not provide consistent estimation unless $\boldsymbol{x}$ and z are independent, where x is the matrix of explanatory variables, and $z = x'\gamma + \varepsilon$ is a latent indicator variable.

If $\boldsymbol{x}$ and z are not independent, we move on to the Zero-Inflated Negative Binomial (ZINB) model for which estimation is consistent. According to Desmarais et al. (2013), one can use the Vuong test "to determine whether the zero-inflated model fits the data statistically significantly better than count regression with a single equation". However, according to recent research conducted by Wilson (2015) and Silva et al. (2015), the Vuong test is no longer valid as its limiting distribution is no longer considered to be standard normal but is instead unknown, meaning the Vuong test cannot be used for inference.

Therefore, in lieu of using the invalid Vuong test, we can approach the problem more practically based on our data. For now, we ignore which NBP model is statistically best, and instead consider the general class of negative binomial models (NB). The largest difference between the ZINB and NB is in the assumed underlying data generating processes (DGP's), namely that the ZINB assumes two different DGP's, while the NB assumes only one. The ZINB contains one DGP which acts as a classification, dictating whether an individual participates in an event or not. The second DGP then stipulates to what extent the individual takes part in the event. As outlined in Winkelmann (2008), the ZINB can be split into a logit specification for the extra zeroes, and we can use for example a NB2 specification for the count data. In that case, we combine two distinct DGP's to account for the extra zeroes in our model that we do not capture when we look at the NB2 model in isolation. The first data generating process would be modeled as follows:

$$\boldsymbol{\pi} = \frac{e^{Z\beta}}{1 + e^{Z\beta}}, \tag{17}$$

which is a logistic regression model where $\boldsymbol{\pi}$ denotes a vector of probabilities, Z a matrix of explanatory variables, and β a vector of regression coefficients. In this manner, we use the logistic regression model to predict whether flow occurs between two regions. The second DGP is given by:

$$\boldsymbol{y} = exp(X\beta), \tag{18}$$

which is a simple negative binomial regression. We have $\boldsymbol{y}$ as the observed outcome, in this case the flows between regions. X denotes the matrix of explanatory variables, and β the vector of regression coefficients. Using this regression equation, we model the flows which were predicted to have occurred between regions.

We commence our analysis by looking at arguably the simplest model, namely the Negative Binomial model (NBP). We limit our consideration to the two most commonly used cases, namely $P = 1$ and $P = 2$. The NBP is mostly used in the case of over-dispersed discrete data, which is when the conditional variance is greater than the conditional mean. We see evidence for this in table 16. Using this model, we adopt the same mean structure as a Poisson regression,

and we include an extra parameter to model the over-dispersion, namely θ in equation 16. In the case where there is over-dispersion, we typically find that the confidence bands for a NBP are tighter than those generated from a Poisson regression.

We are interested in the logged transformation of our variables, as we want to extract the nature of the elasticities of the variables and their influence on the number of commuters, and therefore we take the logarithms of all the variables considered. Firstly, to decide between an NB1 or NB2 model, we apply qualitative techniques. We start by looking at the mean and variance of our commuting, broken down in categories based on the ITA of the destination.

Table 16: Mean and Variance of Commuting flows for each different ITA for MBO 2/3 graduates

ITA (Destination)	mean	variance	N
0.87	2722.62	7.29e+07	35
0.90	494.15	1.77e+06	35
1.01	1015.49	1.57e+07	140
1.02	1436.28	4.25e+07	35
1.03	716.91	1.01e+07	140
1.04	1477.23	6.37e+07	140
1.05	901.01	8.90e+06	35
1.06	1386.96	3.81e+07	105
1.07	809.40	2.31e+07	70
1.08	579.06	7.74e+06	70
1.09	979.66	2.28e+07	140
1.10	927.34	2.54e+07	35
1.11	1514.28	6.02e+07	35
1.12	757.40	1.13e+07	140
1.13	1117.40	3.84e+07	70
Total	1055.86	2.72e+07	1225

Here, we find significant support for the use of a Negative Binomial model as the mean commuting flows seem to vary based on the ITA. We can further argue for the use of an NB2 model, since the conditional variance exceeds the conditional mean at every ITA found in the data, meaning that the data is over-dispersed. However, due to the nature of the data there is a concern that there are two underlying data generating processes (DGP's). We have one DGP which is a classification of whether people will commute between regions or not. The second DGP then reflects where the commuter will commute to, if they choose to commute. We can see evidence of this by looking at the histogram of the discrete frequency of the commuting flows:

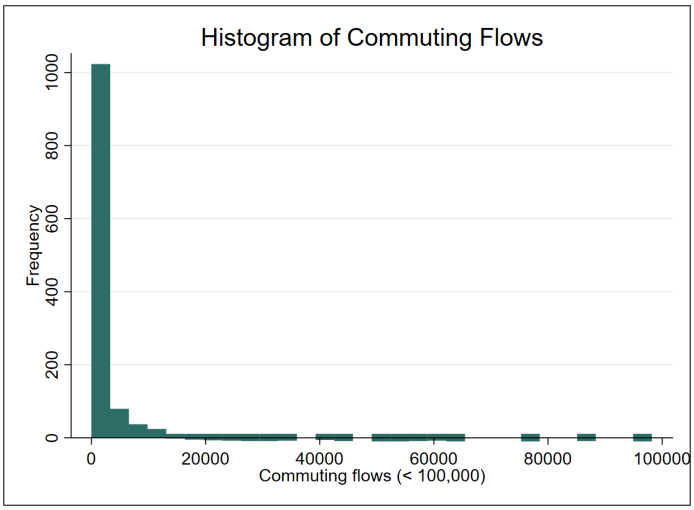

Figure 2: Histogram of Commuting flows for all educational levels

Note that we removed any occurrences of commuting flows that were greater than 100,000. Looking at the histogram, we can see a very high frequency of zero flows, indicated by the tall bar on the left. We see higher flows occurring less frequently, as evidenced by the lower heights of the bars. This seems to indicate that there are indeed two different DGP's generating our data. Because our zeroes seem to be at least partially generated by a different DGP than our other flows, this seems to suggest that the model would be misspecified as a Negative Binomial Model. Therefore, we turn our attention to a variation of the class of NBP models, which takes into account that there are two separate DGP's at play within the data collected. This class of models is called Zero-Inflated Negative Binomial Models (ZINB).

If we continue onwards from the previous analysis, we need to keep in mind the following two characteristics of our dataset: firstly, we remember that the conditional variance of the response variable is greater than its conditional mean. This is reason enough for us to forgo considering a Zero-Inflated Poisson Model (ZIP), since a ZIP model performs better only when data is not over-dispersed, which is not the case here. Secondly, the number of zeroes is excessive, as demonstrated by figure 2. This is supposed to be the case as a result of two distinct DGP's for the commuting flows, and so two different zero-flow responses may exist.

Spatial Econometric Models

Instead of looking at count models, we could explore literature more directly related to the context of our question at hand, namely that of spatial economics and econometrics. The study of spatial econometrics is an active field of

research, however we will mainly be using the findings of LeSage and Thomas-Agnan (2015) to construct a spatial econometric model. Here, we attempt to model the spatial dependence directly through adding spatial-connectivity matrices, and manipulating them in a particular manner. This is in contrast to what we look at before, where we simply looked at the spatial dependency through the distance between regions. Before we introduce the complete model, we walk through the various components of the model. We start with the $n \times n$ flow matrix Y, which represents the labour flows from each labour market region to the other. We vectorize Y by stacking the columns of Y, creating the $N \times 1$ vector y, where $N = n^2$. We then introduce the spatial connectivity matrix $\boldsymbol{W}$, which is a row-stochastic matrix of neighbours to each region. However, other definitions of spatial connectivity can be used as well.

Next, we have ι_N, which is an $N \times 1$ vector consisting of 1's, representing our constant. We introduce our matrix $\boldsymbol{X}$, which is our $n \times k$ matrix of k explanatory variables. In this case, we decide to use the working population as an indication of the size of each region in the X matrix, in order to properly emulate the gravitational model. Lastly, we introduce the matrix G as a $n \times n$ matrix of the distances between regions. We then also stack the columns of matrix G sequentially, resulting in the $N \times 1$ vector g. Having defined all our variables, we will be using LeSage and Pace (2008) as reference material for our analysis. The general model they outline is given as follows:

$$y = \rho_d \boldsymbol{W}_d y + \rho_o \boldsymbol{W}_o y + \rho_w \boldsymbol{W}_w y + \alpha \iota_N + \boldsymbol{X}_d \beta_d + \boldsymbol{X}_o \beta_o + \gamma g + \varepsilon \tag{19}$$

We define $\boldsymbol{W}_d$ as $I_n \otimes \boldsymbol{W}$, $\boldsymbol{W}_o$ as $\boldsymbol{W} \otimes I_n$, and $\boldsymbol{W}_w$ as $\boldsymbol{W} \otimes \boldsymbol{W}$, where $\otimes$ represents the Kronecker product. Then, the product of the $\boldsymbol{W}$ matrices with y result in the average flows from the neighbours of the origin, destination or neighbouring regions to all regions for the product of y with $\boldsymbol{W}_o$, $\boldsymbol{W}_d$, and $\boldsymbol{W}_w$ respectively. Similarly, $\boldsymbol{X}_d$ outlines the characteristics of the destination region, and $\boldsymbol{X}_o$ the characteristics of the origin region. We have that ι_N represents a vector of 1's, meaning that α represents the constant in the model. Our ε is a standard OLS error term.

The spatial connectivity matrix $\boldsymbol{W}$, is produced based on figure 1, and we look at the rook contiguity of each region in the matrix $\boldsymbol{W}$. Rook contiguity means that for each combination of two regions, we have a 1 if the two share a direct border, and 0 otherwise. We then ensure that the spatial connectivity matrix is row-stochastic by giving equal weights to each non-zero entry in the row such that the non-zero entries sum up to 1. Therefore, weights are lower if the number of neighbours increases, and the rationale for this is that the individual influence of one neighbouring region is diminished as the number of neighbouring regions increases. The matrix of explanatory variables, $\boldsymbol{X}$, includes the population and median income for each region.

Regression Results for Alternative Models

Zero-Inflated Negative Binomial Model

Table 17: Summary Results ZINB Regression split by education

ZINB	MBO 2/3	MBO 4	HBO	WO
Working Population (Origin)	3.18e-06***	3.20e-06***	3.80e-06***	5.07e-06***
	(6.13e-07)	(6.17e-07)	(5.10e-07)	(4.69e-07)
Working Population (Destination)	5.72e-06***	5.86e-06***	6.52e-06***	8.47e-06***
	(6.08e-07)	(5.79e-07)	(4.44e-07)	(4.39e-07)
Distance (Equally-weighted with zeroes)	-0.0195***	-0.0187***	-0.0195***	-0.0156***
	(0.000402)	(0.000402)	(0.000357)	(0.000358)
Constant	6.709***	6.818***	7.251***	6.150***
	(0.125)	(0.122)	(0.0865)	(0.0847)
Observations	1,225	1,225	1,225	1,225

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 18: Summary Results ZINB Regression split by categories

ZINB (Count)	Middle Education	Higher Education	Aggregated	Stacked
Working Population (Origin)	4.00e-06***	5.22e-06***	2.79e-06***	2.76e-06***
	(6.01e-07)	(5.00e-07)	(2.68e-07)	(5.80e-07)
Working Population (Destination)	6.82e-06***	8.27e-06***	4.19e-06***	4.73e-06***
	(5.81e-07)	(4.50e-07)	(2.43e-07)	(5.82e-07)
Distance (Equally-weighted with zeroes)	-0.0198***	-0.0182***	-0.0197***	-0.0205***
	(0.000380)	(0.000335)	(0.000330)	(0.000464)
Education × Working Population (Origin)				3.18e-07
				(2.05e-07)
Education × Working Population (Destination)				6.76e-06***
				(2.04e-07)
Education × Distance (Equally-weighted with zeroes)				8.10e-04***
				(1.74e-04)
Constant	7.1125***	7.238***	7.624***	6.833***
	(0.118)	(0.826)	(0.0929)	(0.510)
Observations	1,225	1,225	1,225	4,900

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

LeSage's OLS

Table 19: Summary Results LeSage's OLS Regression split by education

LeSage's OLS	MBO 2/3	MBO 4	HBO	WO
W_o	7.62e-05***	8.00e-05***	5.65e-05***	3.94e-05***
	(9.20e-06)	(9.01e-06)	(6.18e-06)	(9.22e-06)
W_d	7.01e-05***	6.45e-05***	4.37e-05***	3.03e-05***
	(9.24e-06)	(9.11e-06)	(6.46e-06)	(1.01e-05)
W_w	1.90e-05	1.08e-05	-3.98e-06	-4.86e-05***
	(1.23.e-05)	(1.38e-05)	(9.04e-06)	(1.54e-06)
Log of Working Population (Origin)	0.503***	0.492***	0.619***	0.850***
	(0.0616)	(0.0616)	(0.0433)	(0.0491)
Log of Working Population (Destination)	0.708***	0.782***	0.808***	1.135***
	(0.0609)	(0.0611)	(0.0434)	(0.0484)
Log of Distance (Equally-weighted with zeroes)	-1.337***	-1.189***	-1.506***	-1.609***
	(0.0706)	(0.0703)	(0.0631)	(0.0685)
Constant	-3.496***	-4.624***	-4.085***	-10.060***
	(0.997)	(1.002)	(0.731)	(0.808)
Observations	936	858	990	897
R^2	0.625	0.630	0.722	0.703

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 20: Summary Results LeSage's OLS Regression split by categories

LeSage's OLS	Middle Education	Higher Education	Aggregated	Stacked
W_o	3.92e-05***	2.54e-05***	1.42e-05***	7.45e-05***
	(4.81e-06)	(3.78e-06)	(2.07e-06)	(4.19e-06)
W_d	3.27e-05***	1.79e-05***	8.90e-06***	6.32e-05***
	(4.86e-06)	(4.03e-06)	(2.14e-06)	(4.34e-06)
W_w	6.29e-06	1.21e-06	5.09e-06*	-1.33e-06**
	(6.33e-06)	(5.79e-06)	(2.89e-06)	(5.95e-06)
Log of Working Population (Origin)	0.728***	0.848***	1.016***	0.352***
	(0.0611)	(0.0429)	(0.0473)	(0.0483)
Log of Working Population (Destination)	0.942***	1.054***	1.193***	0.542***
	(0.0605)	(0.0432)	(0.0478)	(0.0479)
Log of Distance (Equally-weighted with zeroes)	-1.514***	-1.673***	-1.818***	-1.332***
	(0.0697)	(0.0634)	(0.0611)	(0.0692)
Education × log of Working Population (Origin)				0.006
				(0.0163)
Education × log of Working Population (Destination)				0.0262
				(0.0162)
Education × log of Distance (Equally-weighted with zeroes)				-0.125
				(0.0236)
Constant	-7.435***	-8.218***	-12.580***	
	(0.991)	(0.720)	(0.826)	
Observations	1,046	1,066	1,141	3,681
R^2	0.645	0.747	0.755	0.977

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1