On the estimation of a gravity model for regional population projections: a geographically weighted Poisson regression with fixed effects

Trond Husby, Andries de Jong

13 mei, 2020

1 Introduction

Internal migration is the most uncertain and the most important element of a regional population projection model. Due to falling birth and death rates, fluctuations in the regional population largely follows fluctuations in migration. While top-down models treat internal migration in a fairly rudimentary way, bottom-up models include bilateral migration flows as well as its underlying drivers. This paper reports on the estimation of a gravity model used for the tri-annual regional population projections of the Netherlands Environmental Assessment Agency (PBL) and Statistics Netherlands (CBS).

The workhorse on the regional population projections of CBS and PBL is the cohort-component model PEARL. This is a bottom-up demographic model where internal migration flows between all Dutch municipalities is modelled through a combination of time series forecasting and explicit representation of the migration process. As is common in the demographic literature, a distinction is made between long- and short-distance migration: while the former is primarily motivated by family and job and is fairly stable in time, short-distance migration which is often motivated by housing-market is much more volatile. For this reason, in PEARL long-distance migration is projected with trend extrapolation while short-distance migration is modelled explicitly.

The question is how much detail to include in the modelling of short-distance migration. On the one hand, the inclusion of additional explanatory variables could certainly improve model fit and predictive power. On the other hand, a projection requires that all explanatory variables themselves can be forecasted and that their relationship with the independent variables does not change through time. As is well known in the literature, it turns out that simpler models often offer more robust projections than more complicated ones. For PEARL we use one of the classics from the literature; namely the origin-constrained gravity model.

Constrained gravity models, as discussed in Wilson (1971), extend the traditional gravity models by constraining the estimated parameter values such that estimated flows match the observed total flows. For example, origin-constrained models ensure that the estimated total flows from each origin are match the observed total flows. This is of course a very useful feature in predictive analysis. A further extension was provided by A. S. Fotheringham (1983a), who introduced the competing destination framework to improve the treatment of destination interdependence. Destination interdependence is incorporated in the model through a centrality variable relating the attractiveness of a destination to that of other potential destinations.

In terms of estimation strategy, the most straightforward procedure for the basic gravity model is log-linearisation of both the left- and the right-hand-side and estimation with OLS. The single-constrained gravity models require some additional rewriting of the terms in order to properly account for the constraints (Fotheringham and O'Kelly 1989). However, as perhaps most clearly explained by Silva and Tenreyro (2006) in the context of gravity models of trade, the parameters of log-linearised models estimated by OLS are biased in the context of heteroskedasticity. A simple solution to this problem had already been proposed by Flowerdew and Aitkin (1982); namely to estimate the gravity model using Poisson regression. Constraints can captured in the Poisson regression model by including fixed effects Davies and Guy (1987).

Another issue with estimation is potential non-stationarity of parameters (Fotheringham, Charlton, and Brunsdon 1997). In the context of a gravity model of migration it is entirely conceivable that some relationship differ across space. This may be due to differences in population composition, spatial variation in attitudes or other contextual issues, or simply a result of variations in spatial structure(A. S. Fotheringham 1983b).

For example, Nakaya (2001) found a much stronger distance decay effect in urban regions than in peripheral ones, suggesting that the degree to which distance deters migration increases with the degree of urbanisation. One of the most well-known estimation procedures for such models is Geographically Weighted Regression (GWR) (Brunsdon, Fotheringham, and Charlton 1996).

Although GWR has been used to estimate constrained gravity models (Nakaya 2001; Kalogirou 2016), the simple fixed-effects Poisson regression method has to our knowledge never been applied. Furthermore, usage of panel data in the literature is scant. For our purpose - namely the usage in a projection model - both of these features are essential: by interacting regional and yearly fixed effects we effectively estimate parameters subject to a constraint each year. The resulting parameters capture spatial variation subject to the restrictions imposed by the fixed effects.

In this paper we propose a method for estimating an origin-constrained gravity model of the competing destination flavour, using geographically weighted Poisson regression (GWPR) with fixed effects on panel data. The method is applied to estimate short-distance migration flows (less than 35 km) between 380 Dutch municipalities in the period 2012 to 2017. Since our primary concern here is projections, we evaluate the predictive perfmance of the model on a hold-out sample.

Section 1 discusses

2 Methodology

The topic of this paper relates to spatial interactions, which can be described broadly as any observed (aggregate) flows over space resulting from decisions. Therefore, spatial interaction modelling generally tries to explain geographical flows emanating from individual decision process, rather than the decision processes themselves. This type of models can be employed to examine the characteristics of destinations that experience large inflows of migrants. And, of relevance for this paper, they can be used to make statements on the relative importance of these characteristics and how it varies between origins.

One popular type of spatial interaction models is the gravity model. As the name suggests this model takes its inspiration from physics, depicting the strength of a relationship between two objects as a function of the mass of the two and the distance between them. The gravity model is widely used to study directional graphs such as the migration and trade flows and trips to retail locations. For the analysis of migration flows $M_{i,j}$ is an $n \times (n-1)$ matrix representing the number of individuals who moved from location i to location j, where $i \neq j$.

A basic gravity model of migration can be formulated using only population at destination $P_{j,t}$ and the distance between origin and destination, $D_{i,j,t}$ as explanatory variables. In this model, the population at destination is a general measure of the attractivenes of j as a destination and must therefore be understood as a proxy for a range of other factors that influences the destination choice. The distance between origin and destination has a deterring effect on all kinds of spatial interaction: all else being equal, interaction is less intense when origin and destination are far apart than when they are near each other.

2.1 The competing destinations gravity model

As discussed in the literature (REF!), a shortcoming in the basic gravity model is that it does not properly capture the effects of spatial structure. The competing destinations framework, as formulated in A. S. Fotheringham (1983a), is an answer to this critique, where the dependency between destinations are modelled with a centrality variable $C_{i,j,t}$.

$$C_{i,j,t} = \sum_{k=1;k\neq i;k\neq j} \frac{P_{k,t}}{D_{j,k,t}}$$

Note that we interpreted P_j as a measure of the attractiveness (mass) of the destination j. The centrality measure specified as above then represents the attractiveness of other destinations weighted with the distance to j. A high value of $C_{i,j,t}$ consequently means that destination j is located in a cluster of attractive competing destinations.

2.2 The origin-specific gravity model estimated using Geographically Weighted Regression

Instead of estimating the gravity model on the entire data set, it is quite common to estimate separate models for each origin (Fotheringham, Charlton, and Brunsdon 1998), sometimes referred to as origin-specific parameterisation. As argued in A. S. Fotheringham (1983b), there are good reasons to expect spatial variation in parameter estimates in the setting of competing destination: the distance decay parameter of relatively accessible origins should, in general, be less negative than that of inaccessible origins.

One potential weakness with this approach is that the data set used for each origin-specific model becomes defined by administrative boundaries. In our case this is problematic since the cut-off scheme between long and short-distance migration is the same for the entire sample. Consequently, the origin-specific parameterisation is problematic in certain origins as the reduction in sample size amplifies the effect of influential observations.

Our solution is to define an optimal neighbourhood around each origin using Geographically Weighted Regression. GWR is a non-parametric technique is primarily used for exploring spatial non-stationarity. This technique thus allows for investigating whether the strength of a (statistical) relationship varies between locations in a spatial data set. Essentially, GWR measures the relationships around the geographical location of an observation in the data set, where each set of parameters is estimated by weighted least square. GWR is distinguished by weights being based on distance between regression points.

The final model with three explanatory variables and origin-specific parameters can be written as follows:

$$
\begin{aligned}
M_{i,j,t} &= O_{i,t} A_{i,t} P_{j,t}^{\beta_P(\mathbf{u}_i)} D_j^{\beta_D(\mathbf{u}_i)} C_{i,j,t}^{\beta_C(\mathbf{u}_i)} \\
O_{i,t} &= \sum_j M_{i,j,t} \\
A_{i,t} &= \left(\sum_j P_{j,t}^{\beta_P(\mathbf{u}_i)} D_{i,j}^{\beta_D(\mathbf{u}_i)} C_{i,j,t}^{\beta_C(\mathbf{u}_i)} \right)^{-1}
\end{aligned}
$$

The origin-constraint is written as $O_{i,t}$, while the balancing factior $A_{i,t}$ is included to ensure that the model reproduces flows originating at i. Formulated as a Poisson regression model, we operationalise the origin constraint using region-specific effects γ_i and time-specific effects τ_t.

$$
\lambda_{i,j,t}(\mathbf{u}_i) = \exp\left(\gamma_i \tau_t + \beta_P(\mathbf{u}_i) P_{j,t} + \beta_D(\mathbf{u}_i) D_{i,j,t} + \beta_C(\mathbf{u}_i) C_{i,j,t}\right)
$$

2.3 Fixed effects and Geographically Weighted Poisson Regression

One main reason behind the popularity of the gravity model is that it is easily log-linearised, whereby the parameters can be estimated using OLS. As widely recognised in the literature, the log-normal specification is problematic. Firstly, the log-normal model is unable to deal with zero-valued migration flows, since the logarithm of zero is undefined (Burger, Van Oort, and Linders 2009). Secondly, the logarithmic transformation of M_{ij} has an effect on the estimation process, whereby the antilogarithms of the estimates $\log \hat{M}_{ij}$ tend to be biased, leading to under-predicting of large flows and total flows (Flowerdew and Aitkin 1982; Silva and Tenreyro 2006). A solution to both of these problems is to estimate a regression model based on discrete distributions such as Poisson.

Although it is well established that the fixed effects of a Poisson regression model ensure that constraints hold, it is not immediately obvious that the same holds true for a GWPR model. In this section we show that the origin-constraints follow naturally from the maximum likelihood estimation, also with the GWPR.

Consider the migration flow between municipality i and j, $M_{i,j}$ (the time index is supressed for now). A gravity model with k explanatory variables $x_{k,i,j}$ can be represented as a Poisson regression model as follows (Nakaya et al. 2005):

$$M_{i,j} \sim \text{Poisson}\left[\exp\left(\sum_k \beta_k x_{k,i,j}\right)\right]$$

In this formulation β_k are estimated parameters and Poisson $[\lambda]$ indicates a Poisson distribution with mean λ. We can then adopt the more common formulation used by Flowerdew and Aitkin (1982):

$$\lambda_{i,j} = \exp\left(\sum_k \beta_k x_{k,i,j}\right)$$

Following this formulation, we see that an origin constraint implies that the estimated flows summed over destinations must be the same as the observed flows summed over destinations, for all origins.

$$\sum_j \hat{\lambda}_{i,j} = \sum_j M_{i,j}$$

Constraints (single or double) can be operationalised by specifying some of the $x_{k,i,j}$ explanatory variables as dummy variables corresponding to fixed effects.

The GWPR formulation of this model is similar, but now we let the parameters vary according to the geographical location of the origin, $\mathbf{u}_i = u_{x,i}, u_{y,i}$. We can write the geographically weighted Poisson Regression as

$$\lambda_{i,j}(\mathbf{u}_i) = \exp\left(\sum_k \beta_k(\mathbf{u}_i) x_{k,i,j}\right)$$

and the origin-constraint can be written as

$$\sum_j \hat{\lambda}_{i,j}(\mathbf{u}_i) = \sum_j M_{i,j}$$

The model is estimated by maximising a log-likelihood function written as

$$\ell = \sum_i \sum_j \left[M_{i,j} \sum_k x_{k,i,j}\beta_k(\mathbf{u}_i) - \exp\left(\sum_k x_{k,i,j}\beta_k(\mathbf{u}_i)\right)\right]$$

The derivative of the log-likelihood with respect to the vth structural parameter is

$$\frac{\partial \ell}{\partial \beta_v(\mathbf{u}_i)} = \sum_i \sum_j x_{k,i,j}\left[M_{i,j} - \exp\left(\sum_k x_{k,i,j}\beta_k(\mathbf{u}_i)\right)\right]$$

In order to represent fixed effects, we let the rth explanatory variable be a dummy variable equal to one if $i = I$ and zero otherwise. The derivative of the log likelihood function with respect to the rth structural parameter is

$$\frac{\partial \ell}{\partial \beta_r(\mathbf{u}_I)} = \sum_j \left[M_{I,j} - \exp\left(\sum_k x_{k,I,j}\beta_k(\mathbf{u}_I)\right)\right]$$

The maximum likelihood obviously implies that all derivatives are zero. By plugging in the maximum likelihood estimates of the parameters $\hat{\beta}_k(\mathbf{u}_I)$, we see that the origin-constraint holds for region I

$$\sum_j M_{I,j} = \sum_j \exp\left(\sum_k x_{k,I,j}\hat{\beta}_k(\mathbf{u}_I)\right) = \sum_j \hat{\lambda}_{i,j}(\mathbf{u}_I)$$

It is easy to see that this generalises for all other i. In a panel data setting, we see that a dummy variable of region interacted with dummy variable representing each year operationalises origin constraints for each time period. Note that this likelihood function is somewhat simpler than the local likelihood presented in Nakaya et al. (2005); the same argument applies to that case too.

3 Case study

The Netherlands is a small and densely populated country with good transport infrastructure. The population is concentrated within the Randstad area which covers the urban agglommerations of the four big cities (Amsterdam, Utrecht, den Haag, Rotterdam). This area covers about 20 % of the total surface area of the Netherlands, yet it account for a very large share of total migration flows. As we see in the left panel in Figure @ref(fig: descriptive-data), which shows all destinations' share of total migration flows, the migration is primarily concentrated in the large cities.

If we zoom in on short-distance moves, there is an additional concentration in the Randstad area. In this paper, we define short as moves within 35 km. A previous study found that this corresponded to a cutoff where moves were predominantly motivated by house or neighbourhood (Feijten and Visser 2005). The right panel in Figure @ref(fig: descriptive-data), showing the share of short-distance flows relative to total flows for each origin, shows that short-distance migration is particularly frequent in areas around large cities and less frequent in border regions or in the Frisian islands.

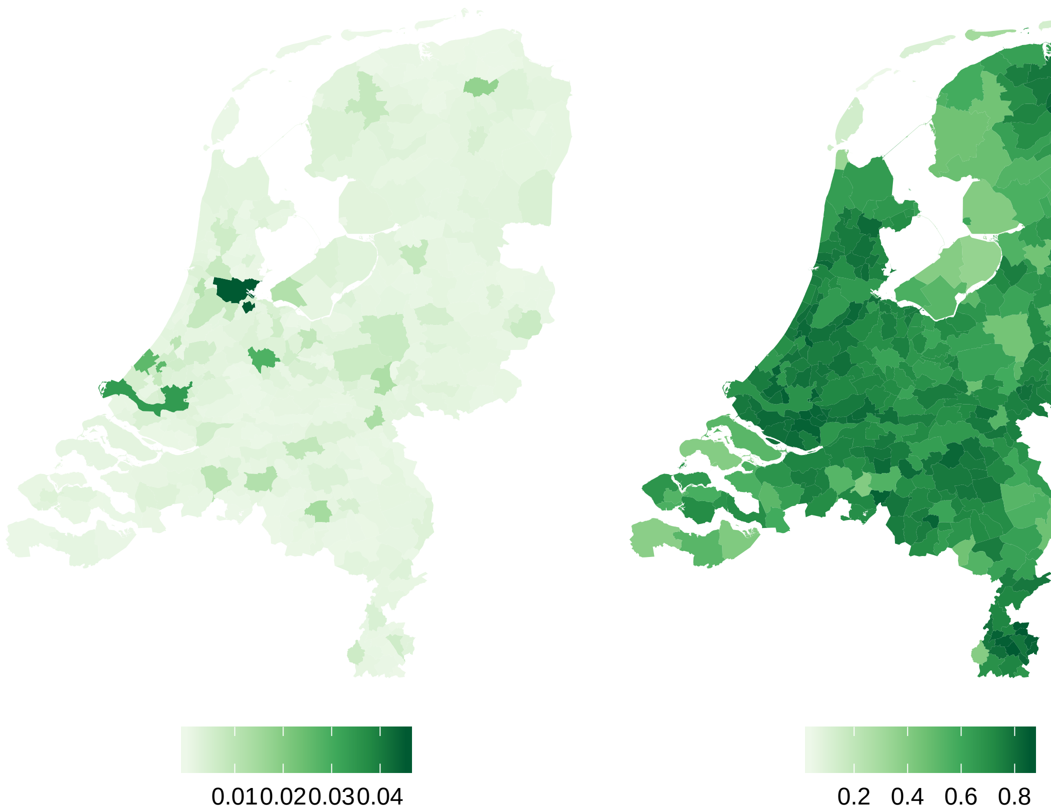

Figure 1: Destination's share of total migration flows (left) and short-distance flows relative to total flows by origin

3.1 Data

The data used to estimated the model are all publicly available from Statistics Netherlands. Data for the dependent variable are origin-destination migration flows between all Dutch municipalities covering the time period 2012 to 2018, and are calculated on the basis of registered individual address changes. In order to account for border changes, we use the municipality definition of 2018 and adjust data for all other years accordingly. Data on population is retrieved and processed the same way.

As was clear from the previous section, he method of GWR relies on distances between origins and destinations. Yet, migration flows are registered by municipalities, meaning we have no information about the precise location of the origin or destination of individual moves. As is common in the literature(Boyle and Flowerdew 1997), we define origins and destinations by the population weighted centroid of the respective municipality, meaning that all outflows from one origin emanate from the same coordinates. This consequently means that distance as used in this paper refer to the distances between the the origin and destination centroids.

Distance tends to be an important predictor of internal migration flows (Champion et al. 1998). In this paper we use the shortest distance on the road network instead of the somewhat simpler Euclidian distance, partly because some origins and destinations are separated by a body of water. Earlier work within the context of PEARL also suggested that a model with road distance fit the data better than models with Euclidian distance (Loke and de Jong 2012). The shortest road distance between all origin-destination pairs is obtained using the Open Source Routing Engine api (Luxen and Vetter 2011).

4 Estimation strategy

In this section we describe the computational strategy of the model calibration.

4.1 Computational issues

The model is calibrated using customised functions from the **GWModel** package in R (Gollini et al. 2015). The package offerings functions for estimating different types of GWR models, including generalised GWR models with Poisson and Binomial options. The implementation of the GWR Poisson model closely follows the GWPR model as discussed in Nakaya et al. (2005), where model parameters are estimated using a local maximum likelihood.

In order to estimate panel data models with fixed effects with the **GWModel** package, we needed to modify the software to avoid negative degrees of freedoms. Since the usage of the discontinous kernels frequently results in a zero weight to a number of observations, we need to ensure that only relevant fixed effects are included. This can be illustrated with a following three region example, where the weighted X matrix with fixed effects, seen from origin number 1, is represented as follows:

$$\begin{bmatrix} i & j & X & FE1 & FE2 & FE3 \\ 1 & 2 & x_{1,2} & 1 & 0 & 0 \\ 1 & 3 & x_{1,3} & 1 & 0 & 0 \\ 2 & 1 & w_{1,2}x_{2,1} & 0 & w_{1,2} & 0 \\ 2 & 3 & w_{1,2}x_{2,3} & 0 & w_{1,2} & 0 \\ 3 & 1 & w_{1,3}x_{3,1} & 0 & 0 & w_{1,3} \\ 3 & 2 & w_{1,3}x_{3,2} & 0 & 0 & w_{1,3} \end{bmatrix}$$

Consider a weighting scheme that gives zero weight to origin number 3 ($w_{1,3} = 0$). This obviously causes the lowest two rows to contain only zeros, but it also means that all entries in the column of the fixed effect of origin 3 become zero. In fact, for the bandwidths considered in this paper, the weighted X matrix ended up consisting primarily of zeroes. By leaving out the all-zero rows and columns, thereby exploiting sparsity in the X matrix, we avoid issues related to singularity of the X matrix and we reduce the computational time of the model dramatically.

4.2 Kernel selection using time series cross validation

In order to complete the fitting of the model we need to determine the weighting scheme$w_{i,j}$, which is done by specifying the type and size of the spatial kernel. Initial tests suggested that estimated parameters did not vary much between different kernel types. Out of convenience and since it is frequently used in the literature, we have chosen the bisquare kernel which can be written as

$$w_{i,j} = \begin{cases} (1 - (D_{i,j}/b)^2)^2 & \text{if } |D_{i,j}| < b \\ 0 & \text{otherwise} \end{cases}$$

Since the municipalities in the Netherlands vary quite substantially in size, we choose an adaptive kernel, meaning that the size of the kernel is determined by number of neighbouring regression points rather than distance. In order to find the optimal bandwith we follow the literature and determine it on the basis of a cross-validation exercise, thereby minimising the simulated out-of-sample forecasting errors.

However, the panel data setting here requires some modifications. As in the setting of a time series, panel data also suffers from strong serial autocorrelation with individual observations. In a cross validation exercise this could invalidate the assumption of independence between test and training sample. The common workaround for this in the time series literature is the so-called evaluation over a rolling origin, where the training data is successively extended in k iterations (Tashman 2000).

This procedure is illustrated in Figure 2. The movement along the x-axis shows how the origins "roll" forward in time, where the model is estimated on the training data (crosses) and evaluated on a hold-out-sample (dot). Since our data set spans the period 2012 to 2017 we can repeat this procedure 5 times, meaning we carry out a 5-fold cross validation. For example, the training set in fold 1 contains observations from year 2012 and its corresponding test set are observations from year 2013, training set in fold 2 contains observations from years

2012 and 2013 and its corresponding test set 2014 and so forth. Forecast perfomance is then evaluated by averaging the forecast errors over different folds.

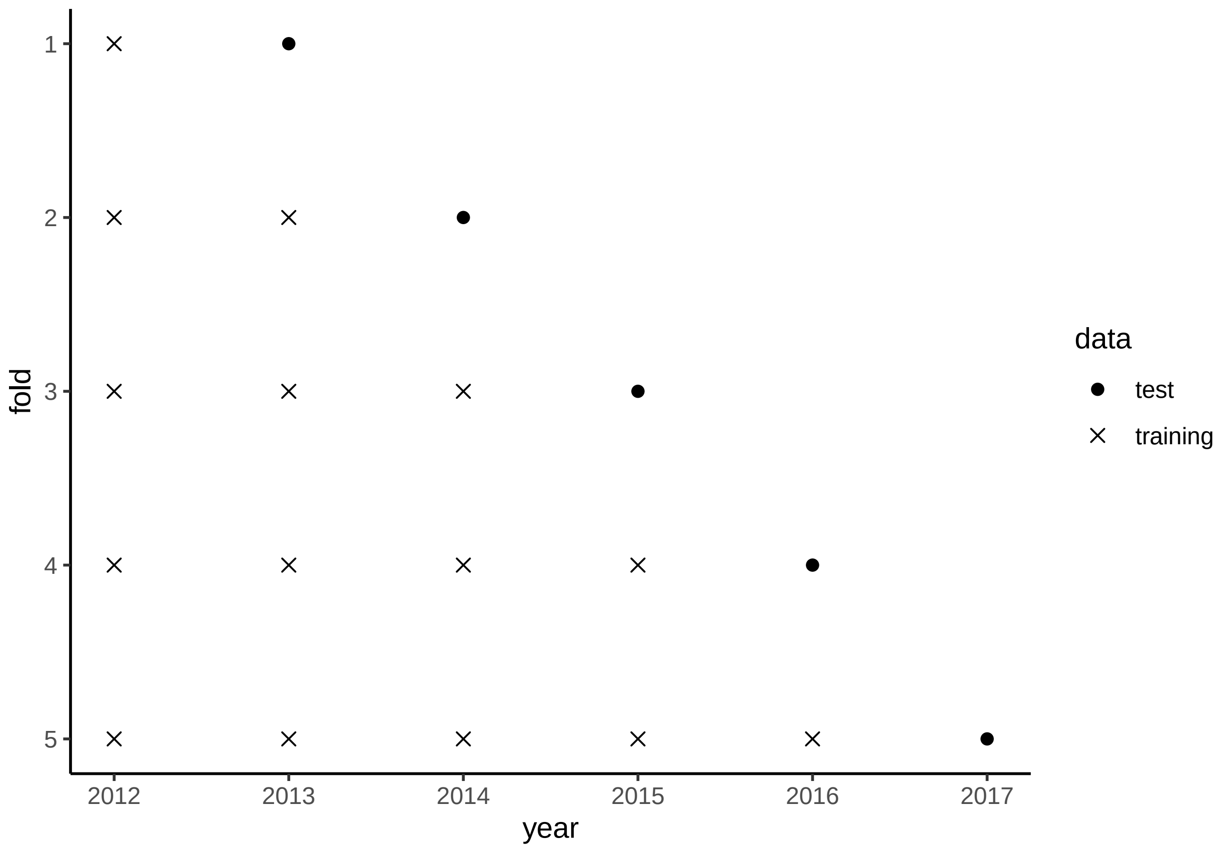

Figure 2: Time series cross validation

In the panel data setting we also need to adjust the bandwith according to the number of observations in the training set. Since the training set in fold number 2 containts twice the number of observations in fold 1, it is logical that an optimal bandwidth for fold 2 is twice the size of the bandwidth of fold number 1. If we let b be the bandwidth in year 2018, then the relevant bandwith in year t can be found using the following formula.

$$B_t = \frac{t - 2011}{7} b$$

The figure below shows the mean Root Mean Square Error (RMSE) and the standard error across folds, for bandwith in the interval between 1150 and 1200. The RMSE for bandwidths between 1180 and 1184 are essentially the same; as such we choose the model with bandiwth 1180. According to the formula above this results in a bandwith of 169 in the model with a single year.

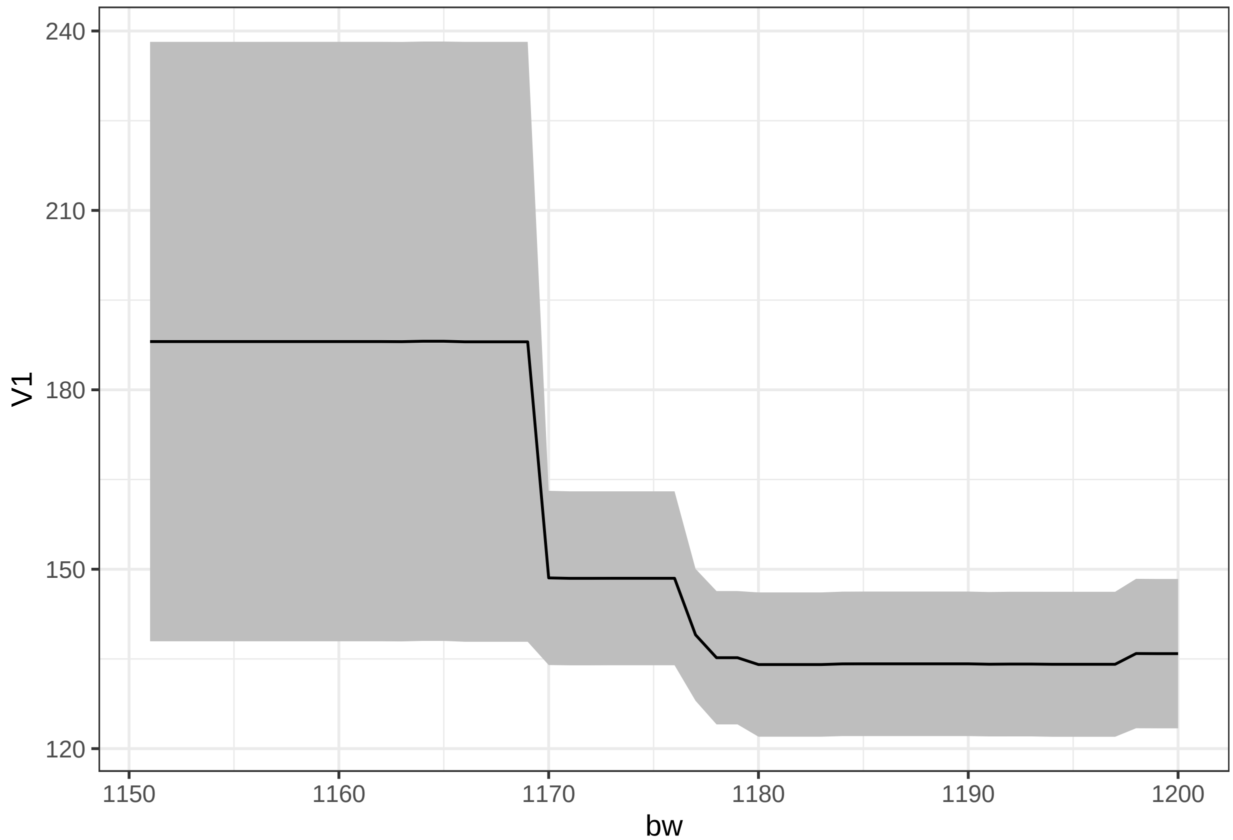

Figure 3: Mean out-of-sample RMSE and standard error

5 Results

In this section we present the estimated model. First we focus on the calibrated parameters, next we

5.1 Parameters estimates

5.2 Prediction

Conclusion

Boyle, Paul J, and Robin Flowerdew. 1997. "Improving Distance Estimates Between Areal Units in Migration Models." *Geographical Analysis* 29 (2). Wiley Online Library: 93–107.

Brunsdon, Chris, A Stewart Fotheringham, and Martin E Charlton. 1996. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28 (4). Wiley Online Library: 281–98.

Burger, Martijn, Frank Van Oort, and Gert-Jan Linders. 2009. "On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-Inflated Estimation." *Spatial Economic Analysis* 4 (2). Taylor & Francis Group: 167–90.

Champion, Tony, Stewart Fotheringham, Philip Rees, Paul Boyle, and John Stillwell. 1998. "The Determinants of Migration Flows in England: A Review of Existing Data and Evidence." *A Report Prepared for the Department of the Environment, Transport and the Regions. Department of Geography, University of Leeds.*

http://www.geog.leeds.ac.uk/publications/DeterminantsOfMigration/report.pdf.

Davies, Richard B, and Clifford M Guy. 1987. "The Statistical Modeling of Flow Data When the Poisson Assumption Is Violated." *Geographical Analysis* 19 (4). Wiley Online Library: 300–314.

Feijten, P, and P Visser. 2005. "Binnenlandse Migratie: Verhuismotieven En Verhuisafstand." *Bevolkingstrends* 53 (2): 75–81.

Flowerdew, Robin, and Murray Aitkin. 1982. "A Method of Fitting the Gravity Model Based on the Poisson Distribution." *Journal of Regional Science* 22 (2). Wiley Online Library: 191–202.

Fotheringham, A Stewart. 1983a. "A New Set of Spatial-Interaction Models: The Theory of Competing Destinations." *Environment and Planning A: Economy and Space* 15 (1). Sage Publications Sage UK: London, England: 15–36.

———. 1983b. "Some Theoretical Aspects of Destination Choice and Their Relevance to Production-Constrained Gravity Models." *Environment and Planning A* 15 (8). SAGE Publications Sage UK: London, England: 1121–32.

Fotheringham, A Stewart, and Morton E O'Kelly. 1989. *Spatial Interaction Models: Formulations and Applications*. Vol. 1. Kluwer Academic Publishers Dordrecht.

Fotheringham, A Stewart, Martin E Charlton, and Chris Brunsdon. 1998. "Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis." *Environment and Planning A* 30 (11). Sage Publications Sage UK: London, England: 1905–27.

Fotheringham, A Stewart, Martin Charlton, and Chris Brunsdon. 1997. "Measuring Spatial Variations in Relationships with Geographically Weighted Regression." In *Recent Developments in Spatial Analysis*, 60–82. Springer.

Gollini, Isabella, Binbin Lu, Martin Charlton, Christopher Brunsdon, and Paul Harris. 2015. "GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models." *Journal of Statistical Software* 63 (17): 1–50. http://www.jstatsoft.org/v63/i17/.

Kalogirou, Stamatis. 2016. "Destination Choice of Athenians: An Application of Geographically Weighted Versions of Standard and Zero Inflated P Oisson Spatial Interaction Models." *Geographical Analysis* 48 (2). Wiley Online Library: 191–230.

Loke, Rob, and Andries de Jong. 2012. "Verbeterd schattingsmodel voor korte afstandsmigratie in het regionaal prognosemodel PEARL[Improved estimation model for short-distance migration in PEARL, the regional prognoses model, in Dutch]." Working Paper. PBL Netherlands Environmental Assessment Agency.

Luxen, Dennis, and Christian Vetter. 2011. "Real-Time Routing with Openstreetmap Data." In *Proceedings of the 19th Acm Sigspatial International Conference on Advances in Geographic Information Systems*, 513–16. GIS '11. New York, NY, USA: ACM. doi:10.1145/2093973.2094062.

Nakaya, Tomoki. 2001. "Local Spatial Interaction Modelling Based on the Geographically Weighted Regression Approach." *GeoJournal* 53 (4). JSTOR: 347–58.

Nakaya, Tomoki, Alexander S Fotheringham, Chris Brunsdon, and Martin Charlton. 2005. "Geographically Weighted Poisson Regression for Disease Association Mapping." *Statistics in Medicine* 24 (17). Wiley Online Library: 2695–2717.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics* 88 (4). MIT Press: 641–58.

Tashman, Leonard J. 2000. "Out-of-sample tests of forecasting accuracy: an analysis and review." *International Journal of Forecasting* 16 (4): 437–50. doi:https://doi.org/10.1016/S0169-2070(00)00065-0.

Wilson, Alan Geoffrey. 1971. "A Family of Spatial Interaction Models, and Associated Developments." *Environment and Planning A* 3 (1). Sage Publications Sage UK: London, England: 1–32.