Supplementary Materials: Modelling the Influence of Regional Identity on Human Migration

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

1. Maps of the sets of regions used

It can be difficult to determine what municipalities should be part of the same identity region, and how many identity regions should be created as a whole. There will not always be one right answer: related regional identities could be combined or split, as evidenced by the relatively small differences in *ICM* values between the set of NUTS 3 regions and the historic regions. Ultimately, we chose to use the Dutch NUTS 2 and NUTS 3 regions, as well as a finer set of historic regions. Maps of these sets of regions are shown in Figures S1, S2 and S3. Ultimately the geographical size of the regions might not matter too much - in other areas of the world identity regions could span much larger areas than in the Netherlands. Sets of regions can be used, as long as the used regions are well researched.

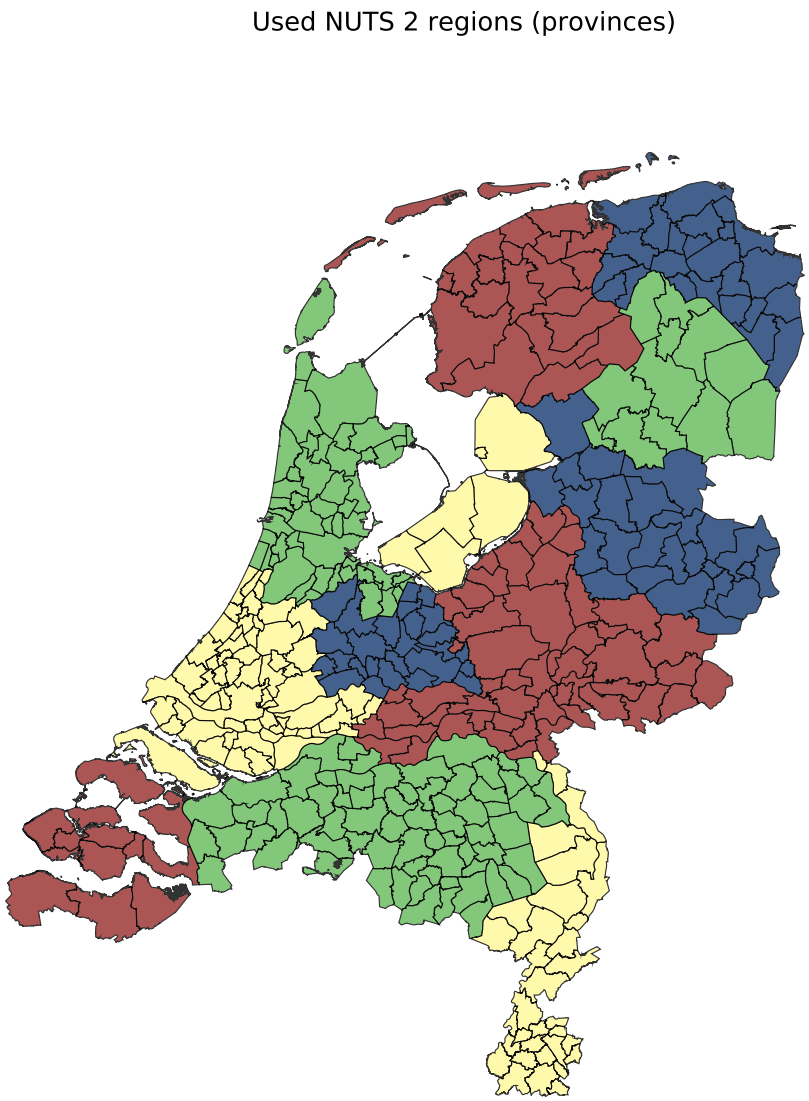

Figure S1. Geographical representation of the used NUTS 2 regions.

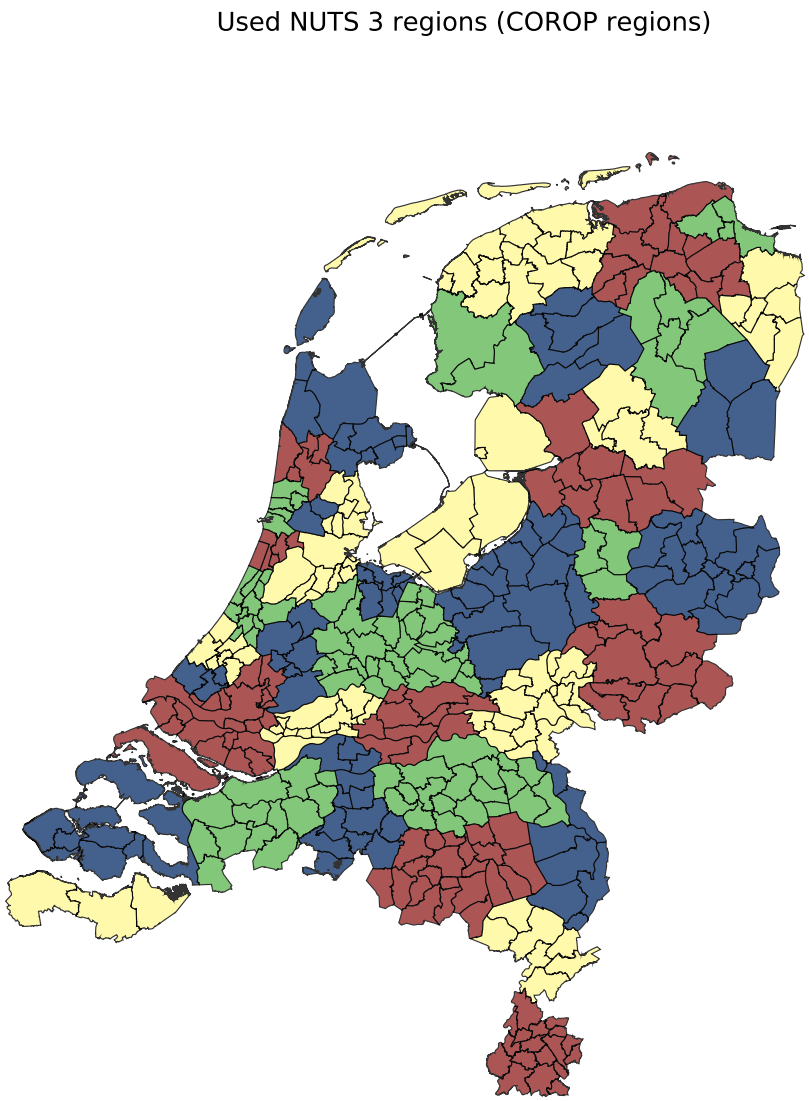

Figure S2. Geographical representation of the used NUTS 3 regions.

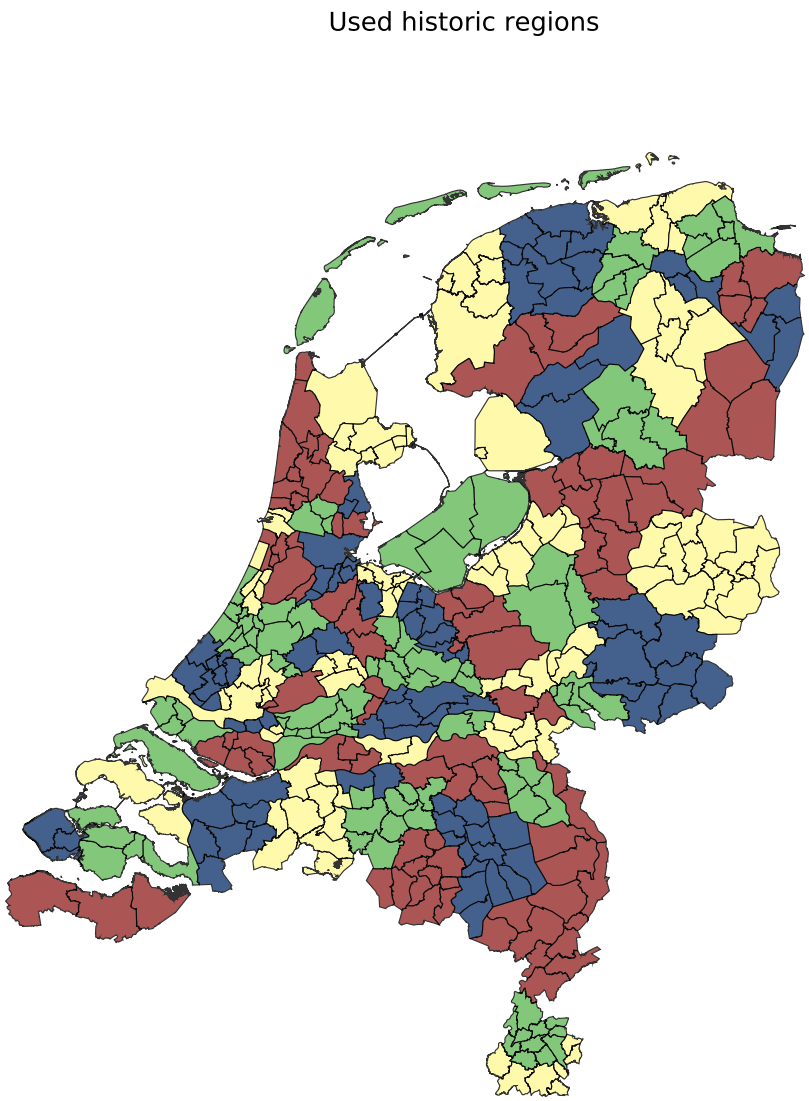

Figure S3. Geographical representation of the used historic regions.

2. Boundary specification

In this paper identity regions are defined as hard-bounded areas. This method makes it easier to specify such regions through literature research, and can be applied almost everywhere. Using hard-bounded identity regions does however also create challenges, as real identity regions are usually not hard bounded. Some municipalities could be part of multiple identities, and in other municipalities there might be small minority groups of people that hold another identity. A simple approach to solve this problem would be to introduce fuzzy boundaries. In this solution municipalities that are located close to a certain region are assumed to house people that also have the same regional identity as people living in that region.

A more sophisticated approach to this problem would be to research the social connectivity between municipalities. By doing this, all different identities that make up a certain municipality could accurately be represented. This would create an identity network as shown in Figure S4. It could however be difficult to use this approach, because detailed data on every individuals social connections has to be acquired. Such data is often hard to acquire or not available at all. As result, this way of incorporating regional identities is not applicable in most situations.

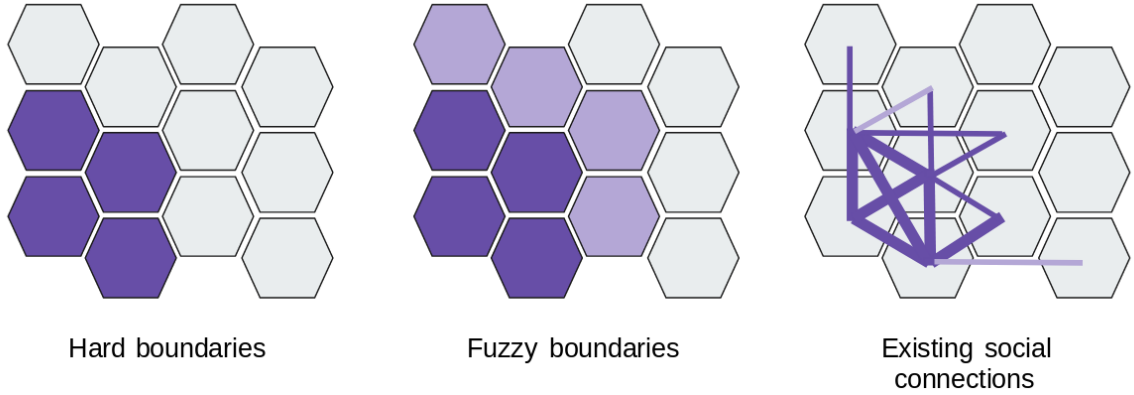

Figure S4. Three different approaches to defining identity region: with hard boundaries, fuzzy boundaries, or by looking at the existing connections between two municipalities.

3. Optimisation of a set of identity regions

Determining the configuration that fully optimises the *ICM* value is however practically impossible for larger numbers of municipalities, as this problem can be reduced to the clique problem. The clique problem has been proven to be NP-complete [**Karp1972**]. To overcome this problem we specified an algorithm to increase the *ICM* values in Algorithm 1, under the constraint that every resulting region has at least two municipalities. Without this constraint it would become impossible to calculate the *ICM* value.

Data: a set of regions, each containing at least two municipalities
Result: a set of regions with a better average *ICM* value than before
initialise current regions filled with municipalities;
do
  initialise new regions empty;
  for *every municipality in the Netherlands* **do**
    determine current region;
    determine the regions neighbouring the municipality;
    determine optimal region using Equation 1;
    if *optimal region different than current region* **and** *current region will have at least two municipalities in the new regions* **and** *fifty percent chance* **then**
      add municipality to the optimal region in the new regions;
    else
      add municipality to the current region in the new regions;
    end
  end
  current regions become the new regions;
while *not every municipality in optimal region* **and** *these municipalities are not the same for the last five iterations*;

Algorithm 1: optimisation algorithm used to increase the average ICM value for a set of regions.

The function optimised to find the optimal region for a certain municipality is shown in Equation 1. In this function the notation $\tilde{M}_{a \to b,years}$ is used to describe the median number of migrants between two municipalities a and b in a certain range of *years*. By using median values in this equation, outlier data will have very little influence on the optimisation process. In this formula the effects of a municipality relocation are evaluated by looking at the change in the *ICM* values of the municipalities in that region R, and the change in the *ICM* value of the municipality itself. Since identity regions are usually not scattered all over the country, this optimisation algorithm is limited to only assign a municipality to one of the regions that also contains a municipality that is located within a distance of twenty kilometres. This distance includes the maximum distance between centre points of neighbouring municipalities,

which makes it the best approximation to determining the neighbouring municipalities by determining whether two municipalities share borders.

$$\frac{\sum_{\forall b \in \mathrm{R}, b \neq a} \tilde{M}_{a \to b, years} + \tilde{M}_{b \to a, years}}{|\{b \mid \forall b \in \mathrm{R} \, | b \neq a\}|} \tag{1}$$

In order to prevent the algorithm from creating a deadlock situation, there is only a fifty percent chance of reassigning a municipality to the determined optimal region, given that there will still be two municipalities left in the region the municipality belonged to. When this probability is not introduced, situations can occur in which two municipalities that should be in the same region can never end up together. To increase performance, the algorithm could be adjusted to allow parts of regions to move towards other regions, instead of single municipalities. Once no municipalities can be relocated to create more optimal regions or the same set of municipalities is relocated for five iterations, the municipality relocation process is ended.

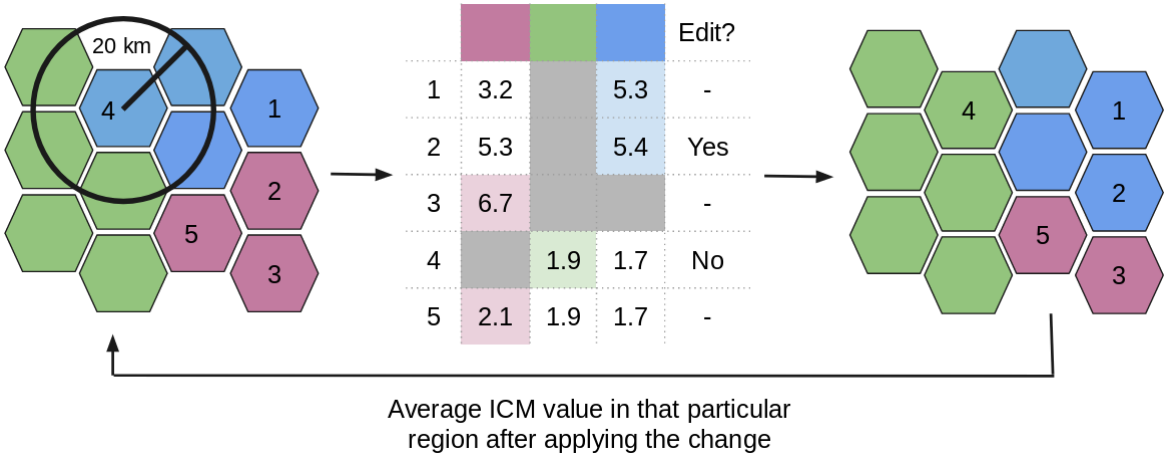

Figure S5. A visual representation of the optimisation algorithm. Given a certain starting configuration, it is determined what regions are located within a distance of 20 kilometres from each existing municipality. For each of these municipalities, the change in the global average *ICM* value is measured when a municipality would be part of that region. Every municipality that should be part of another region than it already was, is then relocated with a chance of 50%. This process is repeated until no more municipalities are relocated for three iterations.

The resulting region configuration is a local optimum. Because there are many of such local optima, this means that the algorithm will have to be executed several times to find the optimal configuration that can be reached from a particular starting configuration. Because the algorithm does not accept changes that lower the *ICM* value, this does not necessarily mean that the optimal configuration of regions can be reached. As a result, we cannot say for sure that the most optimal local optimum accessible is in fact the global optimum. We can however use this distribution of local optima as an estimate for the global optimum.

4. Differences in the mean ICM values

		Predefined		Spatially clustered	
		95% CI	Max.	95% CI	Max.
NUTS 2	Default		20.91	[20.06, 21.25]	21.58
	optimised	[21.61, 22.68]	22.68	[21.05, 22.40]	22.78
NUTS 3	Default		59.57	[45.27, 51.04]	52.76
	optimised	[58.53, 59.91]	59.92	[51.63, 57.55]	57.67
Historic	Default		74.02	[60.91, 70.72]	72.73
	optimised	[78.42, 80.95]	81.13	[75.93, 81.54]	81.74

Table S1. A comparison between the mean ICM values of the three different identity region configurations and the mean 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, 250 samples were taken for each of the non-optimised distributions that consist of randomly spatially clustered regions.

A comparison between the mean ICM values of the different specified identity region configurations is shown in Table S1. Just like the median ICM value distributions shown in the main text, the mean ICM values of the NUTS 2 regions are indistinguishable from the mean ICM values of a set of the same number of randomly generated spatially clustered regions.

The mean ICM values of the sets of NUTS 3 and historic identity regions are much larger than their randomly generated spatially clustered counterparts. When optimised further, the distribution of the mean ICM values of the historic regions falls inside of the distribution of mean ICM values of randomly generated regions, but this is not the case for the NUTS 3 regions. Even the mean ICM value of the unoptimised NUTS 3 regions is larger than the mean ICM values measured in the optimised randomly generated spatially clustered regions. When parameters in the model change by 10% these conclusions still hold.

Abbreviations

The following abbreviations are used in this manuscript:

ICM	Identity Comparison Measure
NUTS	Nomenclature of Territorial Units for Statistics