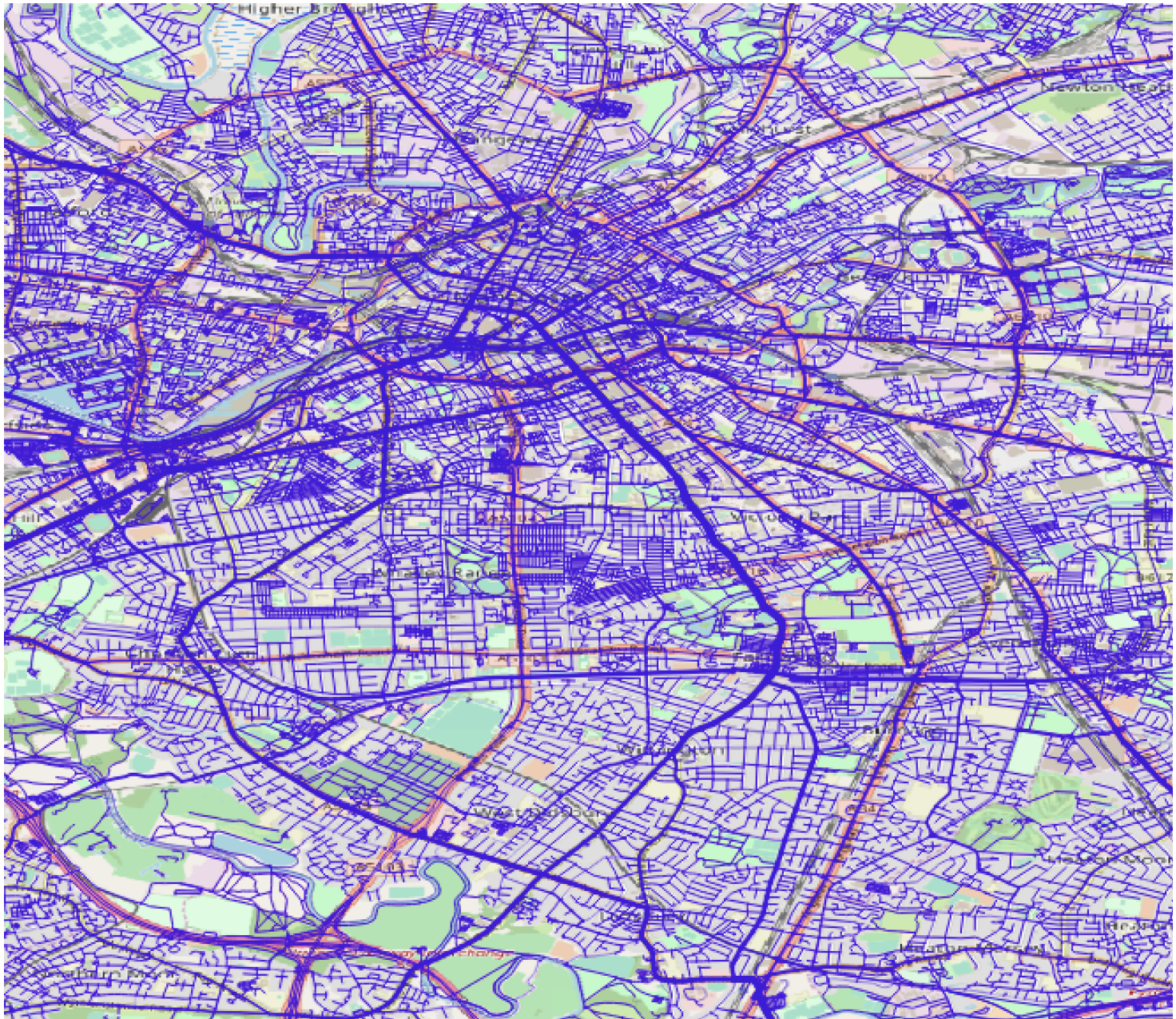
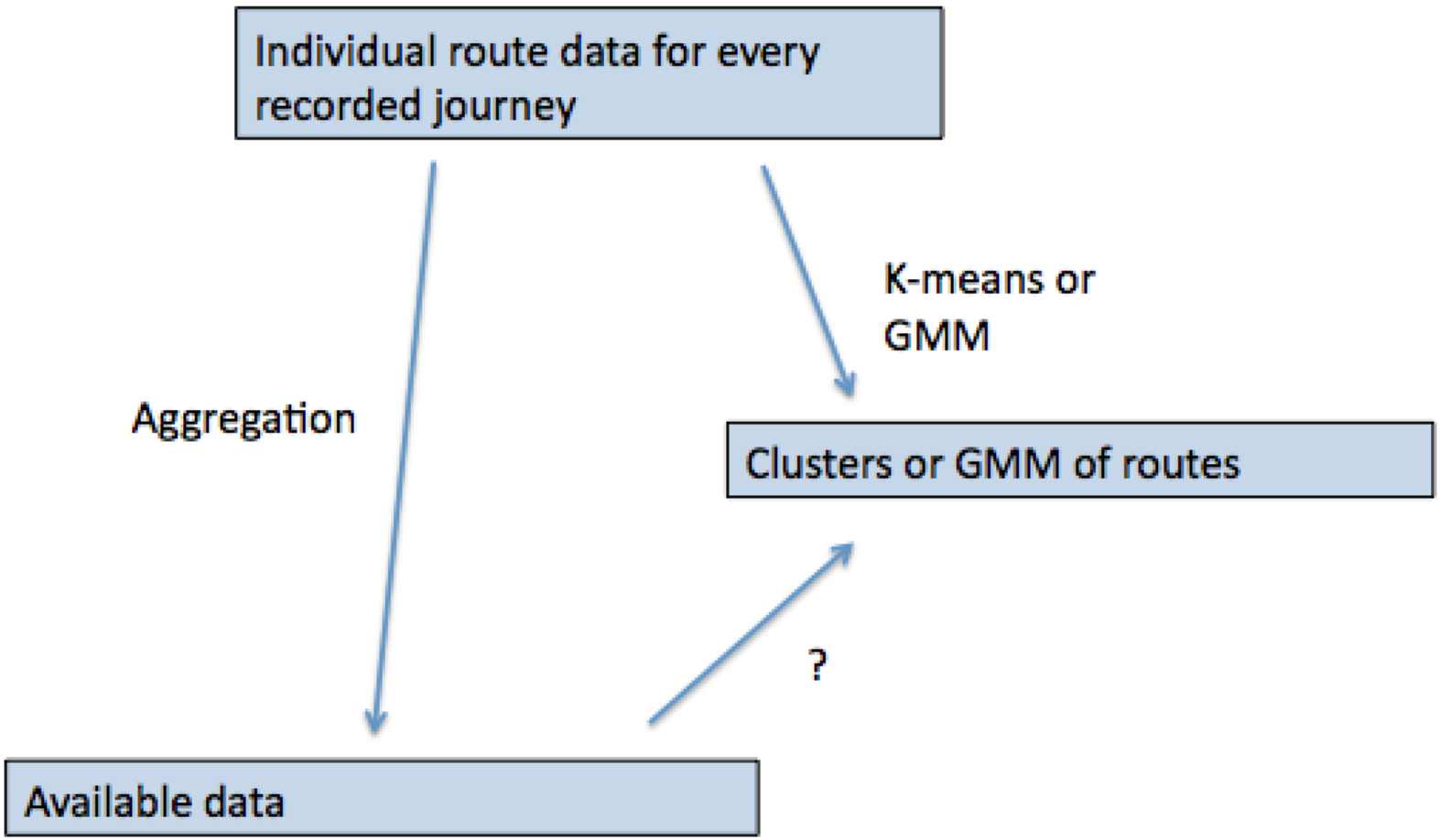
Strava Data Disaggregation

Christina Boididou, James Hook, Lizzi Pitt, Malena Sabate-Landman

13th June 2018



Partially Inverting Aggregation



Forwards Model

Manchester road network $G(V, E)$.

Set of all rides $R = \left\{ \boldsymbol{r}_\ell \in \{0,1\}^E \ : \ \ell = 1, \ldots, N \right\}$.

GMM with k-clusters $R \approx \sum_{i=1}^{k} \phi_i \mathcal{N}(\mu_i, \Sigma_i)$, where $\mu_i \in \mathbb{R}_+^E$.

Let $C \in \mathbb{N}^{k \times m}$ with $c_{ij} =$ number of rides from cluster i on day j.

Our data is given by $D \in \mathbb{N}^{E \times m} \sim \mathcal{N}(\mu_D, \Sigma_D)$, with

$$\mu_D = MC,$$

where $M = [\mu_1, \mu_2, \ldots, \mu_k]$.

Inverse Problem: Given aggregated data D determine the means/centres M.

Inverse Problem

Maximum Likelihood Estimate from Non-Negative Matrix Factorization

$$\min -\log\left(\mathbb{P}(D|M,C)\right) \propto \min_{M\in\mathbb{R}_+^{E\times k}, C\in\mathbb{R}_+^{k\times m}} \|D - MC\|_F.$$

Incorporating prior $p(M,C)$ results in Maximum A Posteriori problem

$$\min_{M\in\mathbb{R}_+^{E\times k}, C\in\mathbb{R}_+^{k\times m}} \|D - MC\|_F^2 - 2\sigma^2 \log\left(p(M,C)\right).$$

Prior should promote 'realistic' routes and route frequencies.

Route Length Prior

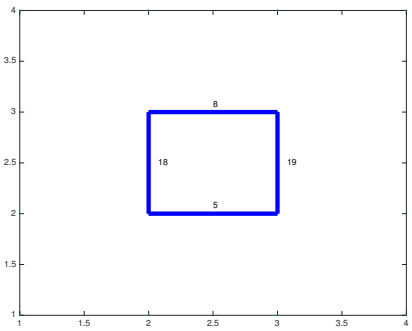
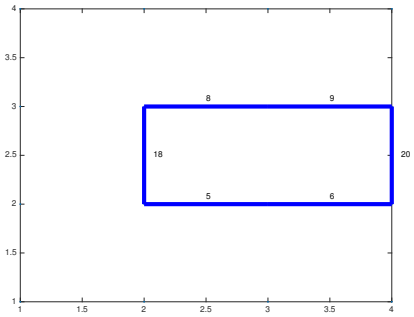
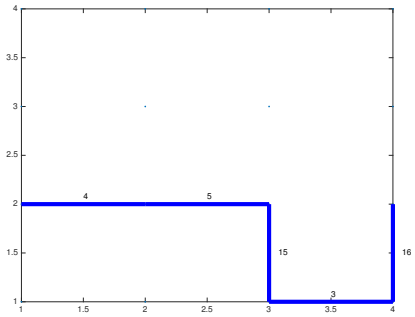
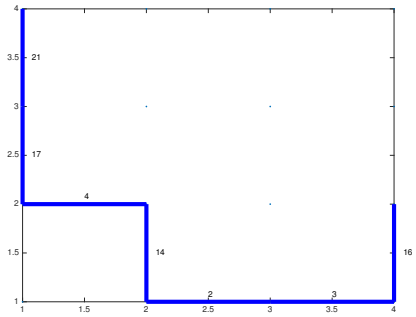
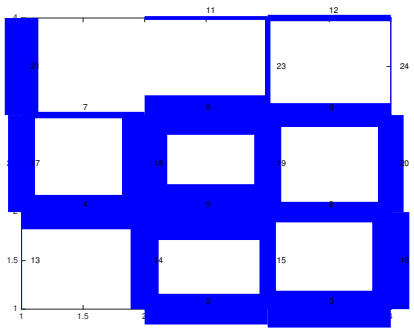
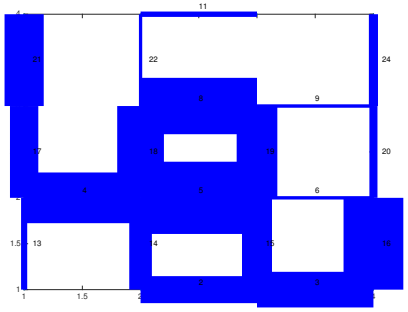
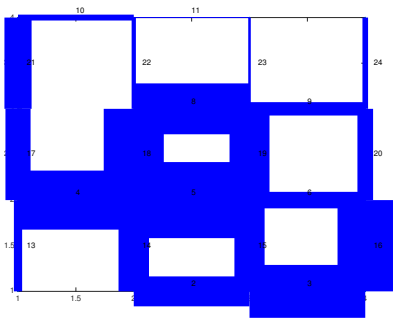
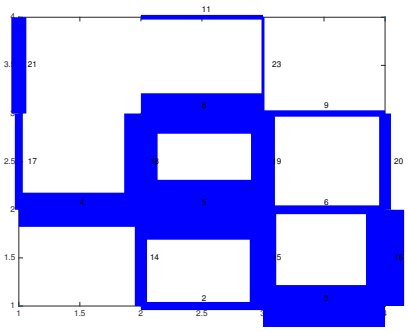
Ride till you crash model.

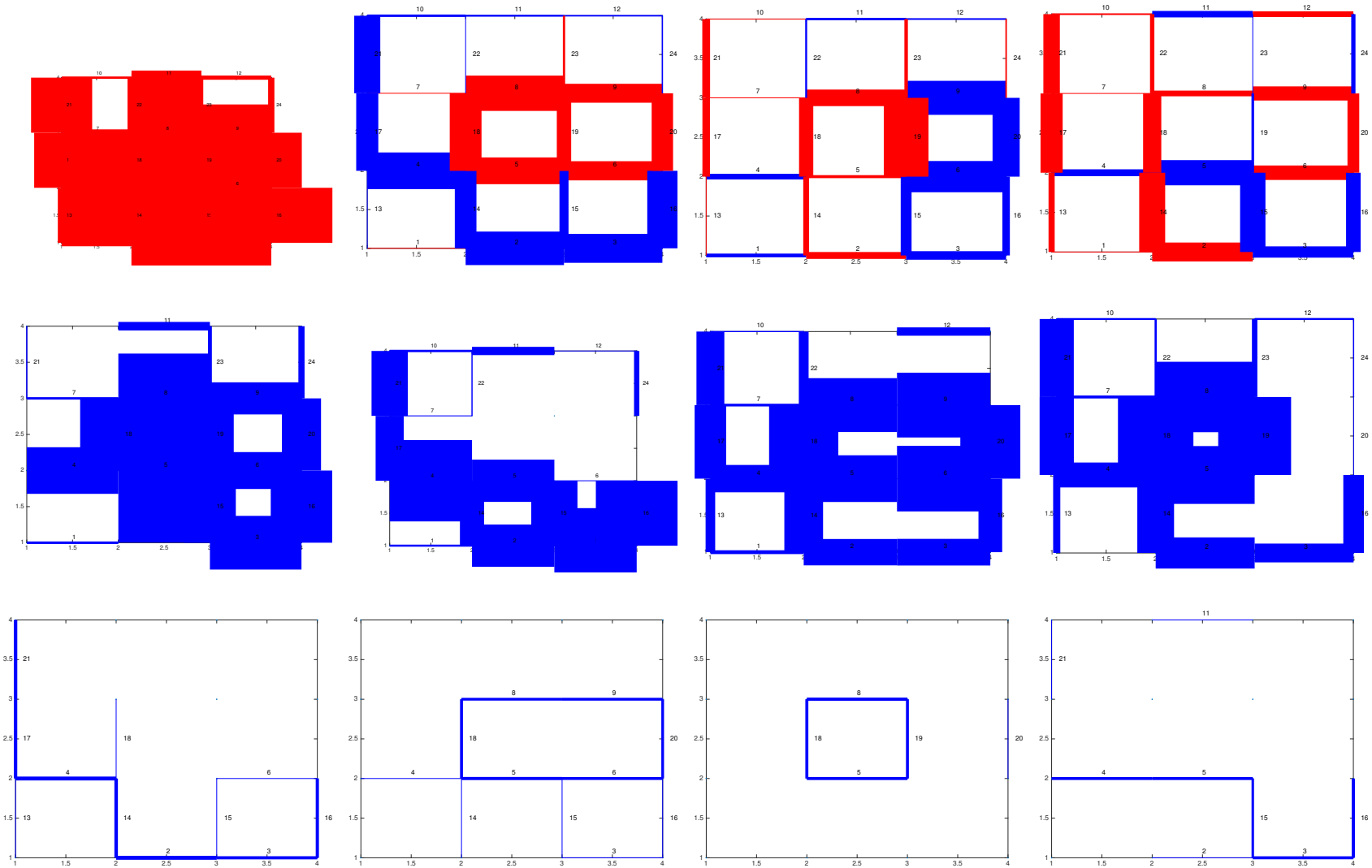
Rider crashes at rate λkm^{-1}, results in exponentially distributed route lengths.

We also know the total number of distinct activities per day $a \in \mathbb{R}^m$.

MAP from $\ell 1$ regularized, linearly constrained NNMF

$$\min_{M \in \mathbb{R}_+^{E \times k},\ C \in \mathbb{R}_+^{k \times m}\ :\ C^\top \underline{1} = a} \|D - MC\|_F^2 - 2\sigma^2 \lambda \|M\|_{1,1}.$$

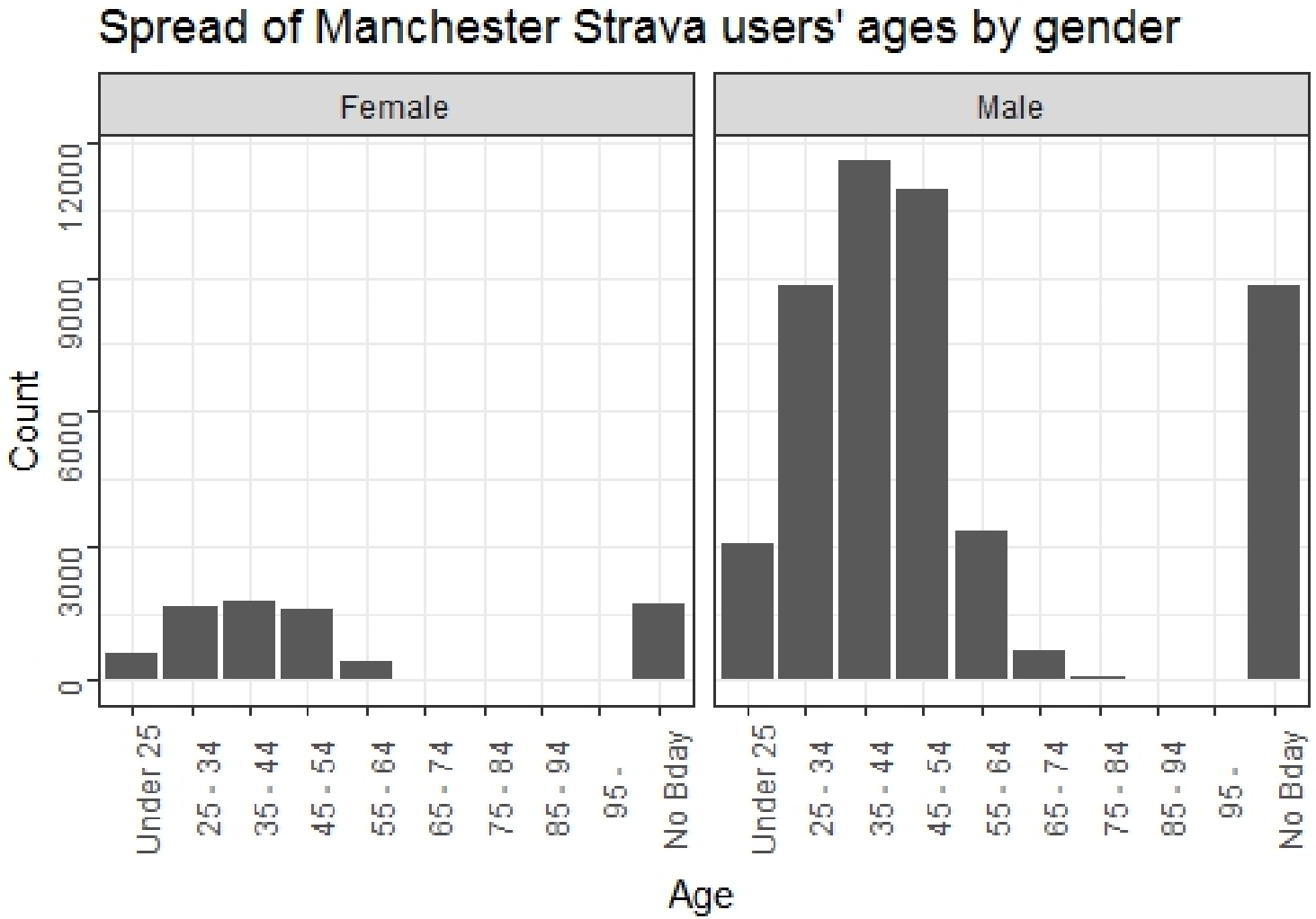











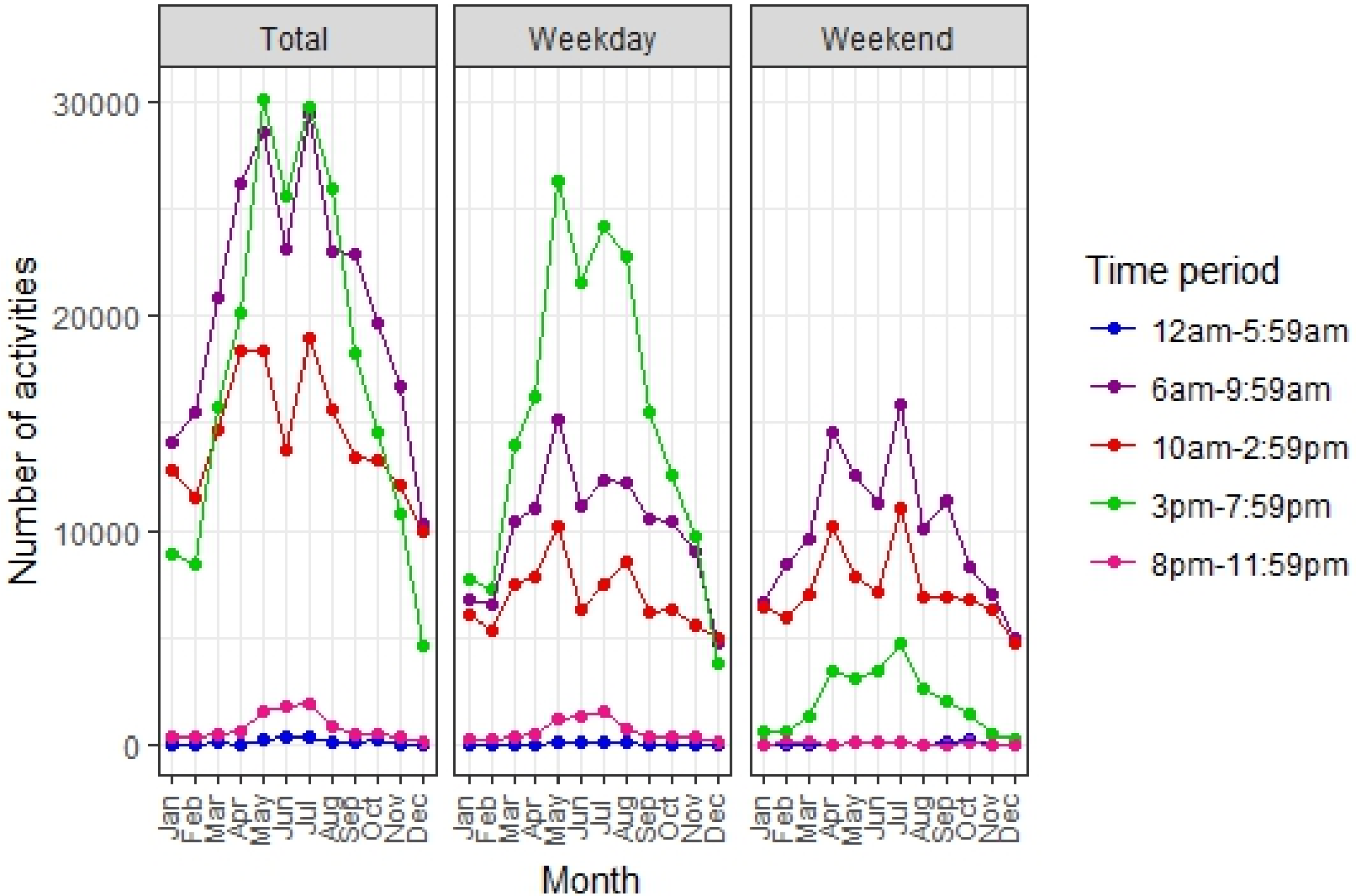
Stronger Prior

- Manchester Strava data has $\approx$ 100,000 roads and activity is only broken down into $\approx$ 100 time windows.
- We expect routes to be organized into $>>$ 100 clusters so need a strong prior for MAP problem to be well posed.
- Route length distribution worked well on smaller problem but won't be enough on full data set.
- Graph theoretic conditions: Basis vectors should be single connected components centred on a distinct path or cycle.
- Use known routes e.g. race routes and club rides to inform our choice of prior and validate our findings.

What affects the usage of these routes?



What affects the usage of these routes?



Conclusion

- Methodology works on toy problem.
- Could be applied to Strava data with a stronger prior.
- GIS software required to visualise routes.
- Could be applied to other problems with aggregated data.

Thanks!

Thanks for listening!
Any questions?