A Bayesian model for data flow: BikeMi

Andrea De Gobbis, Lorenzo Ghilotti, Giorgio Meretti

November 22, 2019

Politecnico di Milano

Bike Sharing

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

We aim to develop a Bayesian model for the analysis of the flow counts on a complex network.

Application to the bike sharing platform BikeMi, in Milan.

The Bike Sharing System

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

- Bike sharing is a service in which bicycles are made available for shared use to individuals on a short term basis for a price or free.
- BikeMi allows people to borrow a bike from a "station" and return it at another station belonging to the system.

Dataset

The data covers a period between 25/01/2016 and 07/03/2016, with a total of 35 thousand data points.

- Date;
- Time of departure/arrival;
- Station of departure/arrival;
- Weekday: binary indicating if the day is in the weekend or not.

Dataset

The data covers a period between 25/01/2016 and 07/03/2016, with a total of 35 thousand data points.

- Date;
- Time of departure/arrival;
- Station of departure/arrival;
- Weekday: binary indicating if the day is in the weekend or not.

Dataset

The data covers a period between 25/01/2016 and 07/03/2016, with a total of 35 thousand data points.

- Date;
- Time of departure/arrival;
- Station of departure/arrival;
- Weekday: binary indicating if the day is in the weekend or not.

Dataset

The data covers a period between 25/01/2016 and 07/03/2016, with a total of 35 thousand data points.

- Date;
- Time of departure/arrival;
- Station of departure/arrival;
- Weekday: binary indicating if the day is in the weekend or not.

Statistics on a Graph

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

Setting $\implies$ complete directed graph $\mathcal{G} = (V, E)$

- vertices V representing the stations
- edges $E = \{(i, j)\}$: the weight of the edges will be the count flow of bikes.

As pre-processing it was introduced a node clusterization using the DBSCAN algorithm, since many stations are very close to one another.

Statistics on a Graph

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

Setting $\implies$ complete directed graph $\mathcal{G} = (V, E)$

- vertices V representing the stations
- edges $E = \{(i, j)\}$: the weight of the edges will be the count flow of bikes.

As pre-processing it was introduced a node clusterization using the DBSCAN algorithm, since many stations are very close to one another.

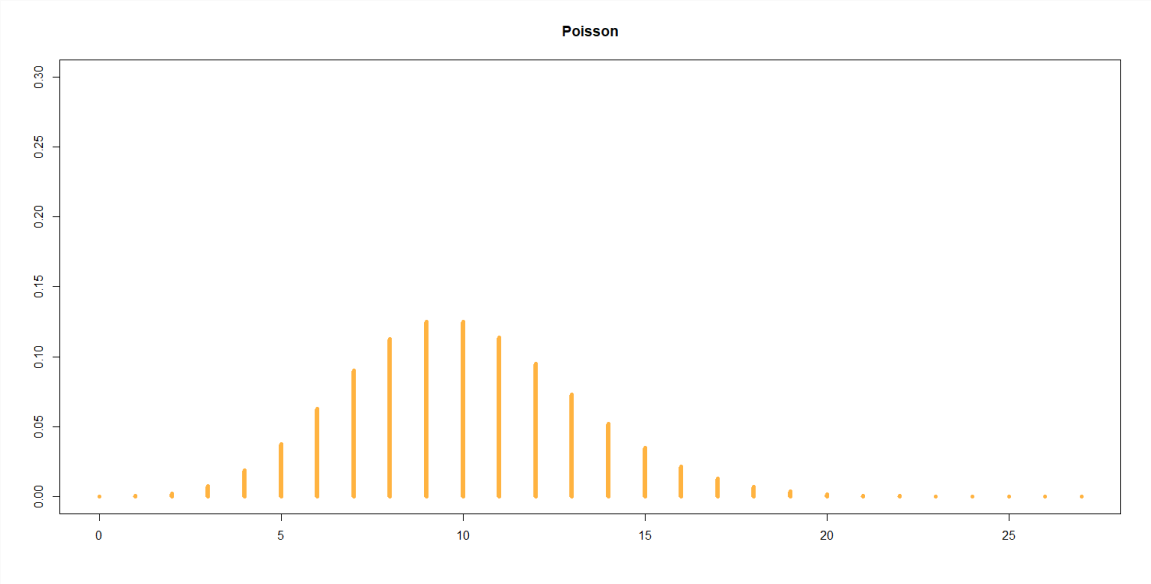
A Basic Model for the Phenomenon

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

The flow counts are modeled with a Poisson distribution:

$$\begin{cases} Y_{ij} \sim \text{Poi}(Y_{ij}|\mu_{ij}) \\ \\ \mu_{ij} = \exp\{\boldsymbol{\beta} \cdot \mathbf{x}_{ij}\} \end{cases} \tag{1}$$

The means are derived through loglinear regression.



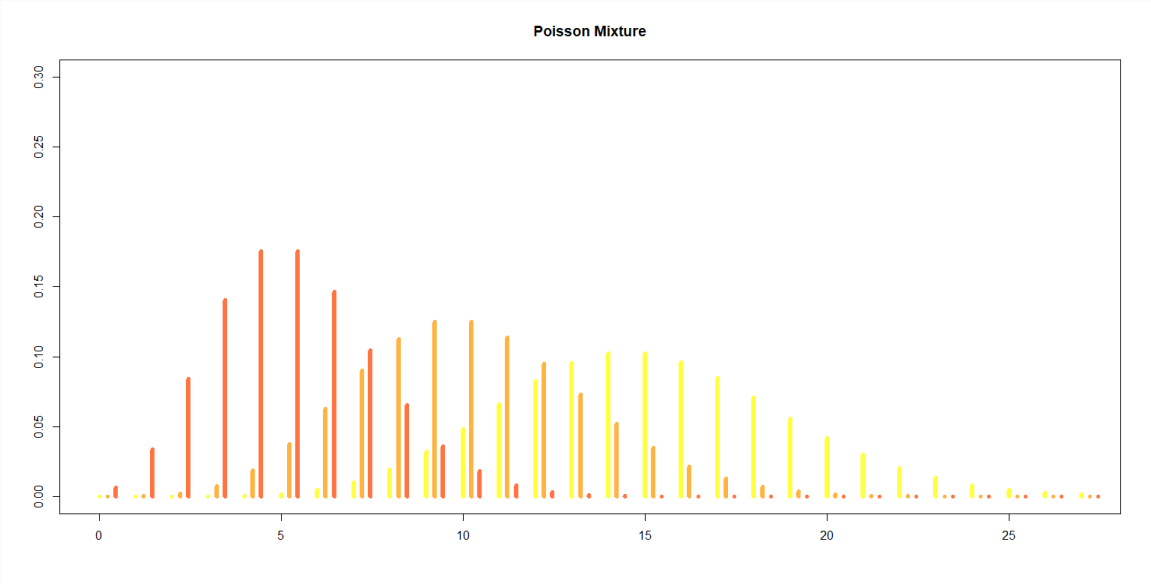
An Improved Model for the Phenomenon

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

We introduce a division in K clusters:

$$\begin{cases} Y_{ij} \sim \mathrm{PM}(Y_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) \\ \mathrm{PM}(\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \lambda_k \mathrm{Poi}(\mu_{ijk}) \\ \mu_{ijk} = \exp\{\boldsymbol{\beta}_k \cdot \mathbf{x}_{ij}\} \end{cases} \tag{2}$$

where λ_k is the probability of belonging to group k.

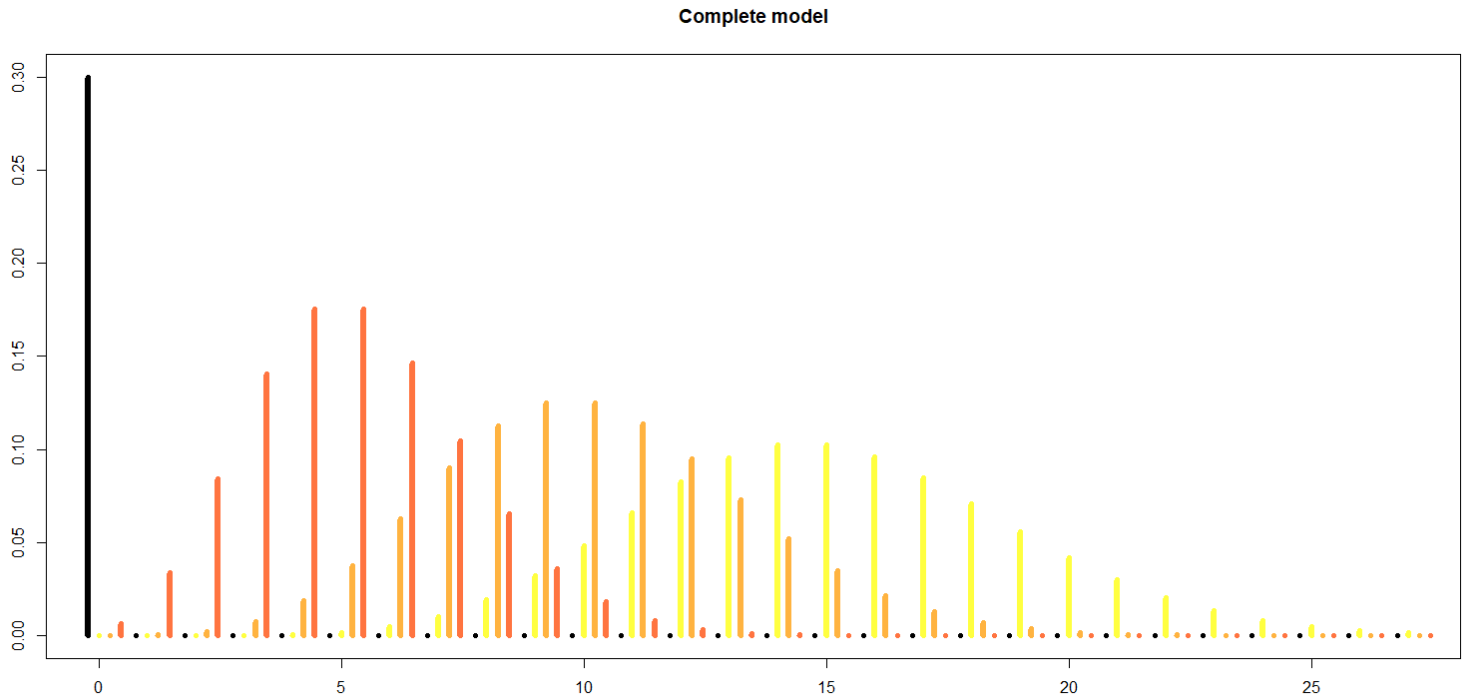


A Proposed Model for the Phenomenon

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

Since the model will determine the topology of the graph we might desire a better control on $\mathbb{P}(Y_{ij} = 0)$, they introduce a delta measure in 0.

$$Y_{ij} \sim \theta\delta_0 + (1-\theta)\mathrm{PM}(Y_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda})$$



Zero-Inflated Poisson Mixture Regression Model

Final model proposed in the paper:

$$\begin{cases} Y_{ij} \sim \theta\delta_0 + (1-\theta)\mathrm{PM}(Y_{ij}|\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) \\ \mathrm{PM}(\boldsymbol{\mu}_{ij}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \lambda_k \mathrm{Poi}(\mu_{ijk}) \\ \mu_{ijk} = \exp\{\boldsymbol{\beta}_k \cdot \mathbf{x}_{ij}\} \end{cases} \tag{3}$$

where $\theta \in [0,1], \boldsymbol{\lambda} \in Simp(K-1)$, $\boldsymbol{\mu}_{ij} \in \mathbb{R}^K, \forall i, j$, $\boldsymbol{\beta}_k \in \mathbb{R}^p, \forall k \in 1 \dots K$ are the parameters of the model.

Set priors:

$$\begin{cases} \boldsymbol{\beta}_k \sim \mathcal{N}(\mathbf{m}_0, \Sigma) \\ \boldsymbol{\lambda} \sim \mathrm{Dirichlet}(\alpha) \\ \theta \sim \mathcal{U}[0,1] \end{cases} \tag{4}$$

Covariates: strenghts of the adjacent nodes and the geographical distance between the stations.

Comments on the Model

Pros

- Proper description of flows
- interpretable clustering of the edges inside the network

Cons

- Lack of day-by-day prediction of the flow
- no distinction between different hours of the day or between various days

New Challenges

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

For our project we decided on three main objectives:

- Autoregressive Model
  Adding the *time dependence* to the model $Y_{ij} \to Y_{ij}(t)$, with the objective of predicting the network flow for a new day;
- Introducing new covariates to keep track, for instance, of weekdays and weather;

New Challenges

For our project we decided on three main objectives:

- Autoregressive Model
  Adding the *time dependence* to the model $Y_{ij} \to Y_{ij}(t)$, with the objective of predicting the network flow for a new day;
- Introducing new covariates to keep track, for instance, of weekdays and weather;

New Challenges

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

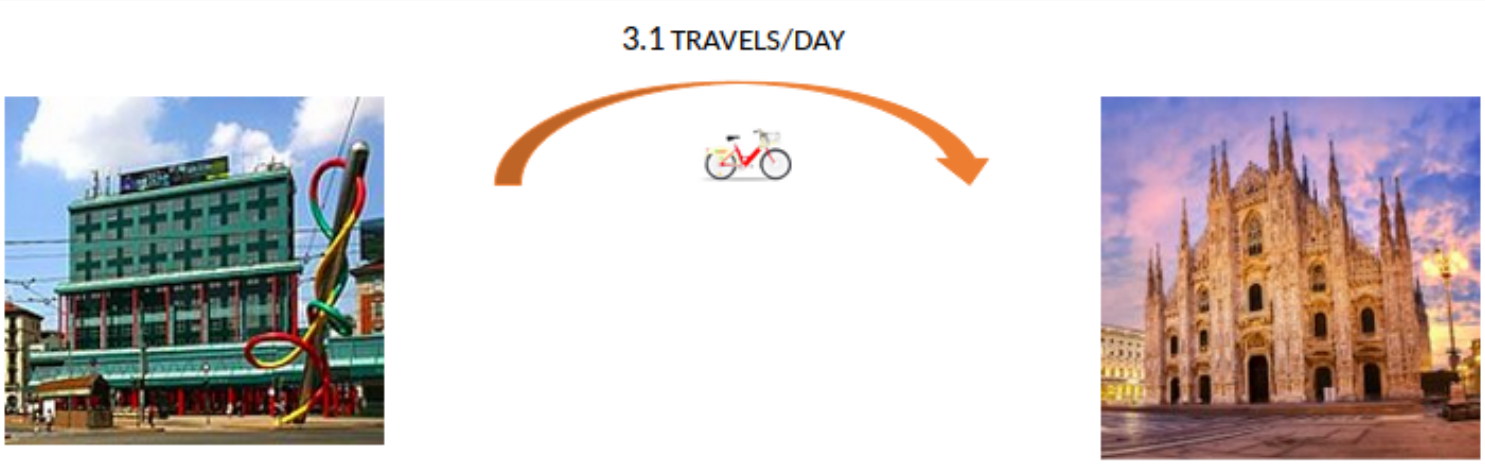
For our project we decided on three main objectives:

- Autoregressive Model
  Adding the *time dependence* to the model $Y_{ij} \to Y_{ij}(t)$, with the objective of predicting the network flow for a new day;
- Introducing new covariates to keep track, for instance, of weekdays and weather;

Impact of Time

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

WEEKENDS



Impact of Time

Andrea De Gobbis,
Lorenzo Ghilotti,
Giorgio Meretti

WEEKDAYS

New Challenges

For our project we decided on three main objectives:

- Autoregressive Model
  Adding the *time dependence* to the model $Y_{ij} \to Y_{ij}(t)$, with the objective of predicting the network flow for a new day;
- introducing new covariates to keep track, for instance, of weekdays and weather;
- leaner computations: a mixed R/C++ implementation.

Initial Bibliography

- *A Bayesian model for network flow data: an application to BikeMi trips*; Giulia Bissoli, Celeste Principi, Gian Matteo Rinaldi, Mario Beraha and Alessandra Guglielmi, 2019