Statistical modelling of river run and river network data

Craig Wilkie*, George Vazanellis*, Ruth O'Donnell, Claire Miller, Marian Scott, Adrian Bowman

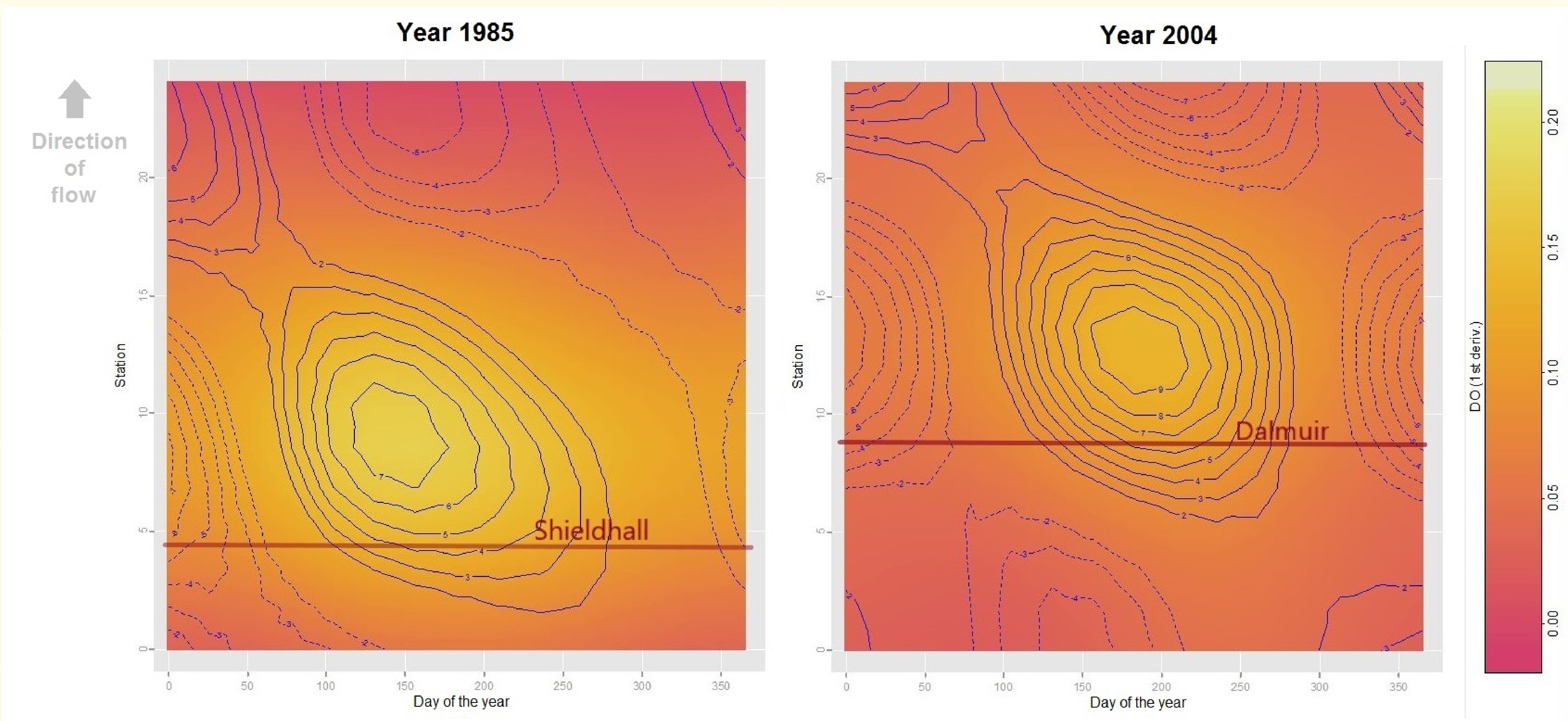
School of Mathematics and Statistics, University of Glasgow

*E-mail: craig.wilkie@glasgow.ac.uk, g.vazanellis.1@research.gla.ac.uk

Background: Rivers are vital parts of the freshwater ecosystem, sustaining aquatic and terrestrial life. However, their health is affected by diverse factors, including anthropogenic influences. This poster reviews the statistical methods developed by members of the environmental statistics group at the University of Glasgow, accounting for the complexities of river and network data, giving examples for the Rivers Clyde, Tweed and Trent. These examples demonstrate how the methods enhance our understanding of complex changes and drivers of river health.

River run data in the River Clyde

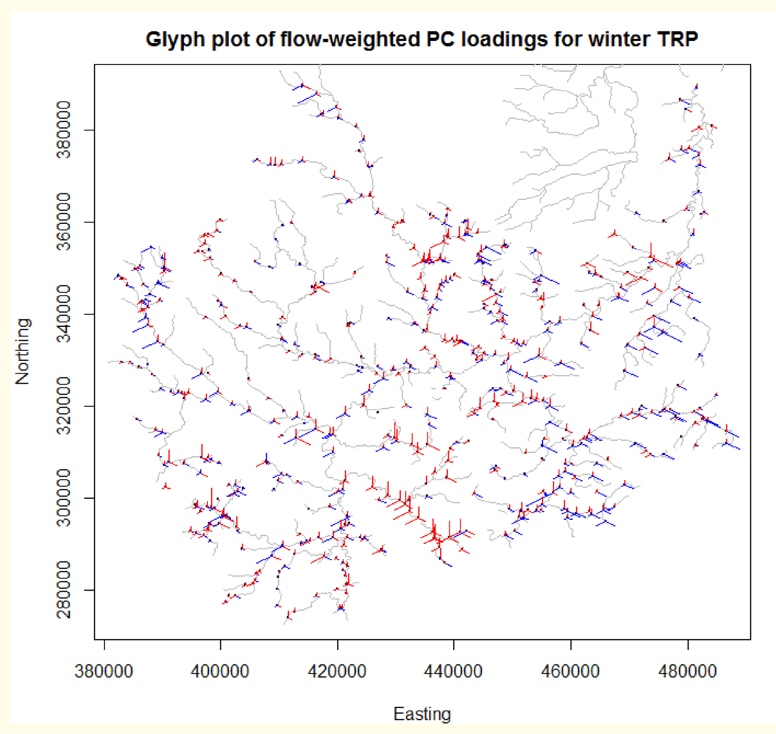
- The River Clyde has suffered from poor water quality over many years [5], due to anthropogenic pollutants, but dissolved oxygen (DO) levels have increased in the last four decades.
- "River run" data have been collected by SEPA at various depths and stations, approximately fortnightly since 1970.
- An ***additive mixed model*** was developed to assess the nonlinear effects of explanatory variables on DO levels.
- This complex spatiotemporal model uses ***partial derivatives*** to detect abrupt changes in DO levels.

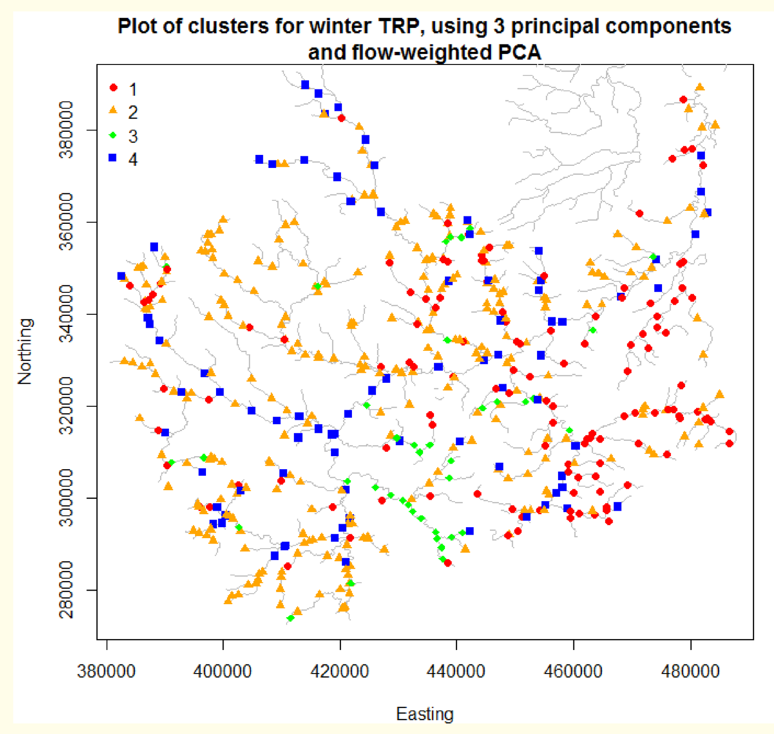


- These plots show snapshots of an `rpanel` [1] app, developed to assess the effects of two sewage treatment works upgrades on changes in DO levels (left: Shieldhall, 1985; right: Dalmuir, 2004). These show smooth changes in DO with respect to the interaction between year, station and day of the year.
- The contours in both plots show positive changes in DO levels in summer, downstream of the upgraded works (locations shown with horizontal lines), showing the power of the model to detect the impacts of external events on improvements in DO.

Spatiotemporal patterns in TRP and TON in the River Trent network

- There are 564 sampling stations measuring TRP and TON in the River Trent catchment, operated by Environment Agency.
- Can this network be reduced, retaining sites contributing most to explaining variation in spatiotemporal patterns?
- ***Flow-directed principal component analysis*** [2] identifies sites with common temporal patterns, accounting for the river network structure of the data.

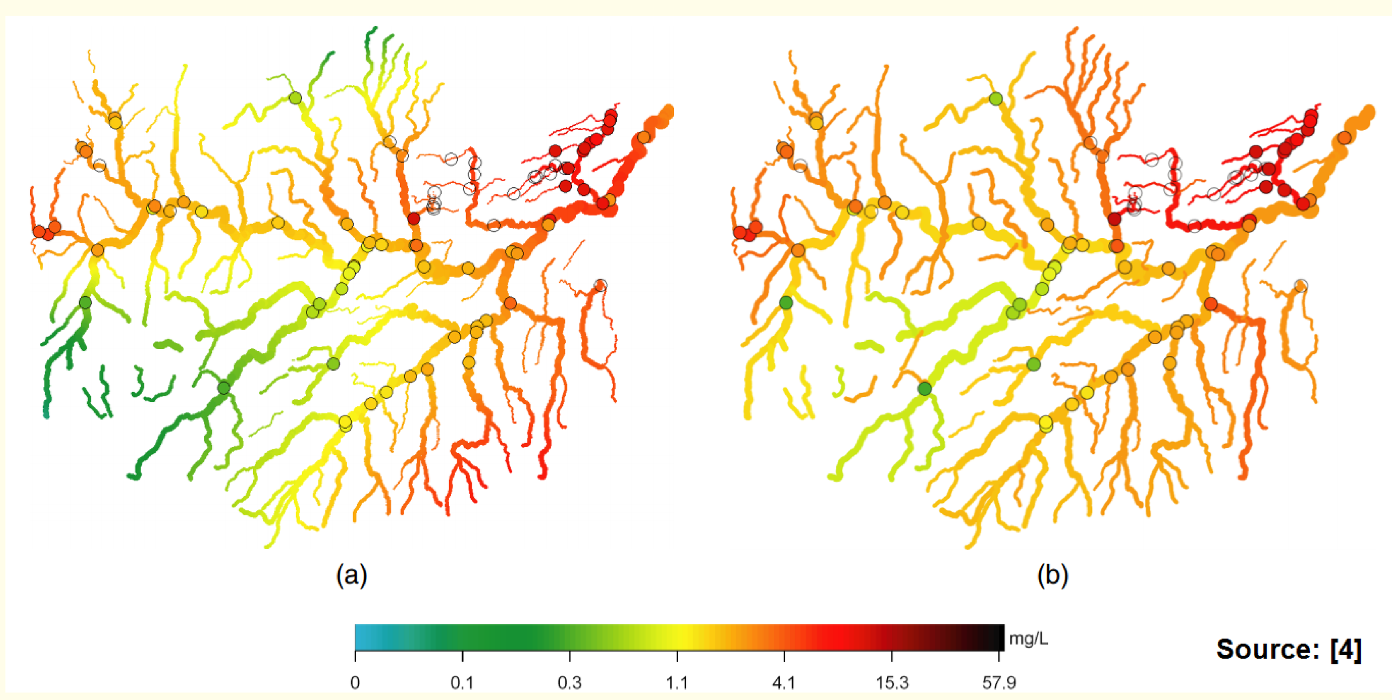




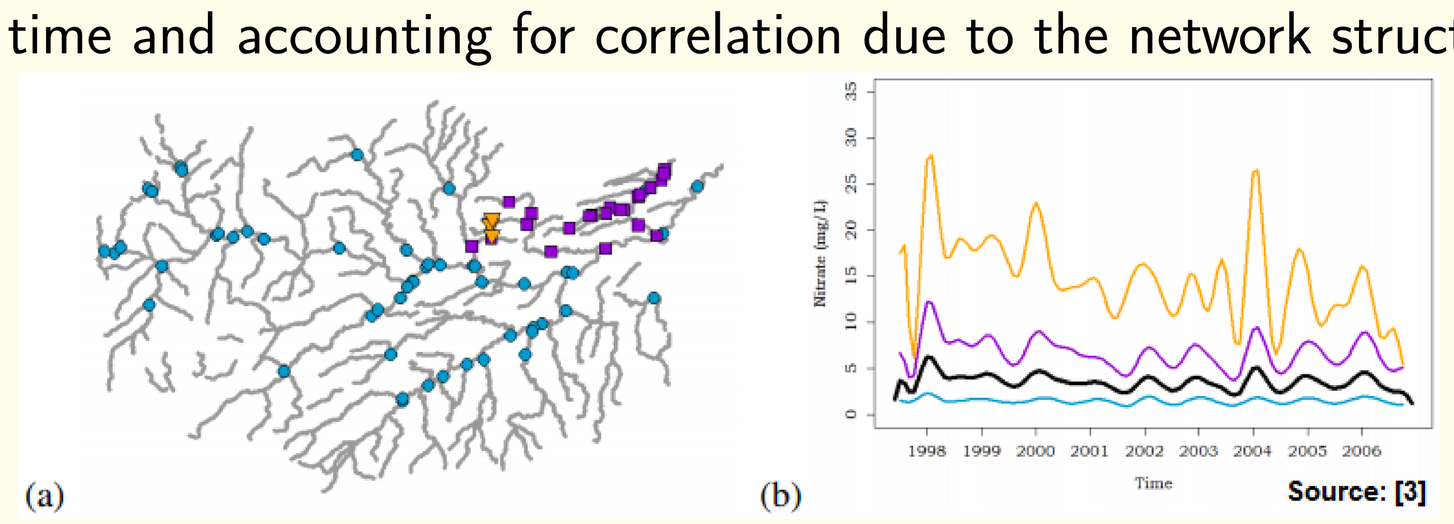
- The left-hand plot shows loadings for PCs 1 to 3, with colour representing sign and length representing magnitude. We can identify groups of locations with similar temporal patterns (i.e. those with loadings of the same sign and similar magnitudes).
- The right-hand plot shows clusters of these loadings, from k-means clustering. We can identify redundant sites, e.g. those in the same cluster providing similar information.
- **Reducing networks:** Stratified sampling, using these clusters as strata, leads to lower standard errors of predictions than random sampling ignoring the PCA results.

Prediction of nutrients across the River Tweed network

- Nitrate levels in the River Tweed basin are monitored by SEPA.
- Modelling these data over space and time is challenging: distance is measured along complex paths and relative flows into and out of confluences must be accounted for.
- ***Flexible regression*** [4] models the underlying spatial trends, allowing the fitting of smooth but flexible patterns, including spatiotemporal patterns, seasonal terms and interactions.



- Figure (a): Predictions based upon Euclidean distance, with data overlaid. Figure (b): Predictions based upon flow-weighted distance. Flow-weighting results in more sensible predictions, as only connected sites are related.
- ***Spatially-weighted functional clustering*** [3] enables the grouping of stations with similar trends and seasonal patterns, through treating data as observations of smooth functions over time and accounting for correlation due to the network structure.



- Figure (a): De-trended stream distance covariance-weighted clustering results. Figure (b): Cluster mean curves and overall mean curve (black).

References [1] Bowman A, Crawford E, Alexander G, Bowman R. *J. Stat. Softw.* 2007; **17**:9; [2] Gallacher K, Miller C, Scott M, Willows R, Pope L, Douglas J. *Environmetrics* 2017;**28**:e2434; [3] Haggarty R, Miller C, Scott M. *JRSSC* 2015;**64**:491–506; [4] O'Donnell D, Rushworth A, Bowman A, Scott M, Hallard M. *JRSSC* 2014;**63**:47–63; [5] SEPA. *State of Scotland's Environment 2006.*

Acknowledgements The Trent and Tweed examples are partly based upon work by Kelly Gallacher, David O'Donnell and Alastair Rushworth, as part of the Environmental Statistics group at the University of Glasgow. Data were provided by SEPA and the Environment Agency.