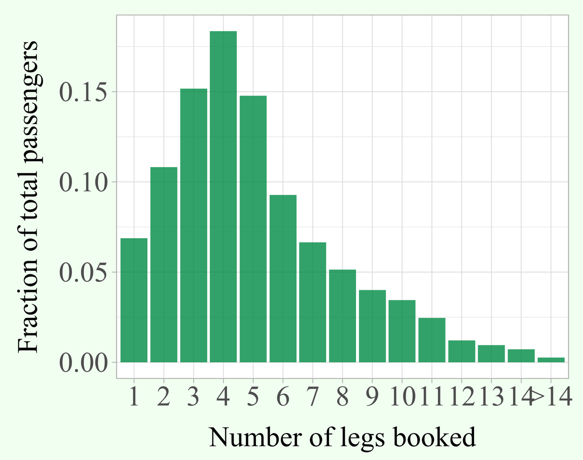
Detecting outlying demand in multi-leg bookings for transportation networks

Nicola Rennie[1], Catherine Cleophas[1,2], Adam M. Sykulski[1], Florian Dost[1,3]

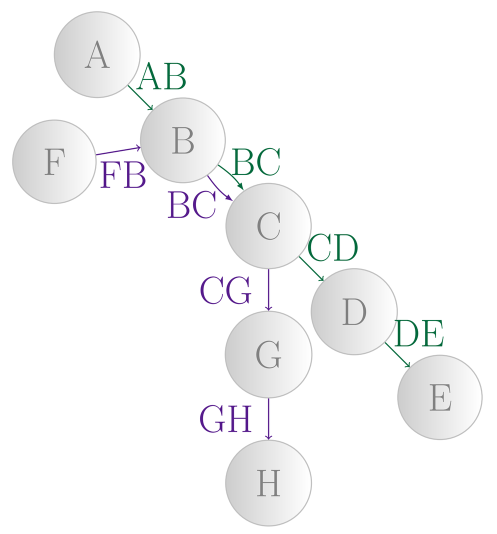
[1] Lancaster University, [2] Christian-Albrechts-Universität zu Kiel, [3] Brandenburg University of Technology

1. Multi-leg bookings

Transport service providers offer a large number of interconnected legs that let passengers travel along a multitude of itineraries. The distribution of the number of legs that passengers booked shows that only 7% of passengers booked single-leg itineraries. Almost half of all bookings spanned five or more legs.



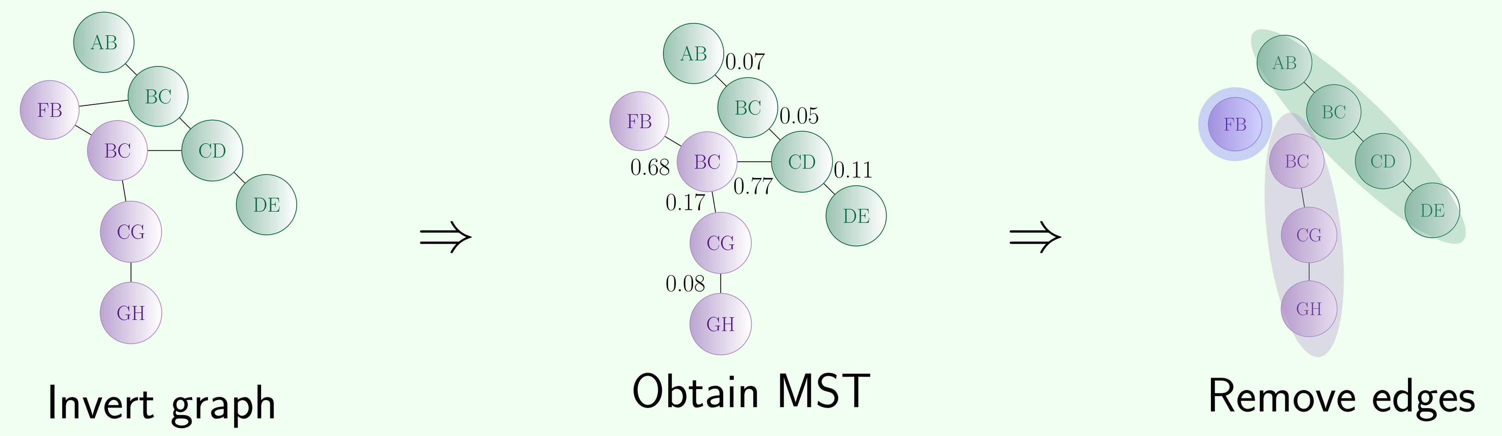
2. Outliers in transportation networks



Certain legs will share common outliers, as a common set of passengers traverses them. We represent the transportation network as a graph where nodes represent stations, and edges represent legs.

3. Clustering highly correlated legs

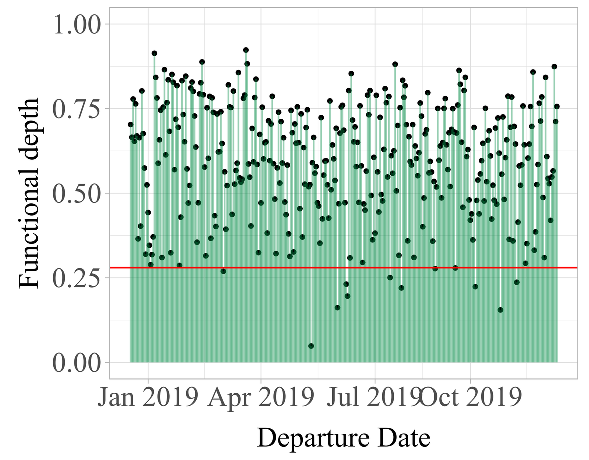
Neither considering each leg independently, nor jointly considering the network as a whole will create the best results. Therefore, we use a **minimum spanning tree** (MST) clustering algorithm to partition the network.



The edge weights are defined as $w(ij, jk) = 1 - \rho(ij, jk)$, where $\rho(ij, jk)$ is the **functional dynamical correlation** between adjacent edges. A correlation threshold of 0.5 is used to remove edges to form the clusters.

4. Functional depth for outlier detection

Functional depth quantifies how close to the most central trajectory a booking pattern is. The most outlying trajectories have the lowest depths. We calculate the functional depth for each departure on each leg, d_{nl}. We also calculate a threshold for the depths on each leg, C_l.



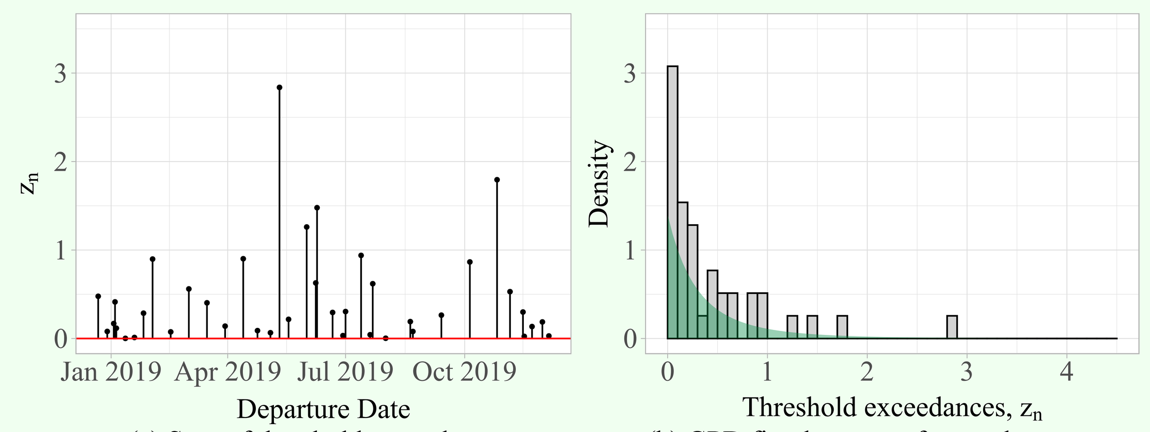
We transform the functional depths to make comparisons between legs with different thresholds:

$$z_{nl} = \frac{C_l - d_{nl}}{C_l}.$$

Departures with a value of z_{nl} above zero are classified as outliers.

5. Aggregating information within clusters

We sum the non-negative threshold exceedances across the L legs within the cluster. That is, $z_n = \sum_{l=1}^{L} z_{nl} \mathbb{1}_{\{z_{nl}>0\}}$. Outliers that are larger, or are detected in multiple legs give larger values of z_n.



(a) Sum of threshold exceedances (b) GPD fitted to sum of exceedances

We fit a Generalised Pareto Distribution (GPD) to the sum of the threshold exceedances.

6. Constructing a ranked alert list

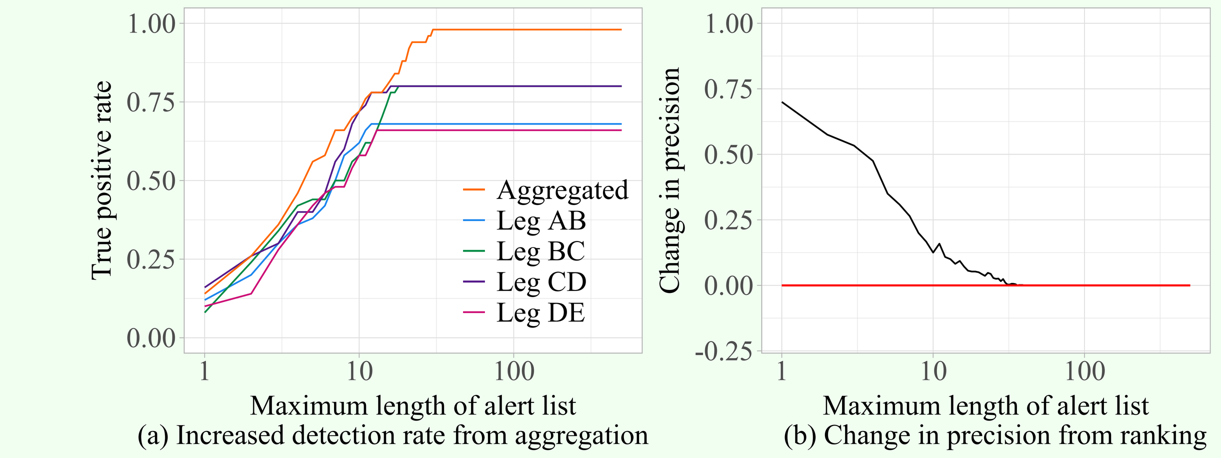
We use the non-exceedance probability from the GPD, θ_n, to quantify the severity of the outlier. Given that an outlier occurs, θ_n is the probability that the sum of threshold exceedances is at least as large at z_n. θ_n is given by:

$$\theta_n = F_{(\mu,\sigma,\xi)}(z_n) = \begin{cases} 1 - \left(1 + \frac{\xi(z_n-\mu)}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - exp\left(-\frac{(z_n-\mu)}{\sigma}\right) & \xi = 0 \end{cases}$$

We then construct an alert list using θ_n to rank the each outliers.

7. Outlier detection performance

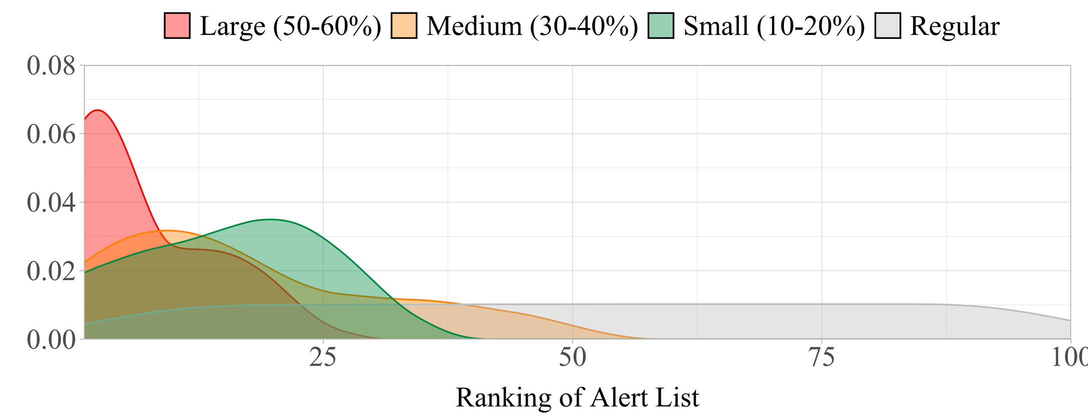
The true positive rate under the aggregated approach is higher than in any of the individual legs. When outlier demand affects multiple legs, the noise from other itineraries means that, when considering the leg's bookings in isolation, the outlier is not detected in every leg.



(a) Increased detection rate from aggregation (b) Change in precision from ranking

Precision is the fraction of classified outliers that are genuine. We look at the improvement in precision when ranking outliers as opposed to listing them in random order. The ranking results in improved precision, especially for short lists, and protects against false alerts.

8. Distribution of outliers in the alert list



We consider the distribution of outliers across the ranked alert list. Larger outliers are ranked higher. The higher variance of the medium-sized outliers can be explained by the fact that the ranking of a medium-sized outlier depends on the other types of outliers that occur.

9. References

- N. Rennie, C. Cleophas, A.M. Sykulski et al. *Identifying and responding to outlier demand in revenue management.* European Journal of Operational Research. 2021. doi.org/10.1016/j.ejor.2021.01.002
- S. López-Pintado and J. Romo. *On the Concept of Depth for Functional Data.* Journal of the American Statistical Association, 104(486):718-734, 2009.

n.rennie@lancaster.ac.uk http://lancaster.ac.uk/∼rennien/