Modelling the Influence of Regional Identity on Human Migration

Vermeulen, W.R.J.; Roy, D.; Quax, R.

DOI
10.3390/urbansci3030078

Publication date
2019

Document Version
Final published version

Published in
Urban Science

urban science

MDPI

Article

Modelling the Influence of Regional Identity on Human Migration

Willem R. J. Vermeulen *, Debraj Roy and Rick Quax

Faculty of Science, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
* Correspondence: w.r.j.vermeulen@gmail.com

Received: 10 June 2019; Accepted: 18 July 2019; Published: 26 July 2019

Abstract: Human migration involves the relocation of individuals, households or moving groups between geographical locations. Aggregate spatial patterns of movement reflect complex interactions among motivations (such as distance, identity, economic opportunities, etc.) that influence migration behaviour and determine destination choice. Gravity models and radiation models are often used to study different types of migration at various spatial scales. In this paper, we propose that human migration models can be improved by embedding regional identities into the model. We modify the existing human migration gravity model by adding an identity parameter based on three different sets of Dutch identity regions. Through analysis of the Dutch internal migration data between 1996 and 2016, we show that adding the identity parameter has a significant effect on the distance distribution. We find that individuals are more likely to move towards municipalities located within the same identity region. We test the impact of regional identity by comparing randomly spatially clustered and optimised identity regions to show that the effects we attribute to regional identity could not be attributed due to chance. Finally, our finding shows that cultural identity should be taken into account and has broad implications on the practice of modelling human migration patterns at large. We find that people living in Dutch municipalities are 3.89 times as likely to move to a municipality when it is located within the same historic identity region. Including these identity regions in the migration model decreases the deviation of the model by 10.7%.

Keywords: human migration; regional identity; gravity models; radiation models

1. Introduction

Migration of human beings has always been an important dynamics of our history and is an important, multidimensional and complex issue [1]. Historical data show that human migration may be temporary or permanent in nature, with the intention of moving to a new land, place or country. Population migration is primarily classified as internal or international, and often analysed separately. However, the underlying economic, social, political and cultural drivers initiating and perpetuating both types of migration are similar and differ predominantly in their relative importance [2]. A fundamental difference, however, is the role of the state and regulations to control international migration. While international migration has gathered more attention due to its policy implications, internal migration, with an estimated total of 740 million people (12% of the world population) in 2000, is much larger than an estimated 200 million international migrants [2]. Research on internal migration is important, as it is a key aspect of the population dynamics of a country. Even though administrative structures can be hard to compare, the United Nations' Department of Economic and Social Affairs shows that five-year migration rates between the smallest administrative units in countries worldwide can vary between 4% and 20% [3]. Migration events within these regions would even further increase these migration rates.

Migration may involve individuals, family units or large groups. Sjaastad [4] beglieved that migration cannot be viewed in isolation, as human decision making is usually collective (influenced by an individuals social network) and shaped by individual behaviour. Broadly, the migration decision entails weighing the costs versus the benefits of migration. It has been discovered that the choice to move to a certain destination is likely influenced by its economic prospects [5], the availability of amenities [6,7], the travel distance [8], information distance [9] and social distance [10] between both locations. The way in which these variables influence the migration decision differs for each individual. After all, each individual has different personal connections [11,12], motivations [9] and aspirations [13,14].

Migration theories can be classified according to the level they focus on. Micro-level theories focus on individual migration decisions, whereas macro-level theories look at aggregate migration trends and explain these trends with macro-level explanations. Neoclassical macro-level migration theories postulate that migration is determined by the expected increase in earnings weighted by the probability on employment [15,16]. A different view is posed in the historical approach of the World Systems theory [17]. In this theory, migration is treated as a consequence of "capitalist expansion of neoclassical governments and multinationals". In the Dual Labour Market theory [18], emphasis is placed on a secondary labour market for unskilled jobs which fluctuates according to the business cycle, making it unstable and unattractive to local workers. This would in turn lead to migration of unskilled labour from poorer regions. Micro-level theories focus on the behaviour of individuals. For example, Wolpert's stress-threshold theory [19], which assumes individuals have a threshold level of satisfaction that is a function of their current state in relation to the environment. The Human capital theory [20] advances the neoclassical macro-theory by incorporating socio-demographic characteristic of individuals. Hein de Haas argued that existing theories on social capital of migration explain the growth of already-established migrant networks, but fail to explain their creation, different trajectories and subsequent collapse [21]. Similarly, there are cognitive theories such as value-expectancy [22] in which migrants make conscious decisions based on various non-economic factors: autonomy, security and self-fulfilment.

Based on the above theories, several types of mathematical and micro simulation models of migration have been created to understand and forecast migration at various temporal and spatial scales. Hoda Rahmati and Tularam [23] presented a comprehensive review of the state-of-art in modelling human migration. Each type of models uses many different variables in a different way. Economic models focus on "economic, non- economic, psychological costs, distance, income, unemployment, educational, population, urbanisation, and human capital investment variables". This can be seen in statistical time series models, which incorporate wage, unemployment and population density. Agent-based models often include other non-economic variables, such as adaptation behaviours, subjective norm, perceived behavioural control, rainfall assets, livestock asset, occupation, experience and general information about migration, age, gender, marital status, rainfall condition, and migration information and consideration. Other types of models have a more long-term focus, such as Markov chain models and optimisation models. The first make a link between the perpetuation of migration and population flows and can be either discrete or continuous, while the latter include return migration but also focus on age, asset and GDP parameters. Another type of model that is popular for migration analysis is the gravity model. The notion of applying the gravity law to population movement was introduced by John Q. Stewart who founded the "social physics" school [24]. The gravity model has been the most frequently-used paradigm for understanding migration flows between cities, countries and regions. The gravity model owes its success to, "firstly, its intuitive consistency with migration theories; secondly, ease of estimation in its simplest form; and, thirdly, goodness of fit in most applications" [25]. Recently, Ramos et. al. used a gravity model for nearly 200 countries between 1960 and 2010 to demonstrate an increase in migratory pressures from European Neighbouring Countries to the EU [26]. Gravity models can easily be augmented to include different additional controls (e.g., economic situation, transportation, etc.) and policy variables

(e.g., economic zones, housing policies, etc.). In these extended gravity models, interactions between different locations are specified as a direct function of their mutual geographical distance and their population mass as proxy to the economic prospects of a location. A new type of migration models that has also become popular are the radiation models [27], which have been applied and extended further [28–30]. In these models, residents create intervening opportunities for migrants, which means that geographical distance only has an indirect influence on the generated migration flows. Both types of models can come in different forms. A good example of this variety in models is an artificial neural network model that included both the traditional variables, as well as intervening variables and amenity variables [31]. However, the above theories and models do not account for regional identity in the analysis of migration.

Identity is a latent variable and often not included in models due to lack of data. In this paper, we overcome this problem by using both Dutch administrative regions as well as historic literature to form identity regions. These identity regions can help separate the effects of distance and identity, and present a more accurate picture of the effect of distance. As distance has a direct effect in the gravity model, we expand this model with identity regions, which in turn allows us to quantify the effect of identity on the migration from a given municipality. Having an easily accessible way of including identity into a migration model could help policymakers, as better migration predictions would allow them to improve their future policies at a low cost.

In this paper, we show that identity plays an important role in Dutch internal migration, as people are 3.89 times as likely to move to a municipality when it is located within the same historic identity region. Section 2 provides a description of the methodologies and data used, and the definition of these regional cultural identities. In Section 3, the models are developed, and parameters are estimated based on survey data. These models are then used in Section 4, where we test the impact of regional identity by comparing randomly spatially clustered and optimised identity regions to show that the effects we attribute to regional identity could not be attributed due to chance. Including these identity regions in the migration model decreases the deviation of the model by 10.7%. Finally, Section 5 concludes with a discussion of the results and a way forward.

2. Methodologies

In this section, we first specify a basic gravity model for human migration, after which this model is fitted on Dutch data collected between 1996 and 2016. Using three different sets of identity regions we specify, we introduce the identity regions to the gravity model. After creating this model, we define a way of comparing the influence of identity on migration for different identity sets, and define ways to create and optimise the definitions of such regional identities.

2.1. Used Migration Data

Part of the internal migration events within the Netherlands take place within municipalities, whereas the other events take place between two different municipalities. This means that both intra-municipal and inter-municipal migration data should be used, as we want to include all internal migration events. Such migration data are available for the years 1995–2016 via Statistics Netherlands [32].

The migration data were collected directly from the Dutch civil registration database. Even though no data are available on the percentage of unregistered movements, this shortcoming is not expected to have a large effect on the outcomes of this research: research by CBS showed that in 2016 96.26% (95% CI [95.72%, 96.81%]) of Dutch citizens were registered at the correct address [32].

For each year, the number of inter-municipal migrants between every combination of municipalities is recorded, as well as the number of intra-municipal migrants. This does not mean that data for the same number of municipalities are recorded each year: because of merges of municipalities, the number of municipalities has decreased from 625 in 1996 to 390 in 2016. To be able to create

maps and compare migration data in different years, we artificially merge municipalities to form the municipalities that existed in 2016 [32].

2.2. Model Specification

As mentioned in the Introduction, several types of models are often used to model migration. Even though other model types might have a slightly better performance, a gravity model is considered here. The decisive factor in this choice is that a gravity model explicitly uses the distance variable, whereas the radiation model uses the distances variable indirectly [33]. The influence of distance on the migration process and the impact of the introduction of identity regions on that distance variable would otherwise be hard to determine.

The most basic gravity model often used to model human migration is shown in Equation (1). Within this equation, the populations p_a and p_b of municipalities a and b are positively related to the number of people that migrate from municipality a to b, $M_{a\to b}$. When more people live in municipality a, a larger number of people can leave that municipality , and when there are more people living in municipality b, it is likely that there are more opportunities in that municipality [9]. This could make people more willing to move there.

Alongside the influences of the population sizes of both municipalities, the distance $\Delta_{a\to b}$ between those two municipalities is also included in this model. This distance can compensate for the influence of travel distance and information distance between two municipalities. As the distance between two municipalities becomes larger, it becomes less likely that individuals move between those two municipalities. The G variable is a proportionality constant that differs depending on the geographical context and time scale in which the function is applied.

$$M_{a\to b} = G \cdot \frac{p_a^{\alpha} \cdot p_b^{\beta}}{\Delta_{a\to b}^{\delta}} \tag{1}$$

This equation can then be rewritten in a linear form through taking the logarithm of both sides, as shown in Equation (2). By doing this, a generalised linear regression (GLM) can be applied to find the values of α, β, γ and δ. The γ variable is introduced to accurately determine the value of G, which equals $\exp(\gamma)$.

$$\ln(M_{a\to b}) = \alpha \cdot \ln(p_a) + \beta \cdot \ln(p_b) - \delta \cdot \ln(\Delta_{a\to b}) + \gamma \tag{2}$$

2.3. Fitting a Standard Gravity Model for Human Migration

To be able to fit the gravity model to the migration data, we also need to have data on the population size of all municipalities and data on the distances between all municipalities. While municipal population data can be acquired through Statistics Netherlands, it is harder to acquire data on the distance travelled by each individual [32].

The distance travelled by every migrant is approximated by the length of the straight line between the geographical centres of both municipalities in kilometres. This can be done because this distance is highly correlated with the travel time between two locations [34]. Even though there are cases where this assumption does not hold, such as when certain geographical boundaries cannot be crossed or the population centre is located far from the geographical centre of a municipality, we assume this does not have a significant impact. This is an assumption that works for the Netherlands, because it is flat and there are many roads and bridges—except perhaps for the two municipalities that are located at opposite shores of the IJsselmeer. When travel times do not necessarily correlate with the geographical distance, it might be better to find the actual travel times between two municipalities instead—taking into account the location of the population centres within those municipalities as well.

This way of approximating the distance travelled cannot be used for the intra-municipal migration data. The distance between the centre of a municipality and the centre of that very same municipality is always zero. Under the assumption that most migration movements take place over shorter distances,

we estimate the intra-municipal travel distance to be about $\sqrt{\frac{1}{2}|\varnothing|}$ instead. This means that the model parameters are calculated using the data sources presented in Table 1.

Table 1. Sources for the values of the different variables used in the regression.

	Intra-Municipal Migration	Inter-Municipal Migration
Migration data	One CBS dataset [32]	Multiple CBS datasets [32]
Distance data	*Estimate*: Distance between the centres of both municipalities	*Estimate*: $\sqrt{\frac{1}{2}\|\varnothing\|}$

The linear form of the gravity model presented in Equation (2) can then be fitted on the data using a Generalised Linear Model (GLM). GLMs are flexible generalisations of linear regressions, in which the response variables can have a non-normal error distribution model [35]. Because logarithms are used in this linear form and it is impossible to take the logarithm of zero, the cases in which no people migrate between two municipalities should be processed before fitting the equation.

There are two options to solve this problem: disregard the migration data when no migrants move between two municipalities in a certain year, or modify all the measured migration data in such a way that all connections have some migrants. Because choosing the first approach would mean that information is lost on municipalities that did not attract migrants, we choose the second option. Every number of migrants is increased by two, as with this value the deviation of the model is minimised. A regression on these data resulted in Equation (3) ($\chi^2(4, N = 4{,}750{,}471) = 1.4232 \times 10^6$, $P \leq 0.001$).

$$M_{a\to b} = \exp(-1.5724) \cdot \frac{p_a^{0.2436} \cdot p_b^{0.2273}}{\Delta_{a\to b}^{0.4748}} \tag{3}$$

2.4. Expansion of the Gravity Model

The impact of the regional identities can be examined by expanding the gravity model with a categorical variable ι that is true if both municipalities have the same regional identity and false if they do not. Following this introduction, the linear version of the gravity model is also adjusted to Equation (5). The used parameters are shown in Table 2.

$$M_{a\to b} = G \cdot \frac{p_a^{\alpha} \cdot p_b^{\beta}}{\Delta_{a\to b}^{\delta}} \cdot I^{[region(a)=region(b)]} \tag{4}$$

$$\ln(M_{a\to b}) = \alpha \cdot \ln(p_a) + \beta \cdot \ln(p_b) - \delta \cdot \ln(\Delta_{a\to b}) + \gamma + \iota^{[region(a)=region(b)]} \tag{5}$$

Table 2. Different parameters used in the extended gravity model.

Parameter	Description
α	Influence of p_a on the number of migrants
β	Influence of p_b on the number of migrants
δ	Influence of $\Delta_{a\to b}$ on the number of migrants
$\gamma = \log(G)$	Normalisation constant of the regression function
$\iota = \log(I)$	Increase in the number of migrants when both municipalities are located in the same identity region
$\Delta_{a\to b}$	Distance between municipality a and municipality b in kilometres
$M_{a\to b}$	Number of migrants between municipality a and municipality b
p_a	Population of municipality a
p_b	Population of municipality b

In this paper, three different sets of regional cultural identities are used. We define a regional identity as a region in which municipalities have a shared political history as proxy for cultural. This shared history is something which can be quantified and documented relatively easily, whereas researching and determining similarities in cultural traditions would be much more difficult. Each of

the three sets consists of a different numbers of identity regions, as shown in Table 3. Comparing the effects that the three different sets have might help to comprehend both the importance of identity in regions of certain sizes and the significance of choosing the right clusters of municipalities.

2.4.1. Specification of Regional Identities

In this paper, three different sets of regional identities are used. The three sets with different numbers of identity regions are shown in Table 3. Comparing the effects that the three different sets have might help to comprehend both the importance of identity in regions of certain sizes and the significance of choosing the right clusters of municipalities. Each of the three sets of regions has a shared history, which would have allowed each of the regions to develop their own cultural identity.

The first two sets of identity regions consist of administrative regions, which are designed to compare regions of certain sizes within the European Union. The first set consists of the twelve NUTS 2 regions or provinces [36], which can be traced back to the historical medieval states that merged to become the Netherlands. Some provinces such as Friesland or Noord-Brabant have strong provincial identities through historical cultural and political traditions, whereas provinces with more heterogeneous populations such as Overijssel and Gelderland do not. The second set of forty NUTS 3 regions [36] can be traced back to the COROP regions defined in the early 1970s. These regions were based on the catchment areas of large Dutch cities, by analysing daily commutes. This can indicate historical dependency. In the creation of these regions, no provincial boundaries were crossed to make the regions easier to use.

Because these two sets of regions were created for an administrative use, it can be the case that multiple cultural identity regions were merged to create the right number of regions with a certain population threshold. Regions within both sets must have a minimal number of residents to allow for accurate statistic comparison [37]. In the case of migration modelling, we might however want to keep the combination of different cultural identities to a minimum.

The third set of identity regions is manually specified through a literature study. It consists of seventy long-standing historical identity regions. The municipalities within each region have a shared history, and therefore often have similar traditions and dialect. Details on these historical regions can be found in [38]. Maps of each of the sets of regions can be found in Section 1 of the Supplementary Materials.

Table 3. Different sets of identity regions embedded in the human migration model.

Data Set	Regions	Specification
NUTS 2 (Provinces)	12	[36]
NUTS 3 (COROP regions)	40	[36]
Historic regions	70	[38]

2.4.2. Introduction of the Different Sets of Identity Regions

Using the same data as before, new models can be fitted on each of the three predefined sets of identity regions. The formula fitted on the NUTS 2 regions is shown in Equation (6) ($\chi^2(4, N = 4{,}750{,}471) = 1.4028 \times 10^6$, $P \leq 0.001$), the formula generated using the NUTS 3 regions in Equation (7) ($\chi^2(4, N = 4{,}750{,}471) = 1.2729 \times 10^6$, $P \leq 0.001$) and the formula that is based on the historic regions in Equation (8) ($\chi^2(4, N = 4{,}750{,}471) = 1.2712 \times 10^6$, $P \leq 0.001$). This means that the deviance is decreased by, respectively, 1.4%, 10.6% and 10.7%. In all cases, the coefficient for the identity influence is significant ($P < 0.001$).

$$M_{a \to b} = \exp(-1.9983) \cdot \frac{p_a^{0.2457} \cdot p_b^{0.2294}}{\Delta_{a \to b}^{0.3961}} \cdot \exp(0.2851)^{[region(a)=region(b)]} \quad (6)$$

$$M_{a\to b} = \exp(-2.3489) \cdot \frac{p_a^{0.2489} \cdot p_b^{0.2326}}{\Delta_{a\to b}^{0.3334}} \cdot \exp(1.1280)^{[region(a)=region(b)]} \tag{7}$$

$$M_{a\to b} = \exp(-2.2415) \cdot \frac{p_a^{0.2490} \cdot p_b^{0.2327}}{\Delta_{a\to b}^{0.3558}} \cdot \exp(1.3589)^{[region(a)=region(b)]} \tag{8}$$

In these three equations, two variables have changed by more than one tenth: γ and δ. The γ variable would previously have contained part of the ι variable, and is likely to have a lower value when the ι variable is introduced. Likewise, the value of δ is lowered because the variable would no longer have to account for the effects that identity has on shorter distance migration. A comparison between the effects of the new equations on the predicted number of migrants at certain distances is shown in Section 4.1.

This way of identity regions assumes that the strength of each of these regional identities is similar, as the expected number of migrants is constantly increased by the same factor when two municipalities are located within the same identity region. In practice, this is not always the case. Identity may for example play a different role in more urbanised environments than in the countryside. Adjusting the model to understand those differences is very difficult, as fitting individual parameters for the influence of each single identity region could lead to over-fitting. Different ways of defining identity regions are discussed in Section 2 of the Supplementary Materials.

2.5. Comparison of the Importance of Identity in Different Sets of Regions

When comparing the importance of identity in the different sets of identity regions, the parameter values of the ι variables that are found in Equations (6)–(8) cannot be compared, because the values of the α, β, γ and δ parameters differ as well. However, we can use the basic gravity model specified in Equation (3) to find and compare the influence of identity on human migration. This can be done by comparing the differences in the predicted numbers of migrants and the actual number of migrants for both the intraregional and interregional migration, the ϵ. This means that the ϵ values are first split into two different categories as specified in Equation (9).

$$\epsilon = \begin{cases} \epsilon_{\text{in}}, & \text{if municipalities in the same identity region} \\ \epsilon_{\text{out}}, & \text{if municipalities not in the same identity region} \end{cases} \tag{9}$$

A two-sample Kolmogorov–Smirnov test between the distribution of all ϵ_{in} values and the distribution of all ϵ_{out} values reveals that both distributions are not the same for each of the three sets of pre-specified identity regions ($p < 0.001$). This means that there are differences between interregional and intramunicipal migration flows that cannot be explained by the basic gravity model.

As a tool to compare the influence of identity in different municipalities, the formula for a municipalities' Identity Comparison Measure (ICM) is specified in Equation (10). The ICM value takes on positive values when the basic model is worse at explaining migration inside a region than migration outside a region. Considering that the basic model is fitted on all migration data, this means that people in that region are more likely to move towards municipalities that are located in the same area than towards municipalities that are not.

$$ICM = \text{avg}(\epsilon_{\text{in}}) - \text{avg}(\epsilon_{\text{out}}) \tag{10}$$

This ICM value can thus tell us about the difference between the variance that can be explained in the intraregional migration flows, and the explained variance in the interregional migration data. A positive ICM value indicates that the model has more difficulties to explain the intraregional migration behaviour than it has to explain the interregional migration behaviour. An ICM value can be calculated for each separate municipality. This is done by only using all ϵ_{in} and ϵ_{out} values for all migration flows originating in that particular municipality. Which migration flows are part of the

intraregional migration figures and which flows are part of interregional migration figures depends on the used identity regions.

A comparison between these different ICM values can give an indication as to what regions contain stronger identities, or whether the municipality is part of the right identity region. This does not imply that an ICM value can be translated directly to the influence regional identity has on migration. Instead, the ICM value defines the unexplained differences in the remaining deviance that could not be included in the γ and δ variables.

2.6. Creation of Randomly Generated Identity Regions

Apart from defining the predefined identity regions, we randomly generate identity regions to be able to analyse the quality to those predefined identity regions. The fact that a certain set of identity regions produces positive ICM values does not mean much, as such values could also occur in randomly generated regions, just because the municipalities within each region are located in proximity to one another.

Using entirely randomly generated regions as a comparison would not be realistic; municipalities located within the same identity regions are generally not scattered all over the country, but present in a cluster of municipalities. A more realistic approach to generating random regions would thus enforce that municipalities located within a region should at least form a spatial cluster together.

The k-means algorithm is used to create such spatial clusters of municipalities [39]. Instead of randomly assigning each municipality to a region, a random centre point is assigned to each region. Each municipality is then assigned to the closest centre point, after which the centre point of each region becomes the geographical centre of the municipalities belonging to that region. Some municipalities might then be located closer to the centre point of another region, which means that the that municipality is relocated to that other region. This process is repeated until no more municipalities are reassigned to another region. A visual representation of this algorithm is shown in Figure 1.

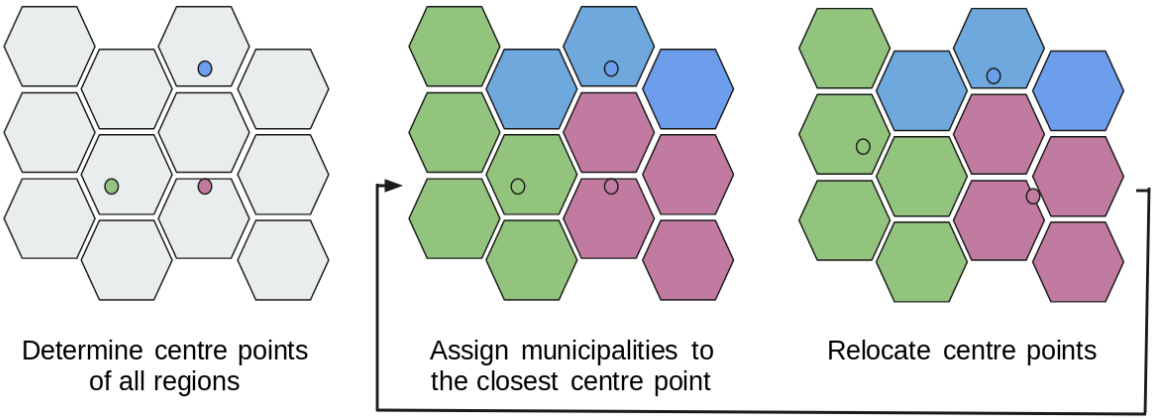

Figure 1. Visual representation of the steps taken to assign municipalities to spatially clustered regions using the k-means algorithm. This algorithm can be applied to generate different numbers of regions. This number of regions is controlled by the variable k. To be able to compare the generated regions with a certain set of identity regions, this k is set equal to the number of regions present in this set of identity regions.

Optimisation of a Set of Identity Regions

By optimising the predefined identity regions and the randomly spatially clustered regions to generate higher ICM values, we can get a better understanding of the regions used. This can be done by comparing the mean and median ICM values of the non-optimised and optimised regions. A set of regions can be optimised by increasing the average ICM value for each region. As this value increases, the differences in the predictive value of the model between the interregional and

intraregional migration are enlarged. When this difference becomes larger, the strength of the identity contained within the defined regions also becomes larger. For the algorithm behind our optimisation strategy, see Section 3 of the Supplementary Materials.

3. Results

3.1. Influence of the Embedding of Identity Regions on the Expected Number of Migrants

When the identity regions are introduced to the model equation, the expected number of migrants between two municipalities changes. In Figure 2, the percentage increase or decrease in the absolute number of migrants is shown when both municipalities have the same number of inhabitants. The figure shows that the predicted number of migrants towards municipalities inside the same NUTS 3 and Historic identity regions is much larger than initially predicted, while the predicted number of migrants towards municipalities inside the same NUTS 2 region is similar. On the other hand, the predicted number of migrants towards municipalities in other identity regions is lower by up to 50% over shorter distances. More surprisingly, the predicted numbers of migrants for larger distances has decreased as well.

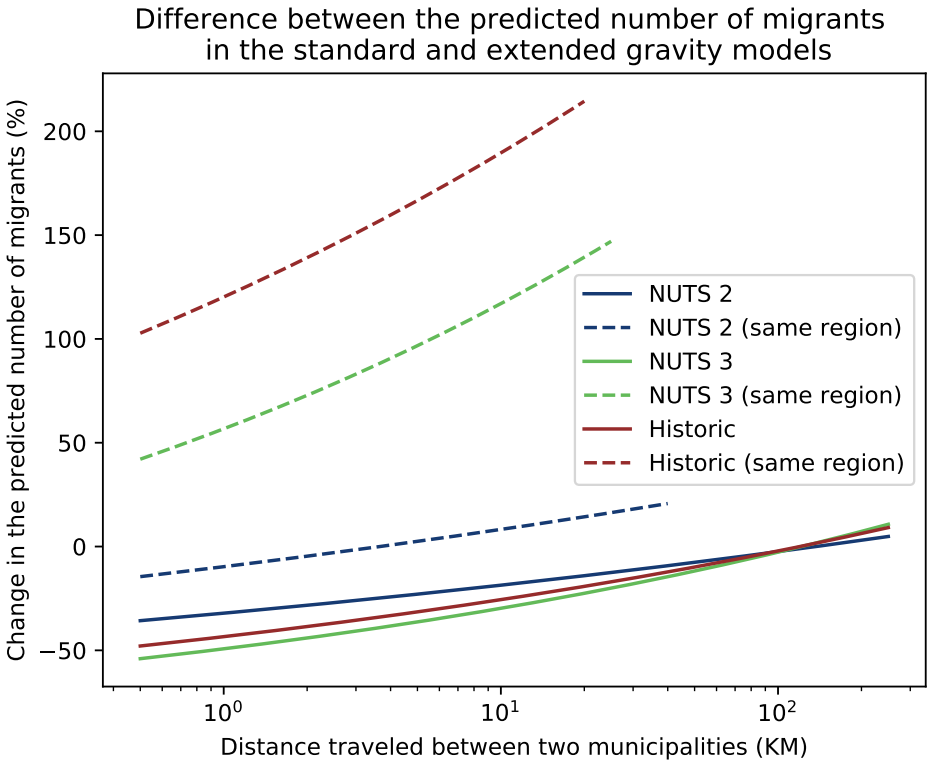

Figure 2. Difference between the expected number of migrants found in Equations (3) and (6)–(8). When two municipalities are located in the same identity region, the number of migrants in the basic gravity model is much lower than expected in the extended models, whereas this is the other way around for municipalities that are not located within the same region. The maximum distance travelled by migrants within the same region is cut to the reflect the size of the regions within each of the sets of regions.

3.2. Comparison of Median ICM Value Distributions

A comparison between the median *ICM* values of the different specified identity region configurations is shown in Table 4. Because the mean value distributions can be useful to better understand the underlying distributions, the mean value distributions for the same configurations can be found in Section 4 of the Supplementary Materials.

A short analysis of the table immediately shows that the NUTS 2 regions seem to perform worse than the sets of NUTS 3 and historic regions. Whereas the median ICM value distributions of the latter two optimised and non-optimised predefined regions fall inside the distributions of the median ICM values of the randomly spatially clustered regions, this is not the case for the NUTS 2 regions. The median ICM value of the NUTS 3 and historic regions is similar to the median ICM value distributions of the randomly generated spatially clustered regions. For the NUTS 3 regions, the median ICM value is even on the upper side of this distribution.

Table 4. A comparison between the median ICM values of the three different identity region configurations and the median ICM value distributions of the same number of randomly generated spatially clustered regions, both optimised and non-optimised. Fifty samples were taken for distributions that involve the optimisation algorithm and 250 samples were taken for each of the non-optimised distributions that consist of randomly spatially clustered regions.

		Predefined		**Spatially Clustered**	
		95% CI	**Max.**	**95% CI**	**Max.**
NUTS 2	**Default**		12.34	[12.60, 13.39]	13.79
	Optimised	[13.02, 13.75]	13.75	[13.43, 14.24]	14.38
NUTS 3	**Default**		33.03	[28.23, 33.33]	34.23
	Optimised	[37.49, 39.12]	39.12	[36.88, 40.54]	40.59
Historic	**Default**		42.34	[40.11, 47.26]	49.73
	Optimised	[54.48, 59.51]	59.94	[53.96, 61.83]	62.05

When the clustering algorithm is applied, the randomly generated regions are likely to partially overlap with the identity regions, because both types of regions are spatially clustered. This means that the ICM values become positive, and the ICM values are larger than they would be in purely randomly generated regions. These patterns did not change after varying the α, β, γ and Δ variables by 10%.

4. Discussion

4.1. Effect on the Influence of the Distance Parameter

When the three different sets of identity regions were introduced to the gravity model, the δ parameter, which controls the influence of distance on the number of migrants, changed. It decreased from 0.4760 in the original model to 0.3987 (NUTS 2 regions), 0.3373 (NUTS 3 regions) and 0.3493 (historic regions). This mainly has a big influence on the predicted number of migrants over smaller distances , as it is at these distances that migrants can choose to move within or outside their identity region. . This means for instance that if the δ parameter were used as a descriptive characteristic of a population, then an error would be introduced in its inference if the confounding effect of identity would not be into account.

4.2. Model Limitations

An assumption made in this model is that we expect amenities to to be uniformly distributed over the country, which means that they should therefore have no effect on the ICM value. Examples of such amenities are industrial areas, natural parks, and available housing. If this were not true, then it would be possible that strong migration patterns due to amenities are mistakenly inferred as being intra-regional or inter-regional, which could, respectively, inflate or deflate the strength of identity regions that we find. However, some amenities can start to have unequal effects when there are large regional differences in the availability of amenities, rather than large local differences. For example, when employment possibilities in a certain identity region are very low, people could be forced to migrate towards another region where they can get a job. This migration would not be related to their identity, but to employment opportunities. Similar situations can occur when certain educational

possibilities are not available within the identity region. On the other hand, certain amenities can also increase the *ICM* value. When the industry in a region is highly specialised, we might find that people are more likely to move within their region than outside the region because they want to keep working in the same industry, and receive specialised education. Taking the availability of different amenities into account could improve the estimate of the relative role of identity regions. However, since the main point of the present article is to demonstrate that the role of identity is significant, not so much its exact value, we leave this for future work.

Forced migration is another case of migration that can have a big negative effect on the *ICM* value of a region. However, in the Netherlands, it is quite rare: in our dataset, the relocation of hundreds of newly registered asylum seekers to other immigration centres is the only example of such forced migration events.

While incorporating more variables will create more exact and accurate *ICM* values, we do not expect this to change our overall conclusion. Identity plays a significant role in the modelling of human migration, and is an effect which is not to be ignored.

4.2.1. Strong Geographical Identities Require Little Displacement

Something that has to be kept in mind when applying this model in other situations is that analysing Dutch cultural identity can differ from analysing other cultural identities. Within the Netherlands, most of these Dutch identities are bound to a specific geographical region, something that might not be the case in other countries. In other areas of the world, where there is a lot of internal displacement of people due to natural disasters, conflicts and forced movement, it can be much more difficult to pinpoint cultural identities to geographical regions. This means that the model might not be applied as is to, for example, African or Southeast Asian migration patterns. An analysis of cultural similarities between different displaced or migrated communities could provide the information needed to include the concept of identity into a model instead. A similar approach might be needed for research in Australia or the United States, areas in which people tend to have less strong ties to certain smaller geographical regions as well.

4.2.2. Identity Regions

In this research, we assumed that historical political boundaries relate to the cultural identity of the residents in an area. This approach allows the idea of identity to be implemented into a model, without requiring extensive research into local and regional culture. However, the used sets of regions are not always perfect. In all three sets of regions, some municipalities can be relocated to other regions in such a way that their negative *ICM* values become positive, or to seriously increase their *ICM* value. This could mean that some municipalities are not located in the right cultural identity region, or that other variables such as the availability of certain amenities outweigh the influence of identity in these areas.

The presence of such municipalities might indicate that the regions could still be improved, but does not mean that those sets of regions are not well defined at all. A comparison between the NUTS 3 regions and their randomly generated spatially clustered counterparts shows that the mean *ICM* value of the NUTS 3 regions set is always larger, and the median *ICM* value in the NUTS 3 regions set is similar to the highest median *ICM* values found in the randomly generated spatially clustered regions. This means that the regions defined in the NUTS 3 set perform better than the randomly generated spatially clustered regions. The same reasoning also holds for the set of historic regions.

4.3. Development of Identity Influence over Time

The same extended migration models can also be fitted on yearly data instead. The resulting developments in both the predicted numbers of migrants and the influence of identity over time are shown in Figure 3.

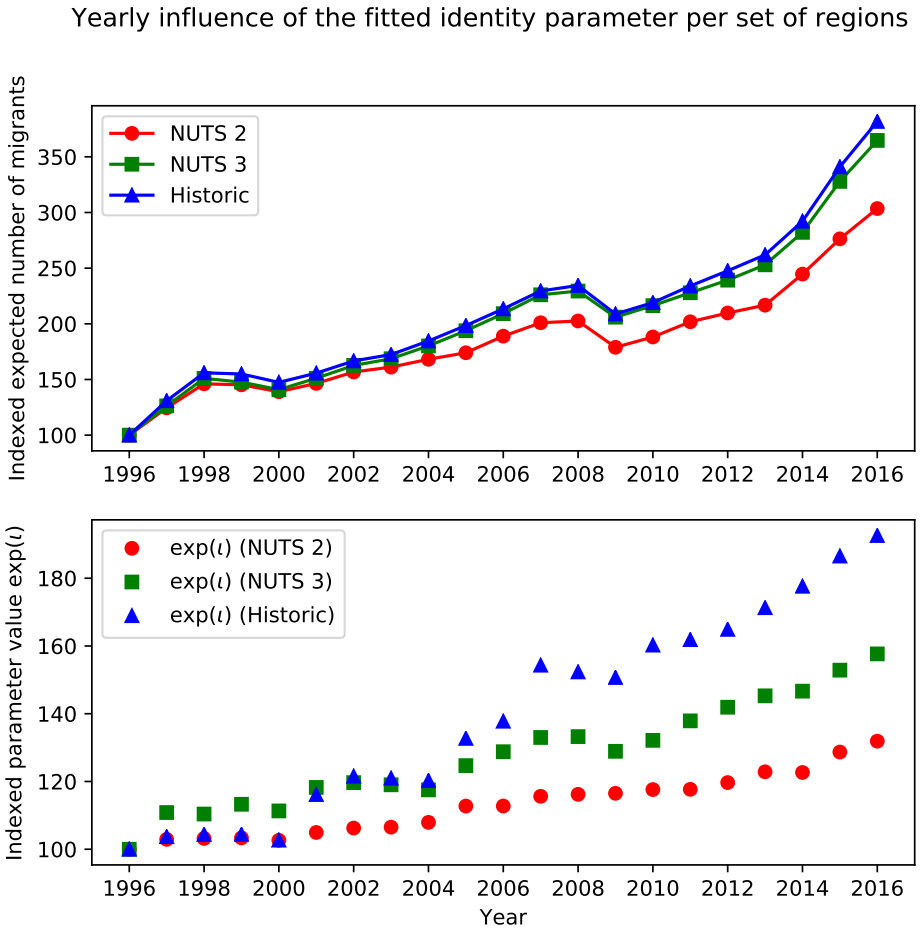

Figure 3. Indexed changes in the predicted number of migrants between two municipalities with 50,000 residents that are located 50 km apart and the indexed changes in the exp(ι) value inside the model for each of the three sets of identity regions between 1996 and 2016. For every year, all municipalities were merged to form the municipalities that existed in 2016 to prevent the different numbers of municipalities from having any influences.

Between 1996 and 2016, we observe a tripling in the expected number of migrants, and a gradual increase in the exp(ι) values. In the case of the NUTS 2 regions, we observe a nearly 60% increase in the value of exp(*iota*). The importance of regional identity in the migration decision is lowered at two different moments: around 2000 and in 2009. This correlates with the moments at which the total numbers of expected migrants is lowered as well. While a link might be made to the economic recession in both years, further investigation on the nature of these patterns is needed.

5. Conclusions

The expansion of the gravity model with one of three different sets of identity regions results in three different sets of parameter values in Equation (6) that showed how the influence of these identity regions should be incorporated into the model. In particular, when the NUTS 3 and historic regions are added to the gravity model, the resulting Equations (7) and (8) show that Dutch people are, respectively, 3.09 and 3.89 times as likely to move to a municipality within the same identity. NUTS 2 regions turn out to have less predictive value, as Equation (6) shows that people were 0.33 times more likely to move towards a municipality within the same NUTS 2 region.

Including either the NUTS 3 or historic regions in the migration model decreases the deviation of the model by, respectively, 10.6% and 10.7%. Again, the larger NUTS 2 regions seem to have

less predictive power: including the NUTS 2 regions only decreased the deviation of the model by 1.4%. Table 4 shows further evidence that the NUTS 3 and historic regions are more relevant; the median *ICM* values of both sets of regions are similar to their randomly generated spatially clustered counterparts, whereas the *ICM* values of the NUTS 2 regions are not. This is strange, as randomly clustered regions should have a very similar structure to the original sets of regions, and the median ICM values should be similar. The same applies to their optimised counterparts. This further supports the use of smaller sized regions whenever this is possible.

We have demonstrated that the incorporation of identity regions does not have to be difficult. By choosing to model the identity regions with strict boundaries, we created a way of incorporating identity regions that can easily be used in other regions as well. Through literature research, it should be possible to specify sets of historic identity regions in these other regions too. Overall, this makes us believe that introducing regional identities in human migration models is an easy way to enhance their performance.

Because regional identity seems to be an important factor in the human migration decision, it would be interesting to embed these identity regions in other types of human migration models as well. When such models are used on regions in different circumstances, this could provide more information about the influence of regional identity itself. It would be interesting to see what would happen in more segregated societies, or in societies where most people regularly travel over larger distances than in the Dutch society. The *ICM* value could also be used to detect the formation of regions over time. This could be particularly useful in areas that were colonised or have seen enormous change in society. This variable could help to map these changes in society onto changes in local migration behaviour, which in turn might help in understanding the regional identities in a certain area.

Supplementary Materials: The following are available online at http://www.mdpi.com/2413-8851/3/3/78/s1.

Author Contributions: Conceptualisation, W.R.J.V., D.R. and R.Q.; Formal analysis, W.R.J.V.; Investigation, W.R.J.V. and D.R.; Methodology, W.R.J.V.; Project administration, W.R.J.V.; Resources, W.R.J.V. and D.R.; Software, W.R.J.V.; Supervision, D.R. and R.Q.; Validation, D.R.; Visualisation, W.R.J.V. and D.R.; Writing—original draft, W.R.J.V.; and Writing—review and editing, D.R. and R.Q..

Funding: D.R. acknowledges the support from the Dutch NWO, eScience project number 027.015.G05 "DynaSlum: Data Driven Modeling and Decision Support for Slums".

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GLM	Generalised Linear Model
ICM	Identity Comparison Measure
NUTS	Nomenclature of Territorial Units for Statistics

References

1. UN. *World Migration Report*; Technical Report; International Organization for Migration: Grand-Saconnex, Switzerland, 2018.
2. Skeldon, R. International Migration, Internal Migration, Mobility and Urbanization: Towards More Integrated Approaches. *IOM Migr. Res. Ser.* **2018**. [CrossRef]
3. Bell, M.; Charles-Edwards, E. *Cross-National Comparisons of Internal Migration: An Update on Global Patterns and Trends*; Technical Report; United Nations Department of Economic and Social Affairs: New York, NY, USA, 2013.
4. Sjaastad, L. The Costs and Returns of Human Migration. *J. Political Econ.* **1962**, *70*, 80–93. [CrossRef]
5. Stark, O. *The Migration of Labor*; Blackwell Publishers: Oxford, UK, 1991.
6. Tiebout, C. A pure theory of local expenditures. *J. Political Econ.* **1956**, *64*, 416–424. [CrossRef]
7. Graves, P.; Linneman, P. Household migration: Theoretical and empirical results. *J. Urban Econ.* **1979**, *6*, 383–404. [CrossRef]

8. Grigg, D. E. G. Ravenstein and the "laws of migration". *J. Hist. Geogr.* **1977**, *3*, 41–54. [CrossRef]
9. Kok, J. Choices and constraints in the migration of families: The central Netherlands, 1850–1940. *Hist. Fam.* **2004**, *9*, 137–158. [CrossRef]
10. Hipp, J.; Boessen, A. Immigrants and Social Distance. *Ann. Am. Acad. Political Soc. Sci.* **2012**, *641*, 192–219. [CrossRef]
11. Bauer, T.; Zimmermann, K. Network Migration of Ethnic Germans. *Int. Migr. Rev.* **1997**, *31*, 143–149.10.2307/2547262. [CrossRef]
12. Massey, D. *Migration: Motivations*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 15, pp. 452–456. [CrossRef]
13. Greenwood, M. Human migration: Theory, models, and empirical studies. *J. Reg. Sci.* **1985**, *25*, 521–544. [CrossRef]
14. Lucassen, J. *In Search of Work*; IISG (International Institute of Social History) Research Papers; International Institute of Social History: Amsterdam, The Netherlands, 2000.
15. Ranis, G.; Fei, J. A Theory of Economic Development. *Am. Econ. Rev.* **1961**, *51*, 533–565.
16. Bauer, T.; Zimmermann, K. *Assessment of Possible Migration Pressure and Its Labour Market Impact Following EU Enlargement to Central and Eastern Europe*; IZA Research Reports; IZA—Institute of Labor Economics: Bonn, Germany, 1999; Volume 3.
17. Wallerstein, I. A world-system perspective on the social sciences. 1974. *Br. J. Sociol.* **2010**, *61* (Suppl. 1), 167–176. [CrossRef] [PubMed]
18. Piore, M. *Birds of Passage: Migrant Labor and Industrial Societies*; Cambridge University Press: Cambridge, UK, 1979. [CrossRef]
19. Wolpert, J. Behavioral aspects of the decision to migrate. *Pap. Reg. Sci.* **1965**, *15*, 159–169. [CrossRef]
20. Todaro, M. A model of labor migration and urban unemployment in less developed countries. *Am. Econ. Rev.* **1969**, *59*, 138–148.
21. De Haas, H. The internal dynamics of migration processes: A theoretical inquiry. *J. Ethn. Migr. Stud.* **2010**, *36*, 1587–1617. [CrossRef]
22. Crawford, T. Beliefs about birth control: A consistency theory analysis. *Represent. Res. Soc. Psychol.* **1973**, *4*, 53. [PubMed]
23. Hoda Rahmati, S.; Tularam, G. A critical review of human migration models. *Clim. Chang.* **2017**, *3*, 924–952.
24. Stewart, J. The development of social physics. *Am. J. Phys.* **1950**, *18*, 239–253. [CrossRef]
25. Poot, J.; Alimi, O.; Cameron, M.; Maré, D. The gravity model of migration: The successful comeback of an ageing superstar in regional science. *Investig. Reg.* **2016**, *2016*, 63–86.
26. Ramos, R.; Suriñach, J. A Gravity Model of Migration Between the ENC and the EU. *Tijdsch. Econ. Soc. Geogr.* **2017**, *108*, 21–35. [CrossRef]
27. Simini, F.; González, M.C.; Maritan, A.; Barabási, A.L. A universal model for mobility and migration patterns. *Nature* **2012**, *484*, 96–100. [CrossRef] [PubMed]
28. Yang, Y.; Herrera, C.; Eagle, N.; González, M. Limits of Predictability in Commuting Flows in the Absence of Data for Calibration. *Sci. Rep.* **2014**, *4*, 5662. [CrossRef] [PubMed]
29. Ren, Y.; Ercsey-Ravasz, M.; Wang, P.; González, M.; Toroczkai, Z. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nat. Commun.* **2014**, *5*, 5347. [CrossRef] [PubMed]
30. Kang, C.; Liu, Y.; Guo, D.; Qin, K. A Generalized Radiation Model for Human Mobility: Spatial Scale, Searching Direction and Trip Constraint. *PLoS ONE* **2015**, *10*, e0143500. [CrossRef] [PubMed]
31. Robinson, C.; Dilkina, B. A Machine Learning Approach to Modeling Human Migration. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '18), San Jose, CA, USA, 20–22 June 2018; ACM: New York, NY, USA, 2018; pp. 30:1–30:8. [CrossRef]
32. Vermeulen, W. *Dutch Migration Datasets—Computational History*; Computational History: Diemen, The Netherlands, 2019.
33. Piovani, D.; Arcaute, E.; Uchoa, G.; Wilson, A.; Batty, M. Measuring Accessibility using Gravity and Radiation Models. *R. Soc. Open Sci.* **2018**, *5*, 171668. [CrossRef] [PubMed]
34. Phibbs, C.; Luft, H. Correlation of Travel Time on Roads versus Straight Line Distance. *Med Care Res. Rev.* **1995**, *52*, 532–542. [CrossRef]
35. Nelder, J.; Wedderburn, R. Generalized Linear Models. *J. R. Stat. Soc. Ser.* **1972**, *135*, 370–384. [CrossRef]

36. Eurostat. *NUTS 2 Regions in The Netherlands, 2010 and 2013*; Eurostat: Luxembourg, 2013.
37. Eurostat. *Principles and Characteristics*; Eurostat: Luxembourg, 2018.
38. Vermeulen, W. *Dutch Historical Regions—Computational History*; Computational History: Diemen, The Netherlands, 2019.
39. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.