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

Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi, Mario Beraha and Alessandra Guglielmi

Abstract We propose a Bayesian model for the analysis of flow counts on a network for an application to a bike sharing platform (BikeMi) in Milano. Incorporating edge-specific covariates, we assume a zero-inflated Poisson mixture regression model, which easily accommodates for the sparse nature of the network under investigation.

Abstract *In questo lavoro proponiamo un modello bayesiano per l'analisi dei conteggi dei flussi su una rete, per un'applicazione relativa ad una piattaforma di bike sharing di Milano (BikeMi). Incorporando covariate specifiche ad ogni arco, assumiamo un modello mistura di Poisson zero-inflated, che permette di catturare facilmente la natura sparsa delle rete presa in considerazione.*

1 Introduction

Self service bike sharing systems have grown in popularity all over the world in recent decades, with dozens of new players entering the market each year. The exploitation of this novel transport system has generated an enormous quantity of data that could lead to a better understanding of city mobility. In this work, we analyze data from BikeMi, Milan oldest bike sharing system. BikeMi has been introduced in 2008, and nowadays encompasses more that 4000 bicycles with more than 250 stations.

Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi
Dipartimento di Matematica, Politecnico di Milano,
e-mail: {giulia.bissoli, celeste.principi, gianmatteo.rinaldi}@mail.polimi.it

Mario Beraha[†], Alessandra Guglielmi
Dipartimento di Matematica, Politecnico di Milano
e-mail: {mario.beraha, alessandra.guglielmi}@polimi.it
[†] Also affiliated with Università degli Studi di Bologna

We focus on the analysis of the flow of bikes from one station to another. By looking at the bike sharing system as a complex network [6], we assume a Bayesian regression model for the flow counts on the edges of the network. In particular, we assume a mixture model [5] for the flow counts, which yields a cluster estimate of the flows that we are able to interpret.

The structure of the paper is as follows: after having introduced the model in Section 2 and presented the dataset in Section 3, we report some posterior inference in Section 4, showing the goodness of fit of our model and the interesting insights from the estimated clusters. Finally in Section 5 we compare our model to competitors.

2 Zero-inflated Poisson mixture regression models

Let $\mathscr{G} = (V,E)$ be a directed graph with vertices V and edges $E = \{(i,j)\}$. In our application, we consider the problem of modelling the data flow on the edges E. For each edge (i,j) we define the variable Y_{ij} as the amount of bike trips from i to j taken on a fixed period of time. We also assume the availability of a set of covariates $\boldsymbol{x}_{ij}$ for each arc, which, a priori, might influence the flow on that particular arc. The Y_{ij}s are thus counts, so that, a standard choice to model these variables would be then to use the Poisson distribution with suitable parameters. We also aim at capturing the topology of the network, i.e. assigning zero flow to the edges that should not be in the network; in addition, our model should be flexible enough to represent a wide class of scenarios, from routes travelled very often to the ones seldom used.

Combining these insights together led us to consider the following zero-inflated Poisson mixture regression model

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

$$\log \mu_{ijk} = \boldsymbol{\beta}_k \boldsymbol{x}_{ij}, \tag{2}$$

for any couple of nodes (i,j). We assume that, conditionally to all parameters, the Y_{ij}s are independent. Here PM stands for Poisson Mixture, i.e. $\text{PM}(\boldsymbol{\mu},\boldsymbol{\lambda}) = \sum_{i=1}^{K} \lambda_i \text{Poi}(\mu_i)$ with $\boldsymbol{\mu} = (\mu_1,\dots,\mu_K)$, $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_K)$, and K is a fixed positive integer. Observe that $\boldsymbol{\mu}_{ij} = (\mu_{ij1},\dots\mu_{ijK})$ is linked to the $p-$dimensional vector $\boldsymbol{x}_{ij}$ via the canonical link function (2) through a p-dimensional regression parameter $\boldsymbol{\beta}_k$.

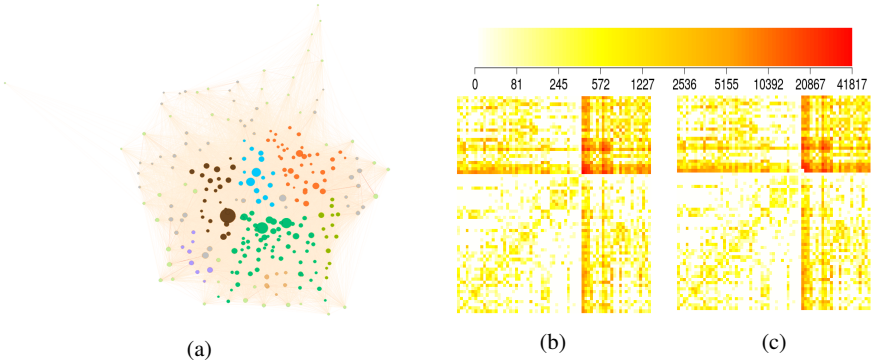
The prior specification assumes parameters $(\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_K)$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ independent with marginal distributions as follows:

$$\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \mathscr{N}(\boldsymbol{m},\Sigma) \quad k=1,...,K \tag{3}$$

$$\boldsymbol{\lambda} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_K) \tag{4}$$

$$\theta \sim \mathscr{U}(0,1) \tag{5}$$

Assuming zero-inflated likelihood thus accounts for the presence of edges in the graph with zero flow, while the Poisson mixture let us easily model a various range



(a) (b) (c)

Fig. 1: (a): The nodes of the clustered network: the size of each dot is proportional to the number of stations it represents. (b): observed flow. (c): predicted flow. The entry of the matrix in row i, column j represents the flow on edge (i, j).

of behaviours in the flows. Moreover, we take into account covariates as regressors in the generalized linear model through the locations μ_{ijk}.

3 BikeMi Dataset

We consider data arising from the most popular and oldest bike sharing system in Milan. The system is composed of 263 stations where users go and pick up and then drop off the bikes. The dataset consists of the records of 350,093 trips between pairs of stations between January 25^{th} and March 6^{th}, 2016. Using the stations as nodes in the graph would, in principle, lead to a graph with $263^2 \approx 69k$ edges, which makes the computational burden unfeasible.

As many stations lie within less than 100m from each others, we perform a clustering on the geo-locations of the stations, using the popular-density based clustering algorithm DBSCAN [4], and have consider the barycenter of each cluster as a node in the graph. In the end, we were left with 67 nodes. We report the clusters obtained using DBSCAN in Figure 1(a), which defines the nodes of the graph we analyze. We have then redefined y_{ij}, the observed flow counts between nodes i and j as the total flow from all the stations which were collapsed through DBSCAN procedure into node i and into node j.

Of course, other strategies are possible to reduce the dimensionality of the problem. For example one could consider Milan's neighborhoods (NILs) as nodes in the graph, as done in [7], and do not rely on the spatial clustering. This approach, however, would imply that trips from one node to itself (within the same NIL) might be longer than trips from one node to another, as two bike stations might be at the opposite sides of a NIL (former case) or just across the border of the two neighboring NILs (latter case). On the contrary DBSCAN clusters together points based on their local density, so that the estimated clusters can genuinely be well represented by their barycenters.

Similarly as in [3], we assume that bike flow from i to j might be depend on: (i) the geographical distance between the stations (d_{ij}) (ii) the outer strength of the pick up node (S_i) and (iii) the inner strength of the drop off node (T_j). The outer (inner) strength of a node is defined as the total amount of trips departing from (arriving at) it. So that, for each arch, the vector of covariates is chosen to be $\boldsymbol{x}_{ij} = (1, S_i T_j, d_{ij})$.

4 Posterior Inference

We fix the hyperparameters in (3) - (4) to be $\boldsymbol{m} = (7,0,0)$, $\Sigma = \text{diag}(2, 1.5, 1.5)$, $\boldsymbol{\alpha} = (2,2,2,2)$. The number of components K was selected among several possible values as the one giving clearer interpretability of the posterior cluster estimate; with $K = 4$ we were able to interpret 4 different behaviours in terms of geolocation of the source and target nodes of the edges.

Posterior inference was computed using Stan software [2]; MCMC chains were ran each for 2000 iterations after 2000 iterations of burn-in. Convergence was checked using both visual inspection of the chains and standard diagnostics available in the CODA package. From the comparison between the observed flow on the network (Figure 1 (b)) and the posterior expected value (Figure 1(c)), we see that our model correctly assigns zero flow to several edges, while predicting accurately the flow on the most travelled ones. It is clear that our model predicts quite well the observed flow.

Figure 2 reports a point estimate of the clustering structure of the edges arising from the Poisson mixture in (1), i.e. through the allocation variables identifying the mixture components. The point estimate has been obtained by finding the MCMC draw that minimizes the Binder loss function [1]. The first cluster includes all arcs with low flow and consists mainly of arcs belonging to the periphery of the city. The second group contains many arcs which are closer to the city center. The third cluster clearly groups together arcs with high flow in the middle of the city. Finally the last one includes only few arcs that represent the most travelled routes. The parameter β_{k3}, that represents the impact of the distance between node i and node j on the flow, is strongly negative for all the clusters, as expected, except the first one, where its values are positive although close to 0.

In all the cluster maps, the central big dot is Milan city center (Duomo), the other recognizable nodes are Cadorna train station (on the left of the map) and Centrale train station (top right), which, as expected, are focal points for the bike sharing system.

5 Competitor Models

Finally, we compare the predictive performances of our approach to the ones obtained by alternative models. Specifically, we consider model (1) without covariates (`0infl`), a standard Poisson mixture regression (i.e. model (1) - (2) with $\theta \equiv 0$) (`Reg`) and our model (`Reg0infl`). The corresponding priors are matched to give same a priori information on the same parameters. We report the comparison in

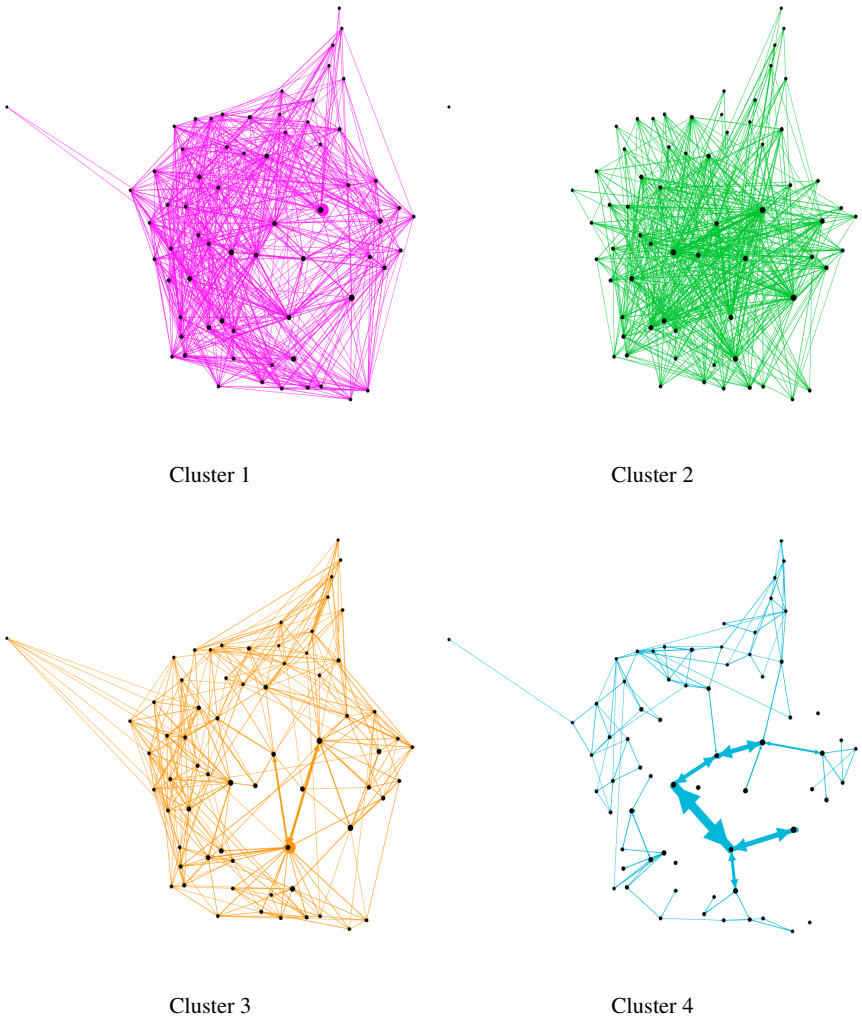


Fig. 2: The clustering of the edges arising from the Poisson mixture model.

Table 1, where for each model the following indexes of performance are shown: (i) *MSE*: the mean square error between the posterior mean and the observed flow. (ii) *LOO-ELPD*: the leave one out estimate of the expected log predictive density (computed using the `loo` package). (iii) *LPML*: the log pseudo marginal likelihood.

`Reg` and `Reg0infl` clearly outperform `0infl`, moreover, `Reg0infl` is slightly better than `Reg`. From a visual inspection of the predicted flows on both models, we can conclude that the main difference between `Reg0infl`'s and `Reg`'s prediction is that `Reg0infl` correctly assigns zero flow on a significant number of edges while `Reg` assigns to the same edges a small but positive flow.

Model	MSE	LOO-ELPD	LMPL
`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 1: Predictive performances comparison.

6 Discussion and Conclusions

In this paper we have presented a full Bayesian model to analyze the mobility of one of the bike sharing systems in Milan. The proposed approach is based on modelling the counts of the number of travels between pairs of (clusters of) bike stations as the flow data on the edges of a complex network. We have assumed a zero-inflated Poisson mixture regression model to capture both the topology of the network and various range of behaviours in the flows, incorporating in the model also information coming from covariates such as the pairwise geographical distance between nodes.

Through MCMC simulations, we have shown how our model compares favorably against possible Bayesian competitors and how it describes the overall structure of the data. Nonetheless, we believe that incorporating into the model other information such as the proximity of a bike station to an underground stop or other places of interest might improve the overall quality of the prediction.

In the future, we aim at applying our model to the whole network and compare the inference. As an alternative we could also develop ad-hoc clustering strategies to reduce the dimensionality of the dataset, aggregating more nodes in the periphery of the city while keeping distinct the ones in the center.

Acknowledgement

We thank Clear Channel for having shared these data with us.

References

1. Binder, D.A.: Bayesian cluster analysis. Biometrika **65**(1), 31–38 (1978)
2. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software, Articles **76**(1), 1–32 (2017). DOI 10.18637/jss.v076.i01. URL https://www.jstatsoft.org/v076/i01
3. Congdon, P.: A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling. Geographical Analysis **32**(3), 205–224 (2000)
4. Ester, M., Kriegel, H.P., Xu, X.: Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In: International Symposium on Spatial Databases, pp. 67–82. Springer (1995)
5. Frühwirth-Schnatter, S.: Finite mixture and Markov switching models. Springer Science & Business Media (2006)
6. Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M., et al.: A survey of statistical network models. Foundations and Trends® in Machine Learning **2**(2), 129–233 (2010)
7. Torti, A., Pini, A., Vantini, S.: Modeling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan. Tech. rep., MOX, Politecnico di Milano (2019)