Generalized additive mixed models for archaeological network data

Nicolas Gauthier

School of Human Evolution and Social Change, Arizona State University

What are GAMMs?

Generalized additive models (GAMs) and their variant Generalized additive *mixed* models (GAMMs) are a flexible family of regression models well-matched to the complexity of the archaeological record.

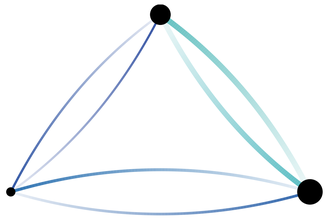
Let's break it down:

- *Generalized* – Model non-normal distributions such as counts or proportions
- *Additive* – Combine linear and non-linear relationships
- *Mixed* – Control for network, spatial, and temporal autocorrelation

What can I do with them?

In archaeology, they're particularly useful for fitting so-called "maximum-entropy" style spatial interaction models. Most archaeological network represent a *flow* of something – goods, information, or people – between spatially-structured populations. Spatial interaction models separate influences from the origin site, the destination site, and the space between them:

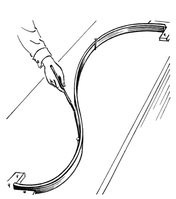
$$flow = f\,(origin) \times f\,(destination) \times f\,(distance)$$



The difficult part comes when we actually have to define all the $f\,()$s. Economic geographers typically estimate spatial interaction models with generalized *linear* models, so must choose the functional forms ahead of time. This is where splines come in handy.

How do they work?

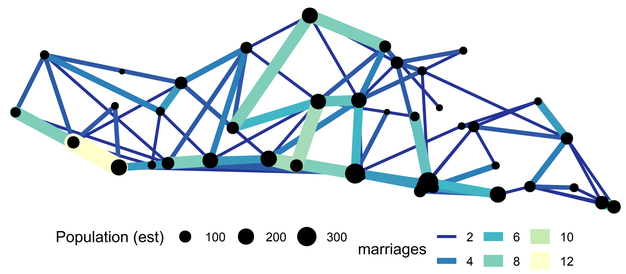
Splines are what make GAMs so useful. Real-world splines are flexible strips of metal or wood used to draw curves. Mathematical splines are complex curves made of many smaller, simpler curves. GAMs limit overfitting by penalizing the "wiggliness" of the function, just as you'd use a more rigid piece of wood to draw a smoother curve.



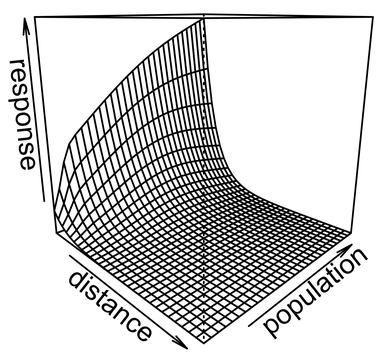
Case Studies

Chumash marriage networks

First we'll look at a dataset of XX840XX marriages between 41 Chumash villages near Santa Barbara, California, as recorded by Spanish missionaries. Populations are estimated from baptism records.



We can model these data by regressing the count of marriages between each village pair on the product of their populations and the distance between them.



The result is just what we'd expect if we had used a GLM without splines: power functions with a negative exponent for distance and a superlinear exponent for population.

Ceramic distribution in Roman Britain

Next is a dataset of Late Romano-British pottery. We have data for the percentage of the ceramic assemblage at 30 sites that were produced at Oxford or New Forest, and want to understand something about how these pots were moved over space.

Figure 1: Percentages of late Romano-British pottery produced in Oxford and New Forest.

We can look at the spline functions fit by the GAM. These suggest that overland travel is subject to logarithmic distance decay, but water transport seems to have no such constraints.

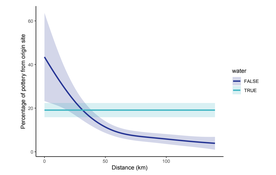


Figure 2: Estimated distance decay functions, with and without water transport.

Finally, by examining the residuals of the basic spatial interaction model, we can focus on the links that are unusually stronger or weaker than expected for further study.

Figure 3: Residuals from a distance-only interaction model. Red ties are stronger then expected, blue ties weaker.

What else?

GAMs can fit many different kinds of functions, with spatial, temporal, and network autocorrelation. Many response types such as counts, percents, binary, categories, ordered categories, hazard models. Use the package `mgcv` in `R`. Bayesian implementations in `brms`. Scan this QR code, or go to [github.com/nick-gauthier/gam-networks] to find the code and data used to make this poster, and more detailed worked examples in R.