Mobile phone data and gravity models

Victor Tuekam[1,2] Sebastian Wichert[1] Oliver Falck[1] Göran Kauermann[2]

[1]ifo Institute [2]Department of Statistics, LMU Munich

November 13, 2024

Outline

Motivation

Data

Model

Results & Interpretation

Motivation

- Mobile phone data as CDR are a record of human mobility
- The can help understand
  - Mass movements in general,
  - Daily journeys and dwells,
  - Dwells at particular places,
  - and much more
- Due to strict data protection and expensive procurement, mobile phone data is hard to come by in most countries.

In this talk, I want to discuss such data to identify flows and a modeling approach based on gravity models.

Data

The Data

- Two days of data (for the whole of Germany): 20th September 2020 and 21st September 2020.
- $\sim$ 2 TB of data.
- Signaling data generated by consumer or IoT devices.
- Anonymized with a 24h stability period.

The Data

Sample

User ID	Event Time	Longitude	Latitude
282359	2020-09-21 02:49:32	11.5723	48.1868
282359	2020-09-21 03:06:52	11.5723	48.1868
282359	2020-09-21 03:39:42	11.5723	48.1868
...	...	...	...
780159	2020-09-21 12:24:31	11.6443	48.1002
780159	2020-09-21 15:43:59	11.6344	48.17
780159	2020-09-21 15:44:24	11.6344	48.17
...	...	...	...
1221982	2020-09-21 17:37:24	11.5594	48.2196
1221982	2020-09-21 18:59:19	11.5587	48.2203
1221982	2020-09-21 18:59:28	11.5594	48.2196
...	...	...	...

- If a user first appears in the dataset connected to cell site A at time t we assume the user connected to mast A up until time t.
- If a user was connected to cell site A at time t and connected to cell site again A at time $t+\delta$ in the raw data, with no other connection information between these two time points, we assume that this user was connected to the same cell, A, at the time periods $\{t, t+1, \ldots, t+\delta\}$.
- If a user was connected to cell site A at time t and connected to a different cell site B at time $t+\delta$ we linearly reconstruct the users locations between times t and $t+\delta$, constrained to the voronoi points.

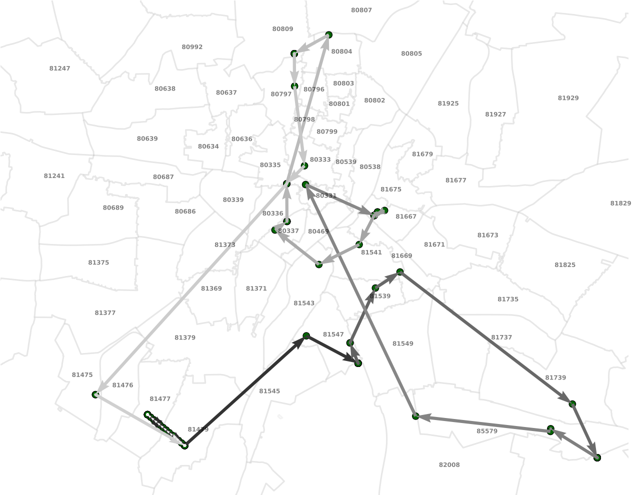


Figure: A trajectory

The Data

Deriving flows

- We now have to map these trajectory points to zipcodes.
- A problem is that these points are either cell locations or pseudo-cells.
- To do that we have to approximate cell coverages $\longrightarrow$ Voronoi partitions.

The Data

Voronoi partitions

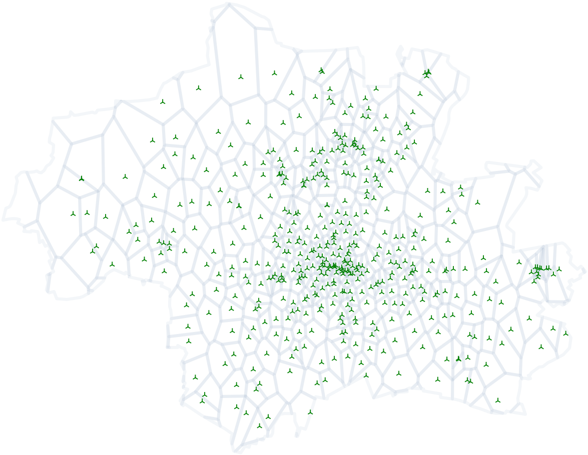


Figure: Voronoi partition of Munich

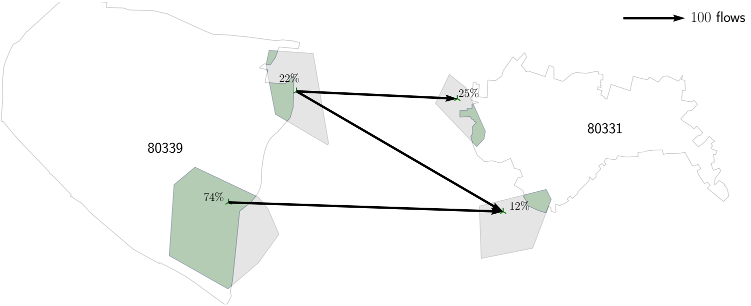
The Data

Flows

▸ Flow equation

- We are interested in flows of interest (spend at least 30 min at a place).
- By mapping every point to its Voronoi region, we can simply count the number of movements from one Voronoi region to the other.
- For flows between postcodes, use the intersection proportions with voronoi regions.

$$f_{80339 \to 80331} = \lceil 0.22 \times 0.25 \times 100 + 0.22 \times 0.12 \times 100 + 0.74 \times 0.12 \times 100 \rceil = 18$$



The Data

Flows

Origin	Destination	Time	Count
...	...	...	...
80807	80805	2020-09-21 08:30:00	94
80807	80809	2020-09-21 08:30:00	167
80807	80933	2020-09-21 08:30:00	0
80807	80935	2020-09-21 08:30:00	0
80807	80937	2020-09-21 08:30:00	24
80807	80939	2020-09-21 08:30:00	108
...	...	...	...

The Data

Flow map

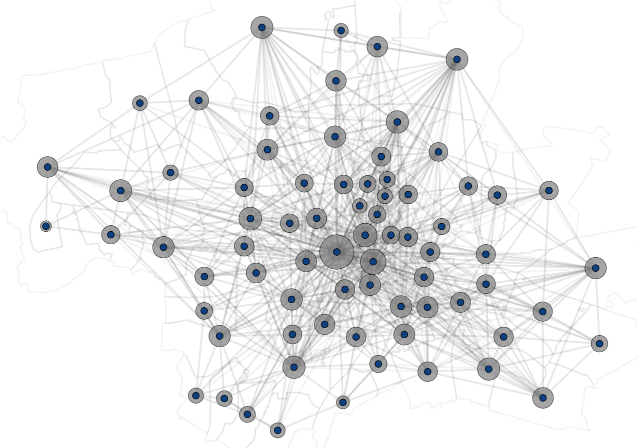


Figure: Flow Map

Model

- Gravity models are a popular class of models for analyzing flow data.
- Gravity equations can be derived using the random utility framework (Anderson, 2011; Beine et al., 2021).
- They often exclude important network effects, thus are necessarily incomplete Lebacher et al. (2020)

Data is commonly fitted to the model using

$$N_{jk}(t) = N_{jj}(t)\exp\{\eta(\boldsymbol{x}_k, t) - \eta(\boldsymbol{x}_j, t) - c_{jk}(t)\}\nu_{jk}(t) \tag{1}$$

such that $E[\nu_{jk}(t) \mid \boldsymbol{x}] = 1$. We also include random effects.
One may assume a Poisson distribution for the counts (Silva and Tenreyro, 2006). But

$$\frac{Var(N_{jk})}{\mathbb{E}[N_{jk}]} = 1$$

Model

A Zero-inflation

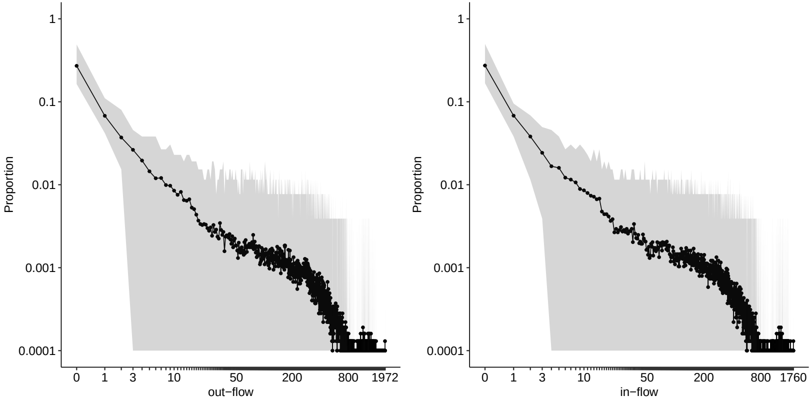


Figure: Distribution of in- and out-flows. Most of the data are zero.

We model the data using a zero-inflated Poisson model as follows:

$$
\begin{aligned}
N_{jk}(t) \mid \mathbf{x} &\overset{\text{i.i.d}}{\sim} 0 && \text{with probability } \pi_{jk}(t) \\
N_{jk}(t) \mid \mathbf{x} &\overset{\text{i.i.d}}{\sim} \text{Poisson}\big(\mu(t) \mid N_{jk}(t) > 0\big) && \text{with probability } 1 - \pi_{jk}(t)
\end{aligned}
$$

$$
\frac{Var(N_{jk})}{\mathbb{E}[N_{jk}]} = 1 + \mu(t)(1 - \gamma)
$$

for some $\gamma > 0$ such that the ratio stays positive. Therefore the model can handle over- and under-dispersion. The classical mixture model formulation can only deal with over dispersion (Tutz, 2011).

Model

Estimation

- Estimation can be performed using Wood et al. (2016), GAM fitting framework.
- This allows us to implement additional families by computing derivatives of the deviance.
- We use a log-link for the non-zero component and a log-log link for the zero component.

Results & Interpretation

Gravity model

Interpretation

▸ More covariate results

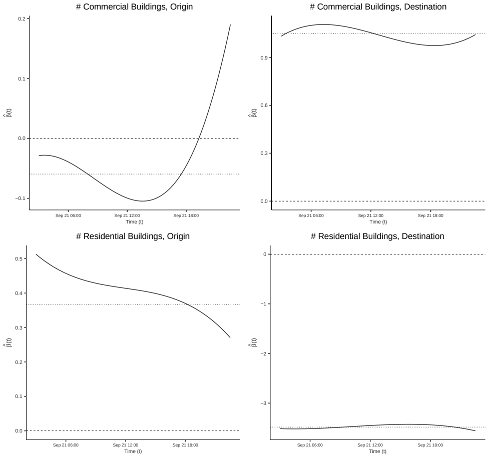
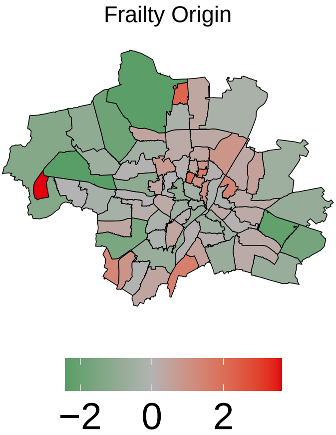


Figure: Results of exogenous statistics relating to some economic factors

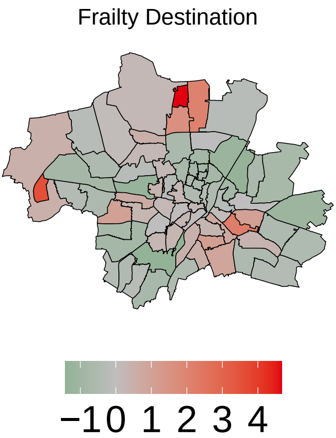
Gravity model

Interpretation



Gravity model

Interpretation



Thanks,

What are your questions?

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics*, 3(1):133–160.

Beine, M., Bertinelli, L., Cömertpay, R., Litina, A., and Maystadt, J.-F. (2021). A gravity analysis of refugee mobility using mobile phone data. *Journal of Development Economics*, 150:102618.

Lebacher, M., Thurner, P. W., and Kauermann, G. (2020). A dynamic separable network model with actor heterogeneity: An application to global weapons transfers. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(1):201–226.

Silva, J. M. C. S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge University Press.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

The Data

Median completeness on Monday, 21st

Letting p_o and p_d be some origin and destination postcodes respectively, and $V_o = \{v \mid v \text{ is a Voronoi region and } a(v \cap p_o) \neq 0\}$ and $V_d = \{v \mid v \text{ is a Voronoi region and } a(v \cap p_d) \neq 0\}$ respectively, where $a(\cdot)$ returns the area of a region, we compute the number of flows from p_o to p_d at time t as

Back

$$f_{p_o \to p_d}(t) = \sum_{v_o \in V_o} \sum_{v_d \in V_d} \left\lceil \frac{a(v_o \cap p_o)}{a(p_o)} \frac{a(v_d \cap p_d)}{a(p_d)} f_{v_o \to v_d}(t) \right\rceil \tag{2}$$

Gravity model

Interpretation

Back

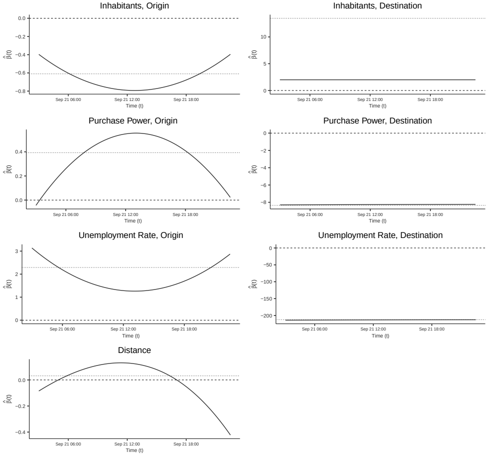


Figure: Results of exogenous statistics relating to some economic factors

Gravity model

Interpretation

◀ Back

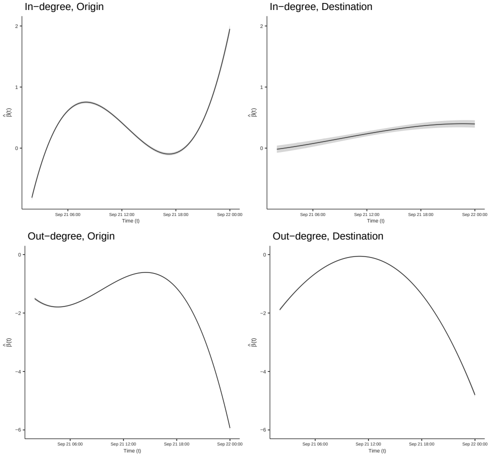


Figure: Results of endogenous network statistics