Odds and sods

Gavin L. Simpson

U Adelaide 2017 · Feb 13–17 2017

Spectral analysis & wavelets

Spectral analysis

Lomb-Scargle power spectrum

Spectral analysis requires evenly spaced data — at least the classical methods do

Could use interpolation to make the data unevenly spaced — introduces strong artifacts into the data

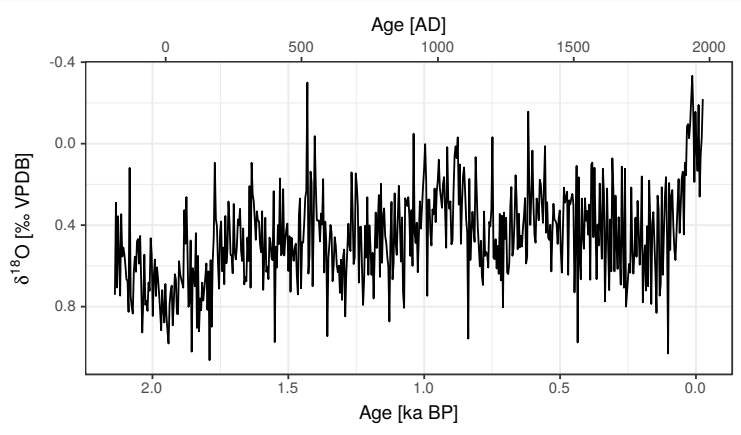
Lomb evaluated the sine and cosine functions only at the observed time points with an offset τ, which makes the $P_x(\omega)$ independent of the time points being shifted around

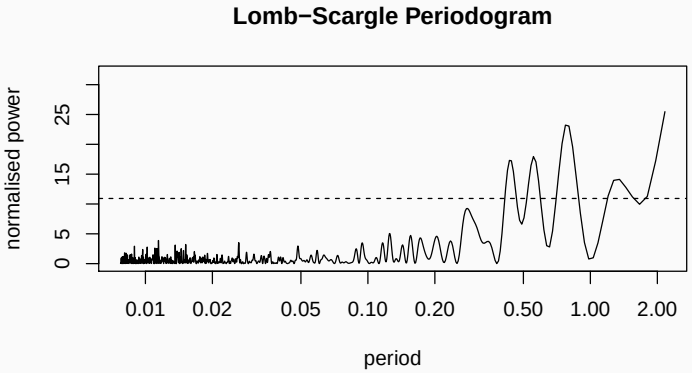
Scargle showed that the choice of τ used by Lomb has the effect that $P_x(\omega)$ obtained using Lomb's method are identical to the least squares fit

$$t_t = A\cos\omega t + B\sin\omega t$$

(ω is the angular frequency)

Lomb-Scargle power spectrum




Lomb−Scargle Periodogram
normalised power
0
5
15
25
0.01
0.02
0.05
0.10
0.20
0.50
1.00
2.00
period

Wavelets

The power spectrum is often a global analysis of a time series; shows the power associated with periodicities over the entire time series

In many time series these periodicities can wax and wane, gaining and loosing strength during different parts of the series

Can look at the power spectrum in small chunks of a series — evolutionary power spectrum

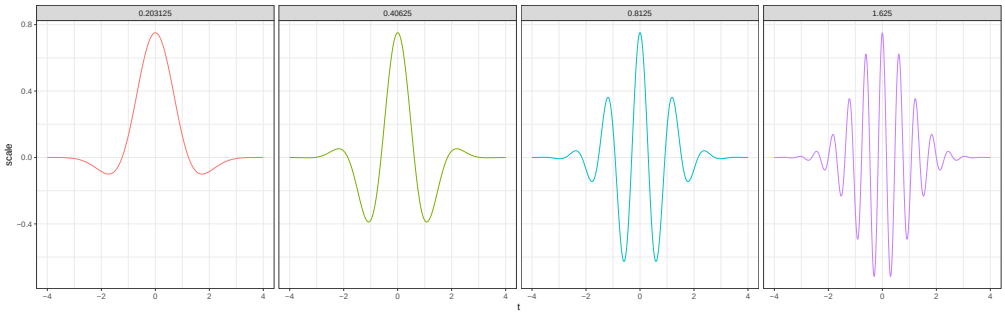
A better approach is to use wavelets

Wavelets

Wavelets are a local transform, describing the intensity of pattern at different scales at a particular location

Wavelets scale a *mother wavelet*, varying it's frequency, & applying it locally to a time series

Coefficients are estimated for the wavelet at each scale and frequency, evaluating the local power at particular frequencies



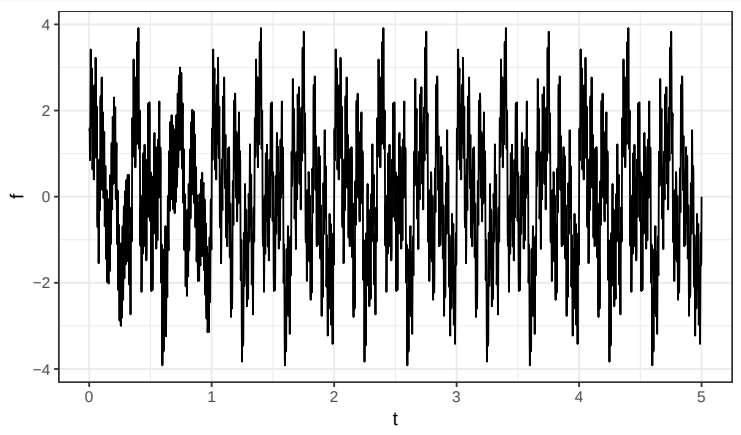
Wavelets

Originally, wavelets were defined for regularly-spaced data

Newer developments, such as “lifting”, allow for wavelets to be applied to irregular series

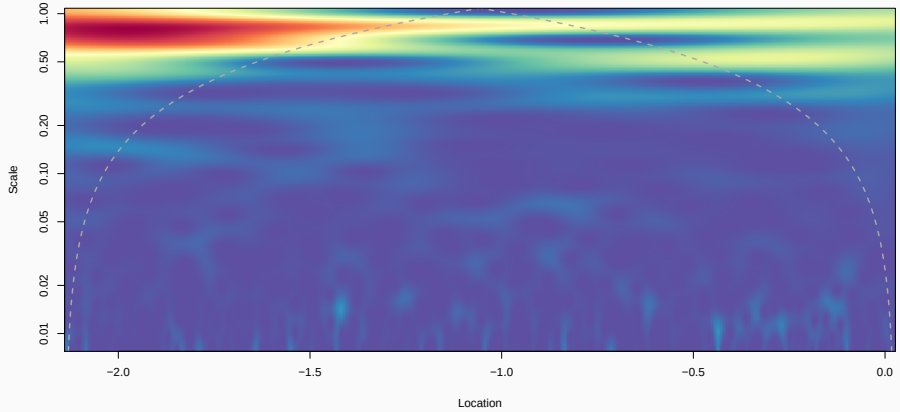
One implementation in R is via the **mvcwt** package

Wavelets example



Wavelets example

```
Warning: executing %dopar% sequentially: no parallel backend registered
```



Classification

Classification

Classification is a supervised learning problem

We have known group labels for each of the *n* samples in the training data

In addition we have a series of variables we wish to use to predict the group labels

Supervised because we know the groups — contrast that with cluster analysis where we wish to find groups

Classic methods include logistic regression and linear discriminant analysis (LDA)

Classification

Classification is an important problem in machine learning and has attracted considerable interest from statisticians

Major advances include

- classification trees
- random forests
- boosted trees

Classification trees

Response is a categorical variable

Search through all variables and all possible locations for a split to find the split that best describes the response

We want splits that result in the most pure nodes as possible

Once we make one split, we repeat the process on the two parts recursively to make the nodes more pure

Use cross validation to decide how many splits are need to predict the groups, without over-fitting

Classification trees

Neil Rose (UCL) collected Spheroidal Carbonaceous Particles from power stations the burned different fuels

- Coal (3000)
- Oil (1000)
- Oil shale (2000)

The SCP surface chemistry was determined on 6000 particles

Can we predict the fuel source from particle elemental composition?

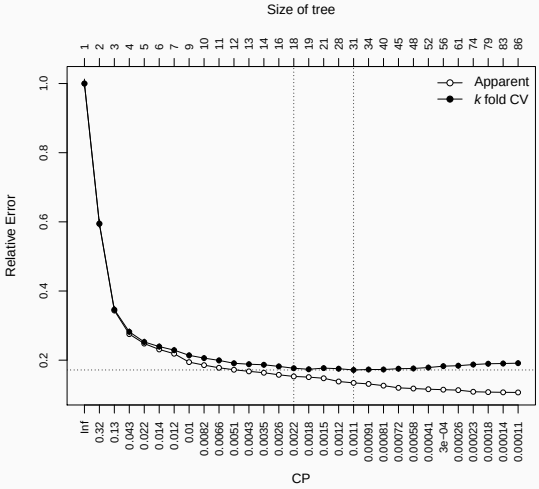


Figure 1: Cost-complexity pruning the fitted tree

Classification tree

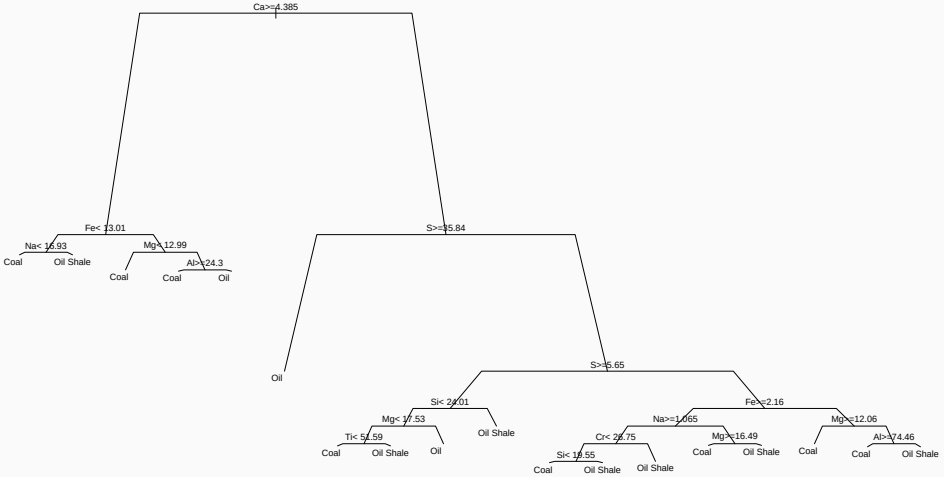


Figure 2: Fitted classification tree

Classification tree

Overall error rate is 0.053

	Coal	Oil	Oil-Shale	Error rate
Coal	2871	49	118	0.055
Oil	1	938	11	0.028
Oil-Shale	113	13	1817	0.063

Random Forests

Trees are high-variance classifiers — had we collected a different data set, we might get different splits due to sampling noise

Can improve trees by bagging; fit m trees, each using a bootstrap sample from the original data & count votes over all m trees

Leo Breiman showed that you can do better than bagging by adding more randomness to the tree building by forming individual splits only considering a small subset of predictor variables

Each split uses a different set of randomly selected variables — trees are grown large without pruning

Fit an entire forest of trees and count the votes over all m trees in the forest

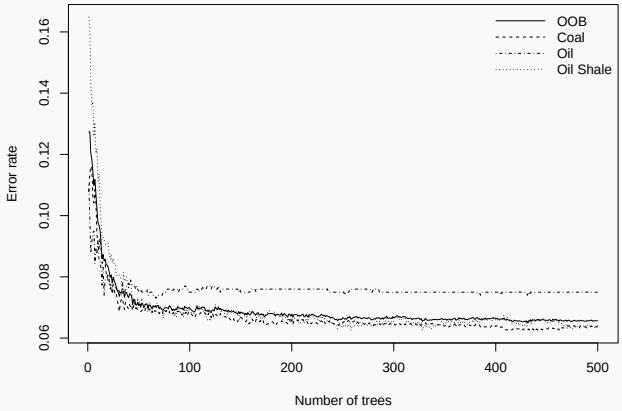


Figure 3: Oout-of-bag error rates as trees added to forest

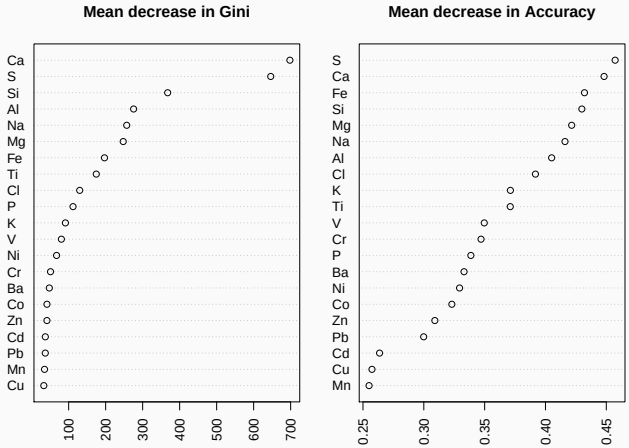


Figure 4: Which variables contribute most to node purity & model accuracy?

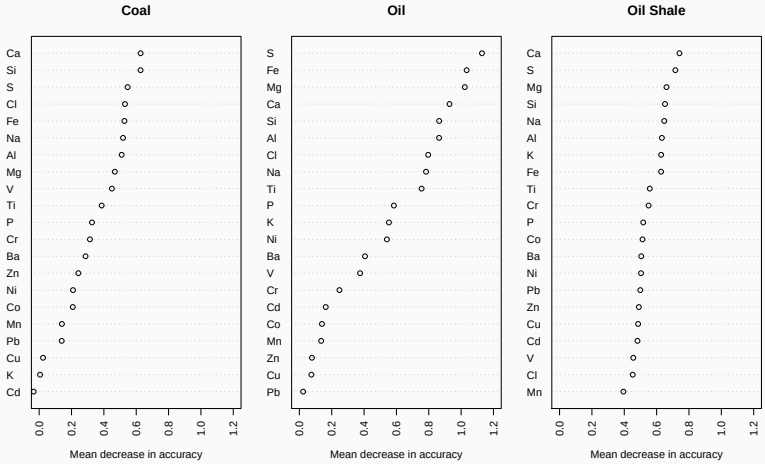


Figure 5: ...same but now for specific groups

Boosted trees

Boosted trees are a similar idea to Random Forests but differ in several important ways

1. individual trees in the ensemble are small trees — often stubs with 1 split
2. each subsequent tree added to the ensemble tries to fit a weighted version of the response
    - samples poorly predicted by the current set of trees get higher weight
    - hence subsequent trees aim to predict the hardest to predict of the observations
3. we learn slowly; we don't take the full prediction but we down-weight them
    - each tree adds only a little bit to the predictive power

Surprisingly, this works and boosted trees are one of the better classification tool available...

...but there is more to tune than with Random Forests

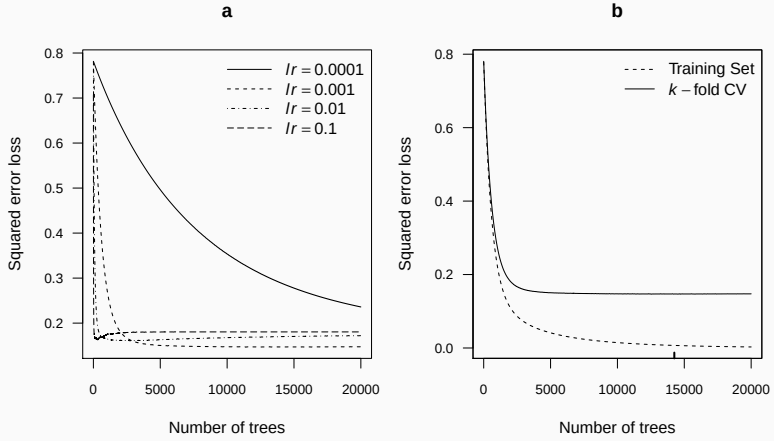


Figure 6: ...same but now for specific groups

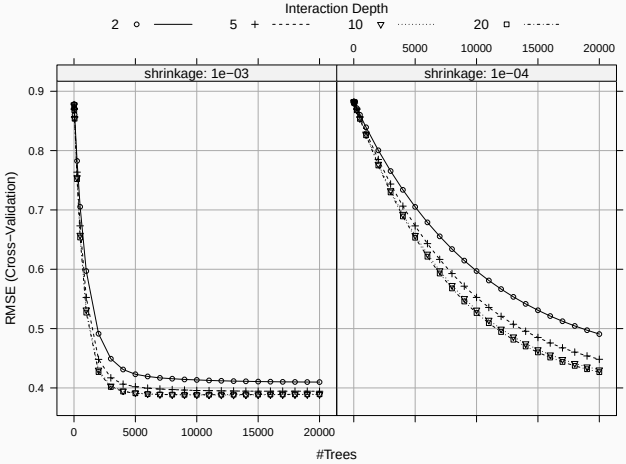


Figure 7: ...same but now for specific groups

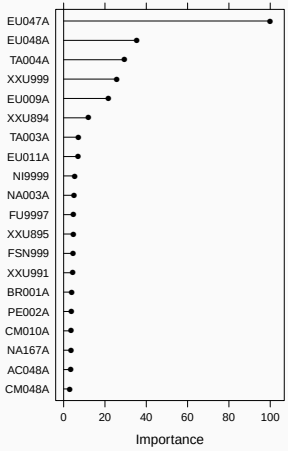


Figure 8: ...same but now for specific groups

Explaining one time series with another

Using one time series to explain another

ARIMA models can include exogenous variables that predict the time series of interest

For unevenly sampled data we can use regression models, esp GAMs, to model the effect of one variable (series) on another

Two examples from Simpson & Anderson (2009) Limnology & Oceanography

1. Model effect of climate on diatom composition in Kassjön, a varved lake in N Sweden
2. Model the effects of acid rain and climate on a diatom record from Fionnaraich, NW Scotland

Using one time series to explain another — Kassjön

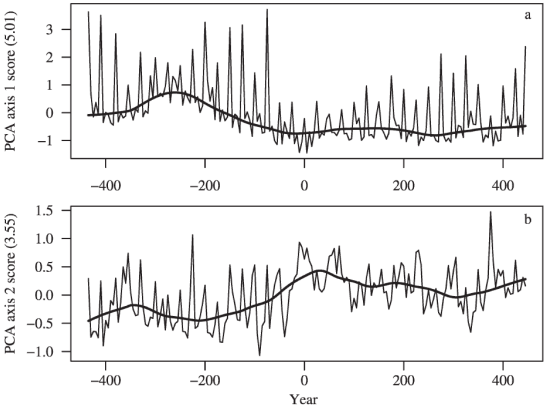


Fig. 1. Time series of (a) axis 1 and (b) axis 2 scores for the PCA of the Hellinger-transformed diatom data from Kassjön. The number in brackets on the *y*-axis label is the % variance explained by each axis. The thick lines are LOESS smoothers fitted through the observations to highlight trends in an exploratory manner.

Figure 9: ...same but now for specific groups

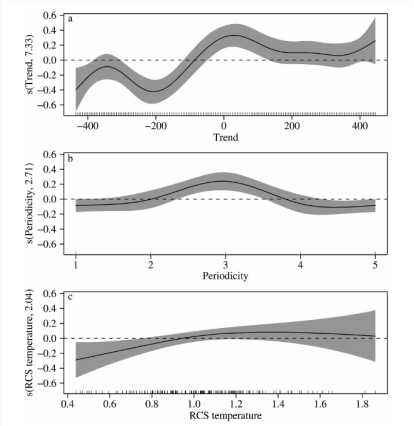


Fig. 2. The fitted smooth functions for (a) trend, (b) periodicity, and (c) RCS temperature from the final AMM for the Kassjön PCA axis 2 scores. The gray bands are approximate 95% pointwise confidence intervals on the fitted functions. The tick marks inside the panels on the *x*-axis show the distribution of observed values for the two covariates. The numbers in brackets on the *y*-axis (7.33, 2.71, and 2.042 for trend, periodicity, and RCS temperature, respectively) are the effective degrees of freedom for each smooth function.

Figure 10: ...same but now for specific groups

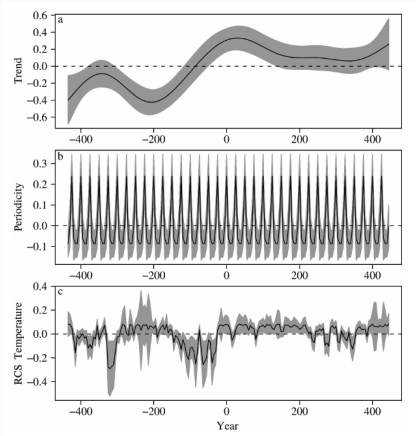


Fig. 3. The contribution of (a) trend, (b) periodicity, and (c) RCS temperature to the fitted diatom PCA axis 2 scores for the final Kassjön model. The gray band is an approximate 95% pointwise confidence interval on the contribution. Where the band includes the dashed zero line, the contribution of the covariate is not statistically significantly different from the intercept.

Figure 11: ...same but now for specific groups

Using one time series to explain another — Fionnaraich

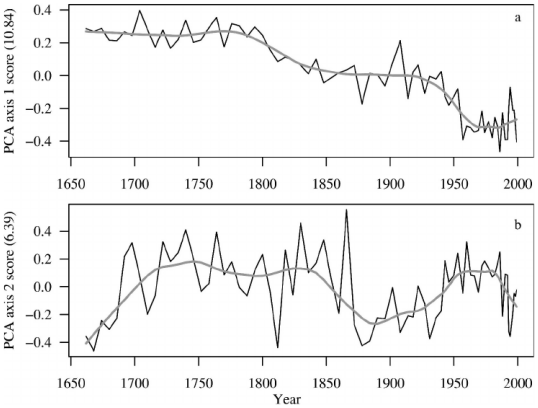


Fig. 5. Time series of (a) axis 1 and (b) axis 2 scores for the PCA of the Hellinger-transformed diatom data from Loch Coire Fionnaraich (LCFR). The number in brackets on the *y*-axis label is the % variance explained by each axis. The thick lines are LOESS smoothers fitted through the observations to highlight trends in an exploratory manner.

Figure 12: ...same but now for specific groups

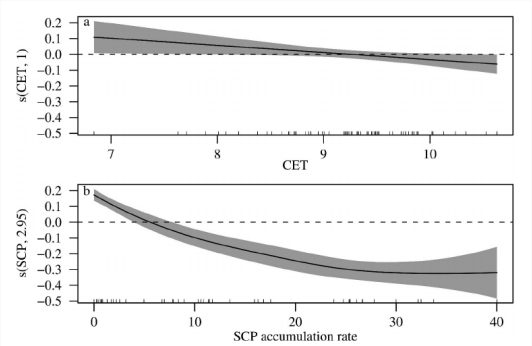


Fig. 6. The fitted smooth functions for (a) CET and (b) SCP accumulation rate from the final AMM for the LCFR PCA axis 2 scores. The gray bands are approximate 95% pointwise confidence intervals on the fitted functions. The tick marks inside the panels on the x-axis show the distribution of observed values for the two covariates. The numbers in brackets on the y-axis (1 and 2.95 for CET and SCP, respectively) are the effective degrees of freedom for each smooth. A value of 1 indicates a linear function.

Figure 13: ...same but now for specific groups

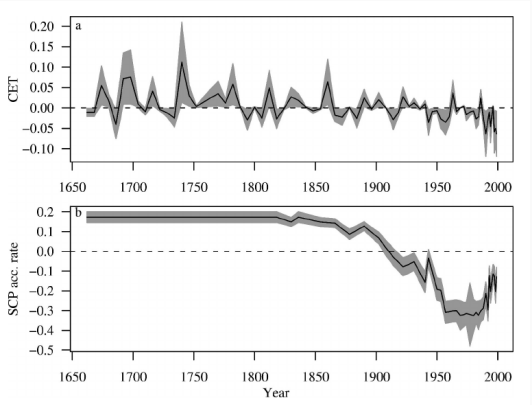


Fig. 8. The contribution of (a) CET and (b) SCP accumulation rate to the fitted diatom PCA axis 2 scores for the final LCFR model. The gray band is an approximate 95% pointwise confidence interval on the contribution. Where the band includes the dashed zero line, the contribution of the covariate is not statistically significantly different from the intercept.

Figure 14: ...same but now for specific groups

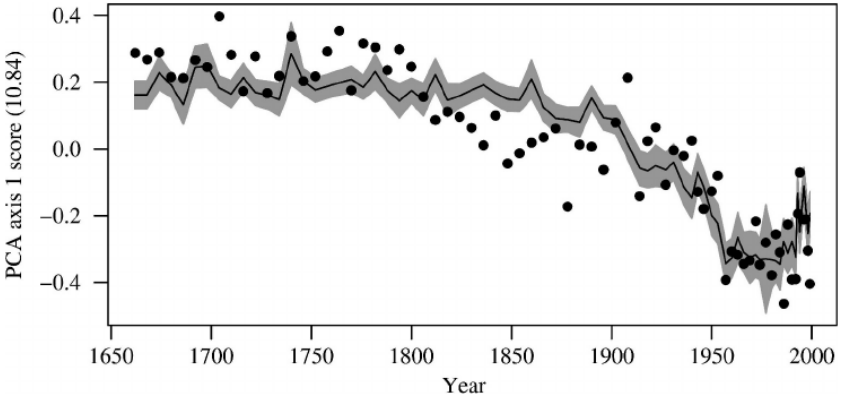


Fig. 7. Observed and AMM fitted values for PCA axis 2 scores for the LCFR core. The gray band is an approximate 95% pointwise confidence interval on the fitted values.

Figure 15: ...same but now for specific groups

Re-use