Planbureau voor de Leefomgeving

Internal migration in the PBL/CBS Regional Population Projections

Trond Grytli Husby, Andries de Jong

Netherlands Environmental Assesment Agency (PBL)

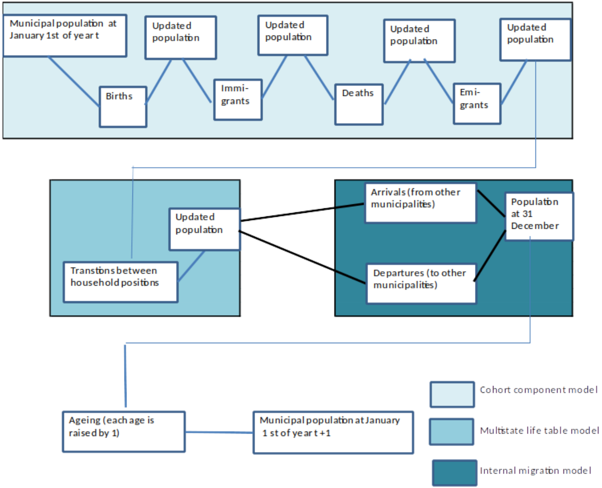
trond.husby@pbl.nl

Amsterdam March 12, 2020

The regional population projections by PBL and CBS

- Projections of population, households and demographic events in Dutch municipalities until 2050
- Carried out every three years: previous edition was in 2019
- Regional projections are made consistent with the national projections (CBS)
- The projections are made with a combination of a cohort-component model (Projecting population Events at Regional Level), multistate life table model and internal migration module
- The projection model is called PEARL: projecting population events at a regional level

Structure of PEARL



Internal migration: arrivals

- Long-distance migration determined on the basis of observed long-distance patterns
- More than half of migration flows are short-distance relocations (<35 km). In PEARL this is represented with a *constrained gravity model*
- The (initial) distribution of flows from municipality i to j is modelled as

$$M_{i,j} = O_i A_i \prod^{k} X_{k,i,j}^{\hat{\beta}_{k,i}} D_{i,j}^{\hat{\gamma}_i}, \quad O_i = \sum_j M_{i,j}, \quad A_i = \frac{1}{\sum_j \prod^{k} X_{k,i,j}^{\hat{\beta}_{k,i}} D_{i,j}^{\hat{\gamma}_i}} \tag{1}$$

- O_i is determined in a previous step in PEARL
- $\hat{\beta}_{k,i}$ and $\hat{\gamma}_i$ are estimated origin-specific parameters

Research Question

- Earlier work focused on local modelling of origin-constrained gravity with *geographically weighted regression (GWR)*
- Empirical strategy: take logs of both sides, rewrite, and estimate with OLS
- **Which specification minimises *out-of-sample* prediction errors? We compare results from a count model with OLS**
- Constraints are usually explicitly modelled, but with certain count models they can be captured by fixed effects. **Does this also apply if they are estimated using GWR?**

Origin-constrained gravity using Poisson regression

$$E(M_{i,j}) = \lambda_{i,j} = \exp(\beta_{0,i} + \sum_{k=1}^{K} \beta_k \log X_{k,i,j} + \gamma \log D_{i,j})$$

- Maximum likelihood estimation with Iteratively Reweighted Least Squares (IRLS)
- Fixed effects $\beta_{0,i}$ ensure that the origin constraints hold: $\sum_j \lambda_{i,j} = \sum_j M_{i,j}$

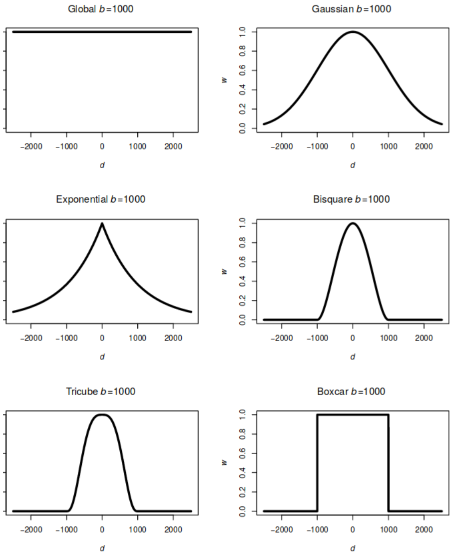
Geographically weighted Poisson regression: origin-specific and origin-constrained

$$\lambda_{i,j} = \exp(\beta_{0,i} + \sum_{k=1}^{K} \beta_k(\mathbf{u_i}) \log X_{k,i,j} + \gamma(\mathbf{u_i}) \log D_{i,j})$$

- Origins i are represented by the population-weighted centroids of each municipality
- The kth parameter for location i, $\beta_k(\mathbf{u}_i)$ is a function of the coordinates $\mathbf{u}_i$
- $\beta_{0,i}$ is the (unweighted) fixed effects of i: origin constraints hold!
- Estimation with local IRLS (Nakaya 2001; Nakaya et al. 2005), using a customised version of the *GWmodel* package in *R*: for estimation we make use of sparsity and the network structure of the data

Geographically weighted Poisson regression

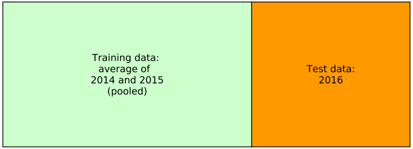
- Weighted regression for each origin *i* with distance-based weights according to a spatial kernel
- This creates a (weighted) neighbourhood around each origin
- Bandwidth is either distance or number of neighbouring points
- Model fitting completed with selection of kernel type and bandwith using cross validation. As a bonus, this allows us to simulate *within-* and *out-of-sample* prediction error, and to investigate the interplay between the two



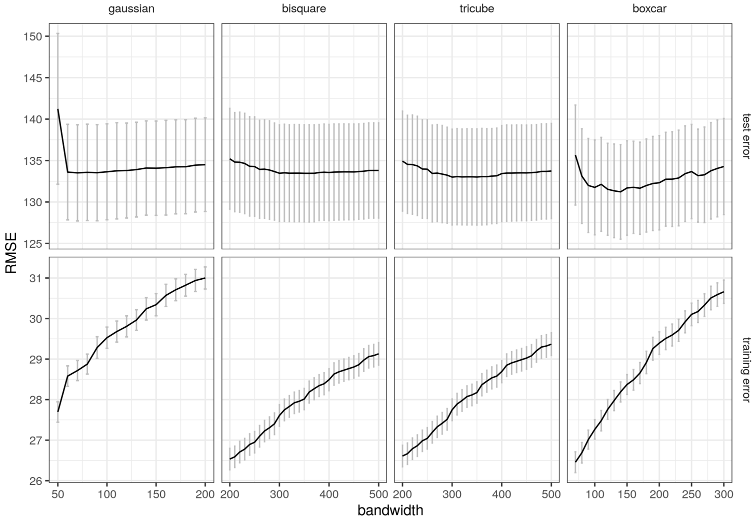
Data and variables

- Short distance: $M_{i,j}$ where Euclidian distance between i and j is less than 35 km
- Network with 390 vertices (municipalities) and 14558 edges (bilateral origin destination flows)
- 2016 used as hold out sample. Evaluation of prediction accuracy by plugging estimated parameters into Equation 1

Explanatory variables	
Population	P_j
Road distance (km)	$D_{i,j}$
Centrality	$C_{i,j} = \sum_{l=1;l\neq i;l\neq j}^{L} \frac{P_l}{D_{j,l}}$
Index of net new dwellings	$H_j = 1 + \frac{\text{Net growth supply}_j}{\text{Stock}_j}$

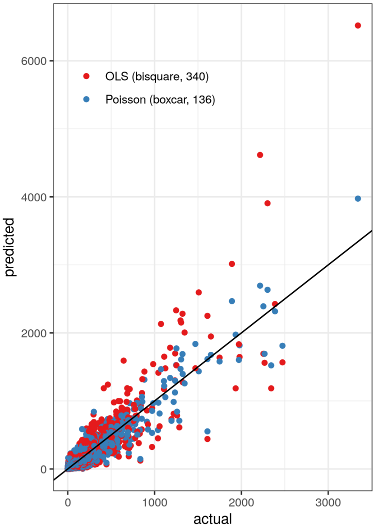


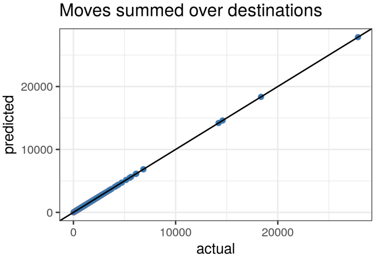
Finding the optimal kernel with cross validation

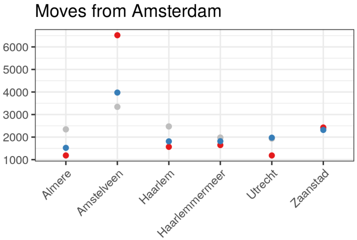


- 10-fold cross validation over a grid of bandwidth and type
- Boxcar, 136, has the lowest RMSE. Mean absolute error (MAE) gives similar results
- Bias variance trade off: test error is convex and training error decreases with bandwith

Out-of-sample predictions: actual versus predicted

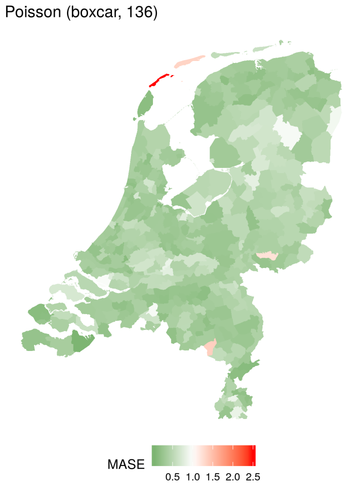






Out of sample predictions: overall model performance

	RMSE	MAE	$\overline{\text{SRMSE}}$	$\overline{\text{MASE}}$
OLS (bisquare, 340)	61.0427	14.0468	0.0304	0.4318
Poisson (boxcar, 136)	34.8364	11.7120	0.0260	0.4013

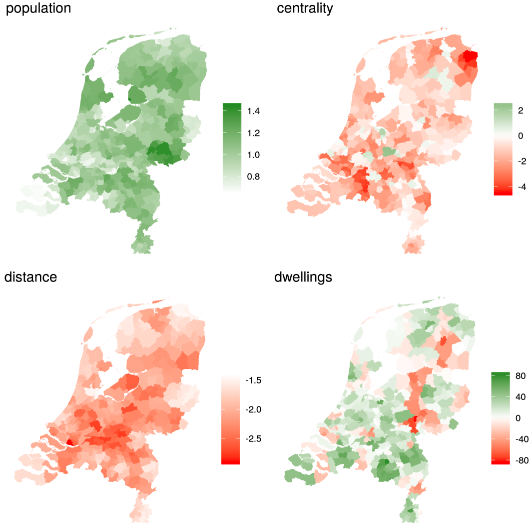


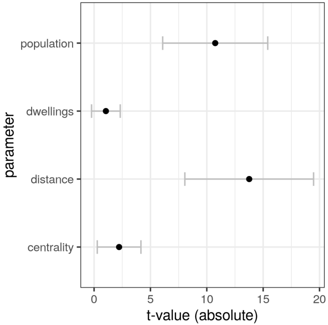
Where scaled errors are defined as:

$$\text{SRMSE}_i = \frac{RMSE_i}{\sum_j M_{i,j}}$$

$$\text{MASE}_i = \frac{1}{J}\sum_{j=1}^{J}\left|\frac{e_{i,j}}{\frac{1}{J}\sum_{j=1}^{J}|M_{i,j}-\bar{M}_i|}\right|$$

Estimated parameters





Conclusions and further work

- Estimated parameters exhibit spatial non-stationarity and cross validation suggests there is a bias variance trade off: need for a spatially explicit model
- Poisson estimator has smaller out of sample errors than the OLS estimator: reduces overprediction of large flows
- Fixed effects effectively ensure that origin constraints hold, also with GWR
- Future work 1: improvement of the dwelling variable
- Future work 2: is 35 km a good cut off for short- versus long-distance?
- Potential future work 3: evidence of overdispersion, information criteria suggest that Negative Binomial model is more appropriate. But predictions are terrible!
- Up for a challenge? Code and data online[1]

[1] https://github.com/trondhusby/gwr

References I

Davies, Richard B, and Clifford M Guy. 1987. The statistical modeling of flow data when the poisson assumption is violated. *Geographical Analysis* 19 (4): 300–314.

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*, 513–516. GIS '11. Chicago, Illinois: ACM. ISBN: 978-1-4503-1031-4.

Nakaya, Tomoki. 2001. Local spatial interaction modelling based on the geographically weighted regression approach. *GeoJournal* 53 (4): 347–358.

Nakaya, Tomoki, Alexander S Fotheringham, Chris Brunsdon, and Martin Charlton. 2005. Geographically weighted poisson regression for disease association mapping. *Statistics in medicine* 24 (17): 2695–2717.

Appendix: proof that Poisson ensures origin constraint (Davies and Guy 1987)

Log-likelihood of Poisson (ignoring the constant) is given by

$$\ell = \sum_i \sum_j [M_{i,j} x_{i,j}^T \beta - \exp(x_{i,j}^T \beta)]$$

where β is a vector of parameters. The derivative wrt to the u^{th} structural parameter is

$$\frac{\partial \ell}{\partial \beta_u} = \sum_i \sum_j x_{i,j,u}^T \left[M_{i,j} - \exp(x_{i,j}^T \beta) \right]$$

Let the r^{th} variable be a dummy variable if $i = I$. Then

$$\frac{\partial \ell}{\partial \beta_r} = \sum_j \left[M_{I,j} - \exp(x_{I,j}^T \beta) \right]$$

Maximum likelihood implies that derivatives are zero. Let $\hat{\beta}$ be the maximum likelihod estimates of β, then we have

$$\sum_j M_{I,j} = \sum_j \exp(x_{I,j}^T \hat{\beta}) = \hat{\lambda}_{I,j} \Rightarrow \sum_j M_{i,j} = \sum_j \hat{\lambda}_{i,j} \quad \forall i$$

Appendix: Origin constraints with fixed effects in GWR

Example: **WX** for three regions

i	j	X
1	2	$x_{1,2}$
1	3	$x_{1,3}$
2	1	$w_{1,2}x_{2,1}$
2	3	$w_{1,2}x_{2,3}$
3	1	$w_{1,3}x_{3,1}$
3	2	$w_{1,3}x_{3,2}$

i	j	X
1	2	$w_{2,1}x_{1,2}$
1	3	$w_{2,1}x_{1,3}$
2	1	$x_{2,1}$
2	3	$x_{2,3}$
3	1	$w_{2,3}x_{3,1}$
3	2	$w_{2,3}x_{3,2}$

i	j	X
1	2	$w_{3,1}x_{1,2}$
1	3	$w_{3,1}x_{1,3}$
2	1	$w_{3,2}x_{2,1}$
2	3	$w_{3,2}x_{2,3}$
3	1	$x_{3,1}$
3	2	$x_{3,2}$

Origin constraints with fixed effects in GWR

Zooming in on $i = 1$: $FE1$ becomes a non-weighted intercept, and it ensures that the origin constraint holds! However at a significant computational cost...

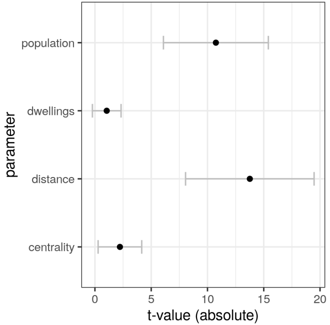
i	j	X	$FE1$	$FE2$	$FE3$
1	2	$x_{1,2}$	1	0	0
1	3	$x_{1,3}$	1	0	0
2	1	$w_{1,2}x_{2,1}$	0	$w_{1,2}$	0
2	3	$w_{1,2}x_{2,3}$	0	$w_{1,2}$	0
3	1	$w_{1,3}x_{3,1}$	0	0	$w_{1,3}$
3	2	$w_{1,3}x_{3,2}$	0	0	$w_{1,3}$

Origin constraints with fixed effects in GWR

... but we can exploit **sparsity** if the weight matrix is full of zeros. This is the case with non-continuous kernel types where bandwith is smaller than global. Setting $w_{1,3} = 0$

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	$w_{1,3}$

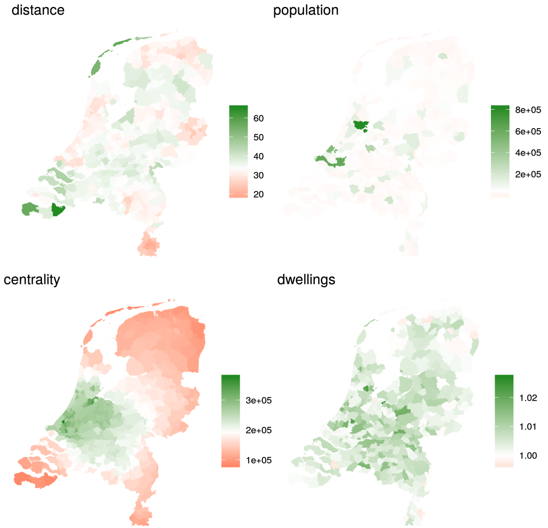
Appendix: What drives the prediction errors? Road distance a potential candidate



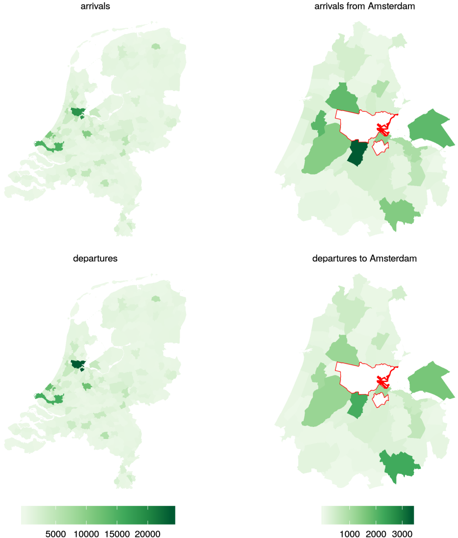
Distance is, in general, the most important variable...

...but road distance can be problematic. Lines in the figure are straight lines between two municipality centroids

Departures and arrivals

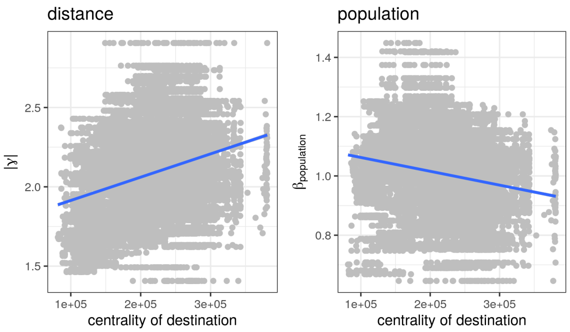


Explanatory variables, average value by destination



Planbureau voor de Leefomgeving

Core-periphery patterns of distance decay and population



- Flows to centrally located areas are characterised by high sensitivity to distance and low sensitivity to population
- The opposite for flows within peripheral areas
- Similar results were obtained for Japan (Nakaya 2001)