A gravity model of short-distance migration

Trond Grytli Husby

Netherlands Environmental Assessment Agency (PBL)

trond.husby@pbl.nl

Den Haag 21 March, 2019

Background

- Short-distance ($<=$ 35km) migration flows between municipalities in PEARL is (partly) determined by a set of estimated parameters
- These parameters allow destination choice to be influenced by population at destination and distance between origin and destination
- Earlier work (Andries and Rob Loke) focused on local modelling using geographically weighted regression using OLS and a block-shaped kernel
- The role of building construction and the use of Euclidian versus road distances have also been investigated
- Recommendations for future research from that study: try other kernel shapes; try other methods than OLS; find more flexible ways of determining best kernel shape and size
- Challenge accepted

Production-constrained gravity model

$$M_{ij} = O_i A_i \prod^{k} X_k^{\beta_k}$$

$$O_i = \sum_j M_{ij}$$

$$A_j = \frac{1}{\sum_j \prod^k X_k^{\beta_k}}$$

- i: municipality with population-weighted centroid u_i, v_i
- M_{ij}: n x (n-1) matrix, $\sum_{t=2014}^{2016} \frac{M_{tij|\text{dist}_{ij} \leq 35}}{3}$
- X_k: n x (n-1) matrices
- k = population, road distance, centrality, new dwellings
- new $\text{dwellings}_{2016} = \frac{\text{dwellings}_{2016}}{\text{dwellings}_{2014}}$

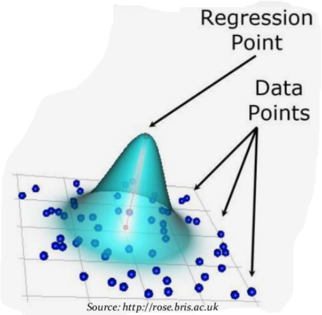
Empirical implementation

1. OLS: $\log M_{ij} + \frac{1}{n}\sum_j \log M_{ij} = \sum_k \beta_k(\log X_k + \frac{1}{n}\sum_j \log X_k) + \varepsilon_{ij}$
2. Poisson regression: $\lambda_{ij} = \exp(\alpha_i + \sum_k \beta_k \log X_k + \epsilon_{ij})$

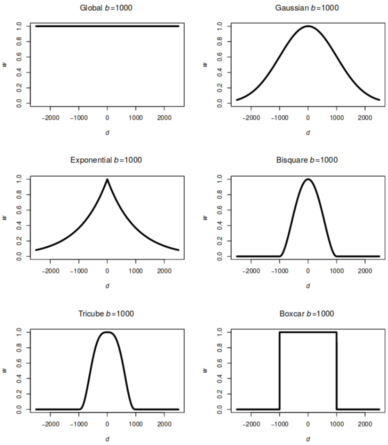
- 1. is a log-log, 2. is linear-log. Antilog transformation in 1. can lead to underestimation of large flows and total flows
- Zero-value M_{ij} in 1. are replaced, somewhat arbitrarily, with 0.1. The count model 2. handles zeros naturally
- Balancing factors O_i and A_i in 2. are captured by fixed effects α_i. Since $n = 390$ we need a big sample for 2. to work well
- 2. requires integerisation of M_{ij}

Geograhically weighed regression (GWR)

- Standard OLS: $y_i = \beta x_i + \varepsilon_i$
- GWR: $y_i = \beta(u_i, v_i)x_i + \varepsilon_i$
- Now $\beta(u_i, v_i)$ is a function of the location of i, where u_i and v_i are the coordinates of i
- To estimate $\beta(u_i, v_i)$ we include neighouring data points in a weighted regression model
- Step 1: select kernel type and bandwidth (distance or points)



Kernel and bandwidth (b) selection



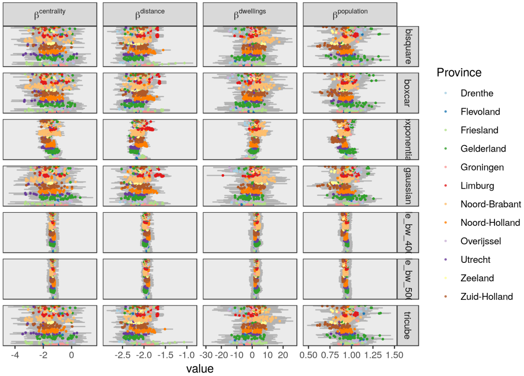
- Bandwidth (i.e. neighbourhood size) either distance or number of neighbouring data points. Since municipalities differ substantially in size we use the latter
- Optimal bandwidth for each kernel selected using leave-one-out cross-validation
- $Min_b \quad CV = \sum_i^n [y_i - \hat{y}_{\neq i}(b)]^2$
- $\hat{y}_{\neq i}(b)$ is the estimated value of y_i after location i is removed

Models and (within-sample) fitness measure

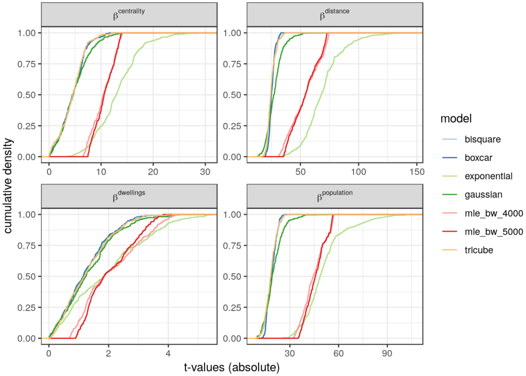
- Models
  - OLS: gaussian(99), bisquare(563), tricube(557), boxcar(328), exponential(400)
  - Poisson: gaussian, bandwiths 4000 and 5000
- $SRMSE = \frac{1}{\sum_{(ij)} M_{ij}} \sqrt{(M_{ij} - \hat{M}_{ij})^2}$
- Ensemble model: select the $\beta_{k,i}$ with the lowest $SRMSE_i$

Planbureau voor de Leefomgeving

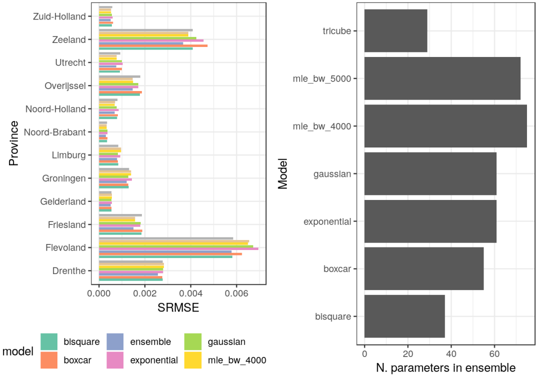
Estimated parameters and CI across municipalities and models



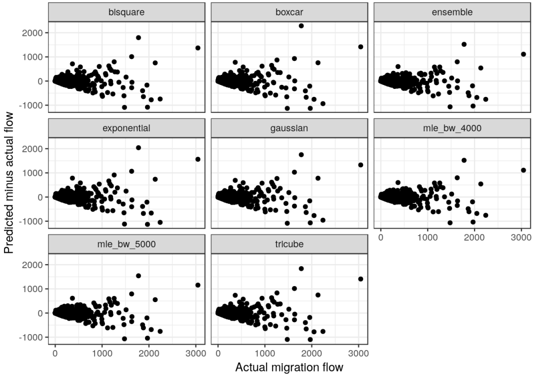
Significance of estimated parameters



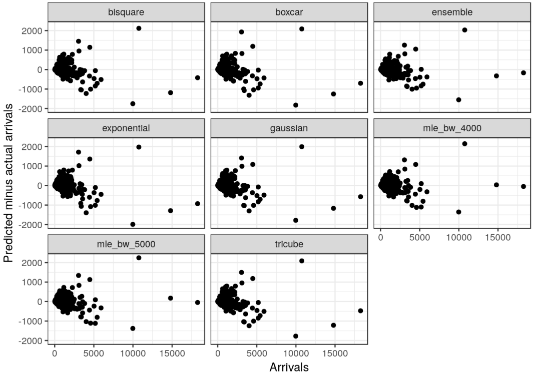
Province-level SRMSE and composition of ensemble model



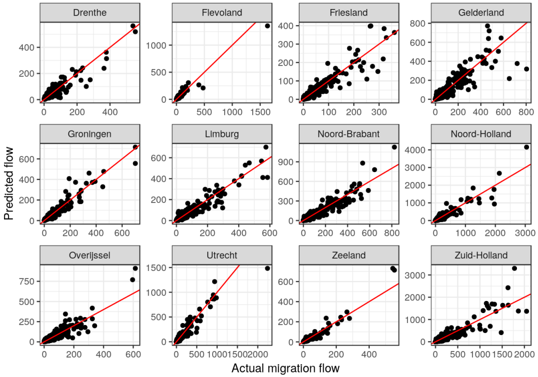
Prediction errors



Predicted minus actual arrivals

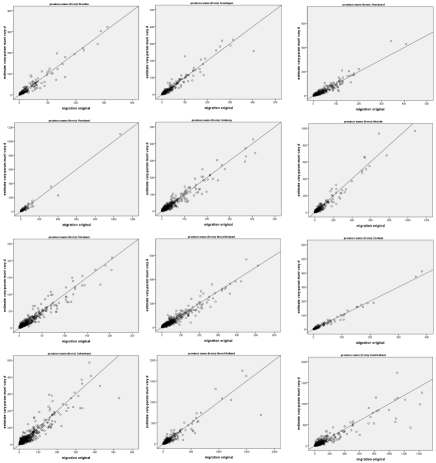


Prediction errors per province (ensemble model)

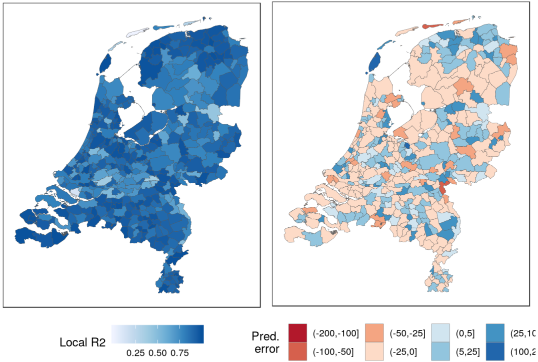


Planbureau voor de Leefomgeving

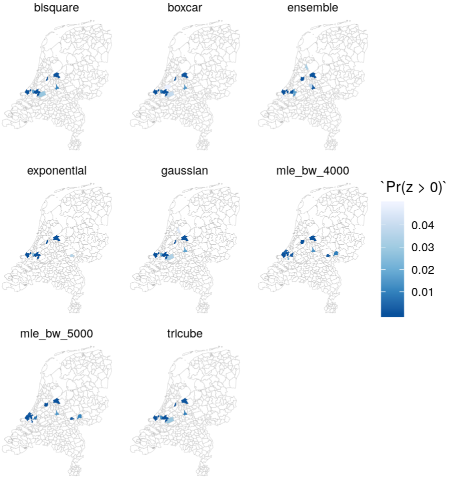
Prediction errors per province (2008)



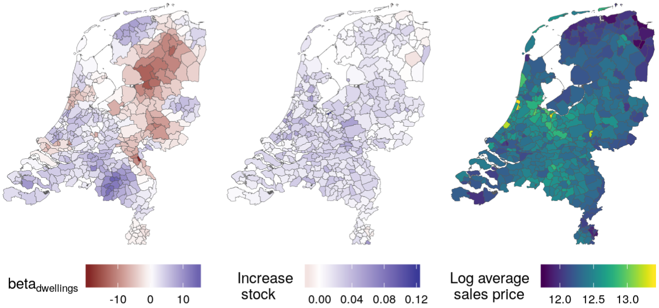
Local R2 and predicted minus actual arrivals (ensemble model)



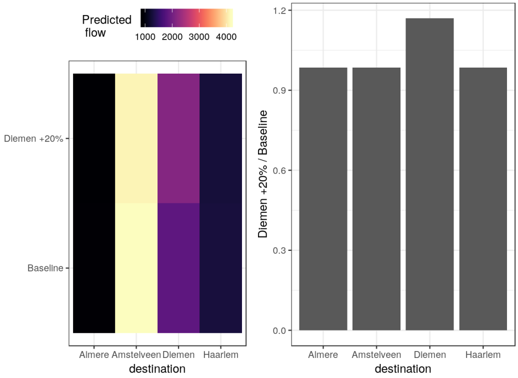
Spatial autocorrelation: Moran's I



The parameter $\beta_{dwellings}$



Experiment: migration from Amsterdam



- 20 % population increase in Diemen. What happens with outflows from Amsterdam?
  - Population increase in Diemen makes it more attractive
  - Other destinations become less attractive through the centrality effect
- $\beta_{population,Amsterdam} = 0.9315$, $\beta_{centrality,Amsterdam} = -1.2634$

Conclusion

- Other kernel types and bandwiths, and combining estimates from different models, help improving model fit
- Prediction errors larger than in 2008, but also higher migration frequency and fewer municipalities (thus larger flows)
- GWR for *predictive* analysis: the two Poisson models generally outperformed the OLS models, with a very large bandwidth and little local variation in parameters
- However, heterogeneity in for example $\beta_{population}$, suggest core-periphery patterns which should be captured by PEARL
- Suggestion 1: consider other model types for highly skewed count data and small samples? Or different regional aggregation levels?
- Suggestion 2: is 35 km the right cutoff point in this setting?
- Suggestion 2: selective migration by education. Potentially interesting?