A Bayesian model for network flow data: an application to BikeMi trips

Mario Beraha

Joint work with Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi and Alessandra Guglielmi

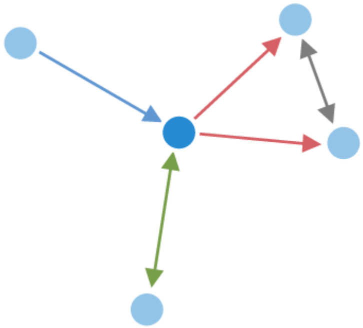
20 June 2019

POLITECNICO
MILANO 1863

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

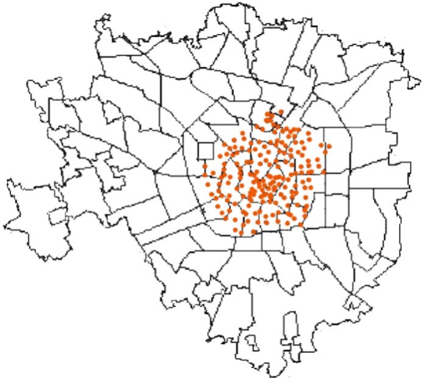
We are interested on making inference on a graph $\mathcal{G} = (V, E)$

- V: vertices
- $\mathrm{E} = \{(i, j)\}$ edges
- Quantity of interest Y_{ij}: flow on the edge (i, j).
  In our case, the flow is the number of trips
  from node i to node j



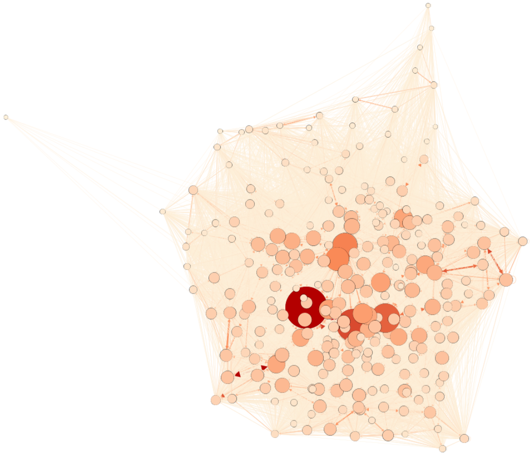
Dataset

- 263 BikeMi station
- **350093** trips between January 25th and March 6th 2016
- length of each trip



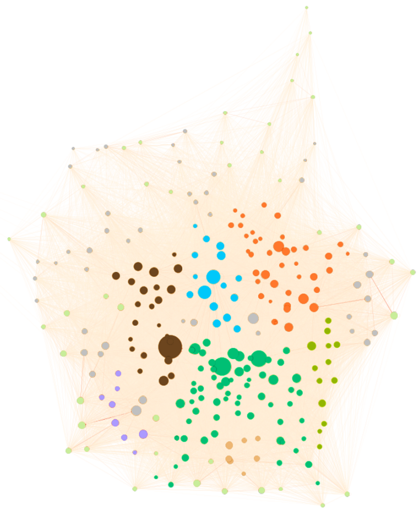
Dataset

- 263 BikeMi station
- **350093** trips between January 25th and March 6th 2016
- length of each trip
- $263^2 = \mathbf{69169}$ possible arcs!



Dataset

- 263 BikeMi station
- **350093** trips between January 25th and March 6th 2016
- length of each trip
- $263^2 = \mathbf{69169}$ possible arcs!
- DBSCAN clustering on the geolocation of the stations
- 67 nodes in the reduced network



The Bayesian Model

We consider the following Zero-inflated Poisson mixture regression model

$$p(Y_{ij} = y)|\theta, \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \begin{cases} \theta + (1-\theta)\text{PM}(0|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } y = 0 \quad i, j = 1, \ldots, 67 \\ (1-\theta)\text{PM}(Y_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

$$\text{PM}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \text{PM}\left((\mu_1, \ldots, \mu_K), (\lambda_1, \ldots, \lambda_K)\right) = \sum_{i=1}^{K} \lambda_i \text{Poi}(\mu_i)$$

$$\log \mu_{ijk} = \boldsymbol{\beta}_k \boldsymbol{x}_{ij}$$

For each arc, the covariates $\boldsymbol{x}_{ij}$ are $(1, S_i \times T_j, d_{ij})$

- S_i: "sourceness" of starting node (outer degree)
- T_j: "targetness" of destination node (inner degree)
- d_{ij}: geographical distance

The Bayesian Model

We consider the following Zero-inflated Poisson mixture regression model

$$p(Y_{ij} = y) | \theta, \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \begin{cases} \theta + (1-\theta)\text{PM}(0|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } y = 0 \quad i, j = 1, \dots, 67 \\ (1-\theta)\text{PM}(Y_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } y = 1, 2, 3, \dots \end{cases}$$

$$\text{PM}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \text{PM}\left((\mu_1, \dots, \mu_K), (\lambda_1, \dots, \lambda_K)\right) = \sum_{i=1}^{K} \lambda_i \text{Poi}(\mu_i)$$

$$\log \mu_{ijk} = \boldsymbol{\beta}_k \boldsymbol{x}_{ij}$$

For each arc, the covariates $\boldsymbol{x}_{ij}$ are $(1, S_i \times T_j, d_{ij})$

$$\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \mathcal{N}_3(\boldsymbol{m}, \Sigma) \quad k = 1, \dots, K$$
$$\boldsymbol{\lambda} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$
$$\theta \sim \mathcal{U}(0, 1)$$

$$K = 4, \;\; \boldsymbol{m} = (7, 0, 0), \;\; \Sigma = \text{diag}(2, 1.5, 1.5), \;\; \boldsymbol{\alpha} = (2, 2, 2, 2)$$

Posterior Inference

MCMC simulation was performed using Stan software.

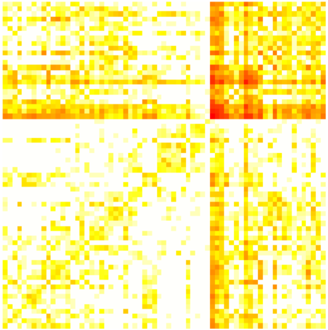


Figure: Observed flow

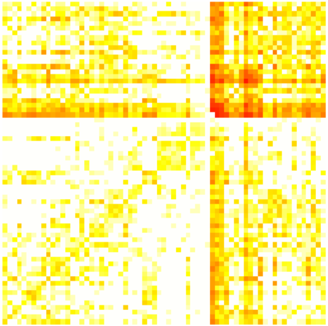
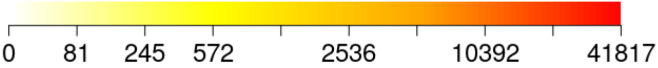


Figure: Predicted flow



The cluster estimate was found minimizing Binder's loss function.

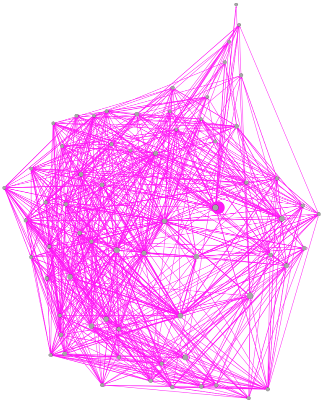


Figure: Estimated cluster 1

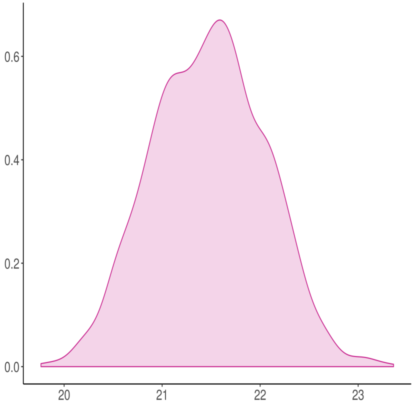


Figure: Posterior mean flow

The cluster estimate was found minimizing Binder's loss function.

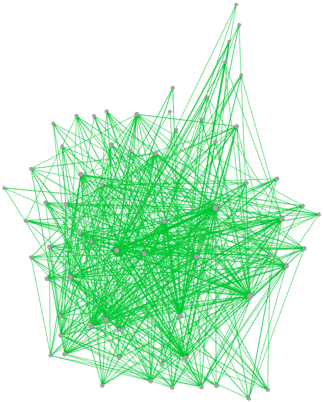


Figure: Estimated cluster 2

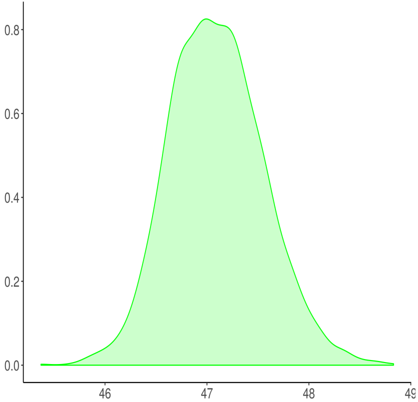


Figure: Posterior mean flow

Clustering

The cluster estimate was found minimizing Binder's loss function.

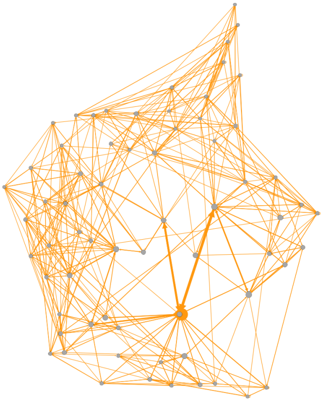


Figure: Estimated cluster 3

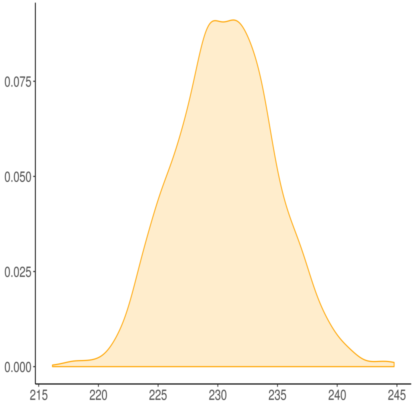


Figure: Posterior mean flow

Clustering

The cluster estimate was found minimizing Binder's loss function.

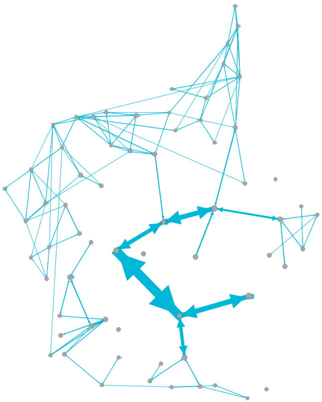


Figure: Estimated cluster 4

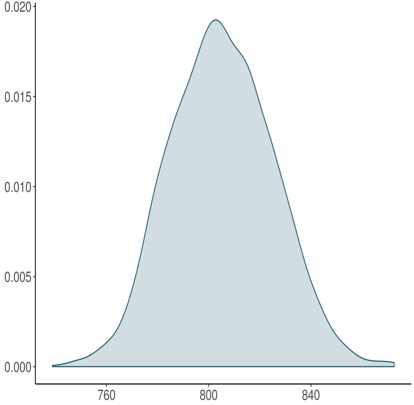


Figure: Posterior mean flow

Comparison

We assess the predictive performances of our model against other possible Bayesian "competitors".

- `Reg0infl`

$$Y_{ij}|\theta, \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \begin{cases} \theta + (1-\theta)\text{PM}(0|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } Y_{ij} = 0 \\ (1-\theta)\text{PM}(Y_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) & \text{if } Y_{ij} > 0 \end{cases}$$

$$\text{PM}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i=1}^{K} \lambda_i \text{Poi}(\mu_i)$$

$$\log \mu_{ijk} = \boldsymbol{\beta}_k \boldsymbol{x}_{ij}.$$

Comparison

We assess the predictive performances of our model against other possible Bayesian "competitors".

- `Reg0infl`
- `0infl`

$$Y_{ij}|\theta, \boldsymbol{\mu}, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \begin{cases} \theta + (1-\theta)\mathrm{PM}(0|\boldsymbol{\mu}, \boldsymbol{\lambda}) & \text{if } Y_{ij} = 0 \\ (1-\theta)\mathrm{PM}(Y_{ij}|\boldsymbol{\mu}, \boldsymbol{\lambda}) & \text{if } Y_{ij} > 0 \end{cases}$$

$$\mathrm{PM}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i=1}^{K} \lambda_i \mathrm{Poi}(\mu_i)$$

Comparison

We assess the predictive performances of our model against other possible Bayesian "competitors".

- `Reg0infl`
- `0infl`
- `Reg`

$$Y_{ij} | \theta, \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \text{PM}(Y_{ij} | \boldsymbol{\mu}_{ij}, \boldsymbol{\lambda})$$

$$\text{PM}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i=1}^{K} \lambda_i \text{Poi}(\mu_i)$$

$$\log \mu_{ijk} = \boldsymbol{\beta}_k \boldsymbol{x}_{ij}.$$

Comparison

We assess the predictive performances of our model against other possible Bayesian "competitors".

- `Reg0infl`
- `0infl`
- `Reg`

Model	MSE	LOO-ELPD	LPML
`Reg0infl`	**1,518.3**	**-14,270.5**	**-11,237.9**
`Reg`	1,685.4	-18,347.7	-18,341.7
`0infl`	300,373.4	-84,862.0	-103,089.3

Table: Predictive performances comparison.

Conclusions

- Presented a class of full Bayesian models to analyze the mobility of BikeMi
- Zero-inflated Poisson mixture regression models captures both the topology of the network and different range of behaviours in the flows
- Incorporate in the model edge-specific covariates

Conclusions

- Presented a class of full Bayesian models to analyze the mobility of BikeMi
- Zero-inflated Poisson mixture regression models captures both the topology of the network and different range of behaviours in the flows
- Incorporate in the model edge-specific covariates

Future developments

- Different clustering
- Considering the whole network
- Node specific covariates such as proximity to points of interest

References

[1] David A Binder. “Bayesian cluster analysis”. 1978.

[2] Bob Carpenter et al. “Stan: A Probabilistic Programming Language”. 2017.

[3] Peter Congdon. “A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling”. 2000.

[4] Martin Ester et al. “Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification”. 1995.

[5] Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. 2006.

[6] Anna Goldenberg et al. “A survey of statistical network models”. 2010.

[7] Michael Hahsler and Matthew Piekenbrock. *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*. 2018.

[8] Agostino Torti et al. *Modeling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan*. 2019.