



Data has a better idea

Palaeolimnologists must rethink their approach to data analysis

Gavin Simpson

Palaeo data are time series

Interest in changes in the data over time — implies the estimation of trends in data

Commonly, trend detection involves eye-balling the data

Fundamentally irreproducible — poor science

Conflates signal and noise

What statistical analysis is done is often wrong

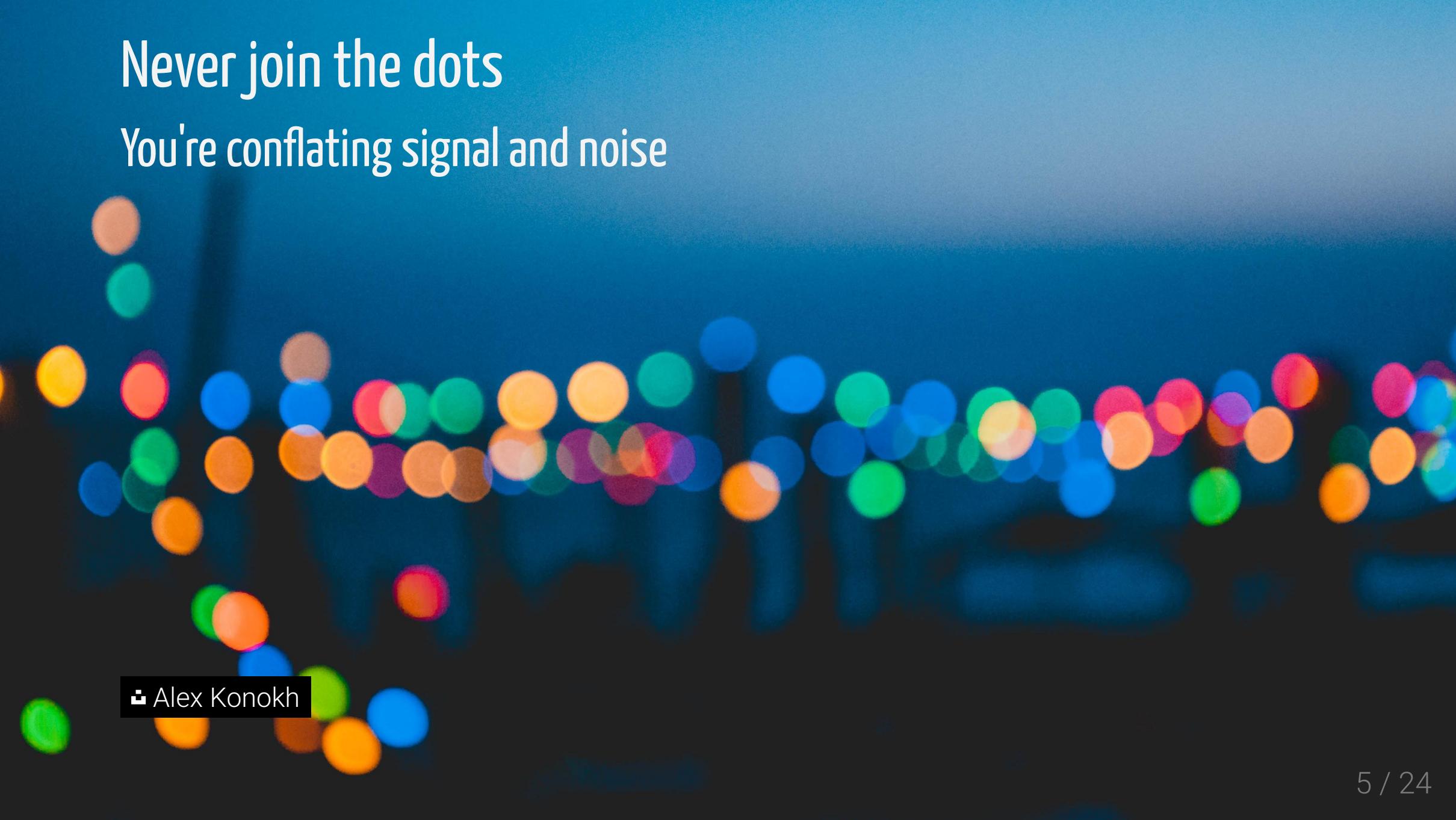


Time waits for no palaeolimnologist



Never join the dots

You're conflating signal and noise



Alex Konokh

Don't correlate time series — Spurious correlation

Two time series may be correlated because

- of their relationship with a third confounding variable
- of nothing, they're just both going up or down for no common reason

Correlation is simply a measure of the strength of association between two series

Correlation ≠ Causation

Vol. LXXXIX.]

[Part I.

JOURNAL
OF THE ROYAL STATISTICAL SOCIETY.
JANUARY, 1926.

WHY DO WE SOMETIMES GET NONSENSE-CORRELATIONS BETWEEN TIME-SERIES?—A STUDY IN SAMPLING AND THE NATURE OF TIME-SERIES.

THE PRESIDENTIAL ADDRESS OF MR. G. UDNY YULE, C.B.E., M.A., F.R.S., FOR THE SESSION 1925-26. DELIVERED TO THE ROYAL STATISTICAL SOCIETY, NOVEMBER 17, 1925.

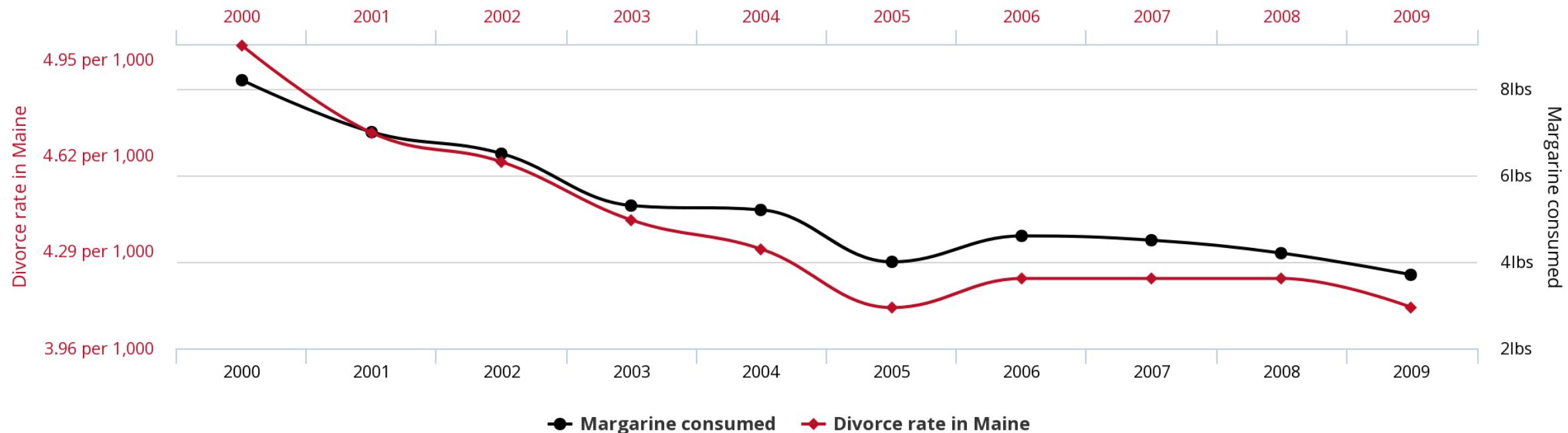
| | PAGE |
|--|------|
| Section I.—The problem | 2 |
| “ II.—The correlation between simultaneous segments of two variables that are simple harmonic functions of the time, of the same period but differing by a quarter-period in phase; and the frequency-distribution of correlations for random samples of such segments.... | 6 |
| “ III.—Deductions from Section II: classification of empirical series | 12 |
| “ IV.—Experimental investigations | 30 |
| “ V.—Serial correlations for Sir William Beveridge's index-numbers of wheat prices in Western Europe; and for rainfall at Greenwich | 41 |
| Appendix I.—The correlations between segments of two sine-curves of the same period, etc. | 54 |
| “ II.—The relations between the serial correlations of a sum-series and of its difference series, when the series may be regarded as indefinitely long.... | 57 |

THE problem which I have chosen as the subject of my Address is one that puzzled me for many years. The lines of solution only occurred to me two or three years ago, and I thought that I could not do better than endeavour to work them out during the Session 1924-25—time and opportunity having hitherto been lacking—and utilize them for the present purpose. As often happens, the country

VOL. LXXXIX PART I.

B

Divorce rate in Maine correlates with Per capita consumption of margarine



Source: Tyler Viglen www.tylervigen.com/spurious-correlations

Non-parametric \neq no assumptions

You can calculate r , ρ , or τ on time series – but what do they mean?

When you *test* r , ρ , or τ you need theory & assumptions

Key is independent observations

Palaeo data almost surely violate those assumptions

Vol. LXXXIX.]

[Part I.

JOURNAL
OF THE ROYAL STATISTICAL SOCIETY.
JANUARY, 1926.

WHY DO WE SOMETIMES GET NONSENSE-CORRELATIONS BETWEEN TIME-SERIES?—A STUDY IN SAMPLING AND THE NATURE OF TIME-SERIES.

THE PRESIDENTIAL ADDRESS OF MR. G. UDNY YULE, C.B.E., M.A., F.R.S., FOR THE SESSION 1925-26. DELIVERED TO THE ROYAL STATISTICAL SOCIETY, NOVEMBER 17, 1925.

| | PAGE |
|--|------|
| Section I.—The problem | 2 |
| “ II.—The correlation between simultaneous segments of two variables that are simple harmonic functions of the time, of the same period but differing by a quarter-period in phase; and the frequency-distribution of correlations for random samples of such segments.... | 6 |
| “ III.—Deductions from Section II: classification of empirical series | 13 |
| “ IV.—Experimental investigations | 30 |
| “ V.—Serial correlations for Sir William Beveridge's index-numbers of wheat prices in Western Europe; and for rainfall at Greenwich | 41 |
| Appendix I.—The correlations between segments of two sine-curves of the same period, etc. | 54 |
| “ II.—The relations between the serial correlations of a sum-series and of its difference series, when the series may be regarded as indefinitely long.... | 57 |

THE problem which I have chosen as the subject of my Address is one that puzzled me for many years. The lines of solution only occurred to me two or three years ago, and I thought that I could not do better than endeavour to work them out during the Session 1924-25—time and opportunity having hitherto been lacking—and utilize them for the present purpose. As often happens, the country

VOL. LXXXIX PART I.

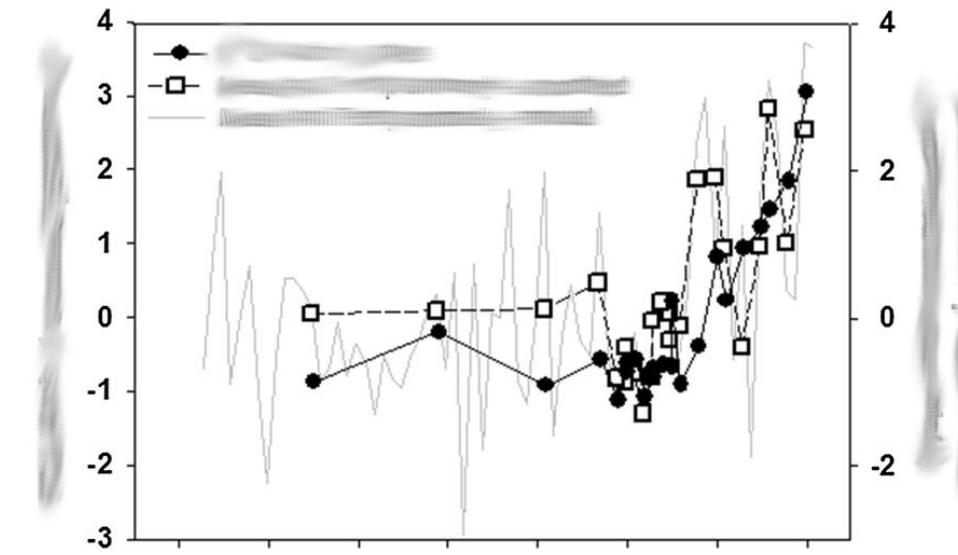
B

Never smooth data you input to a statistical test

Ever — not once

It makes the test result *moar significant*

You just threw away the noise!



Loess must die



Loess must die

Loess is a scatterplot smoother – for EDA

The fit is controlled by the *span*

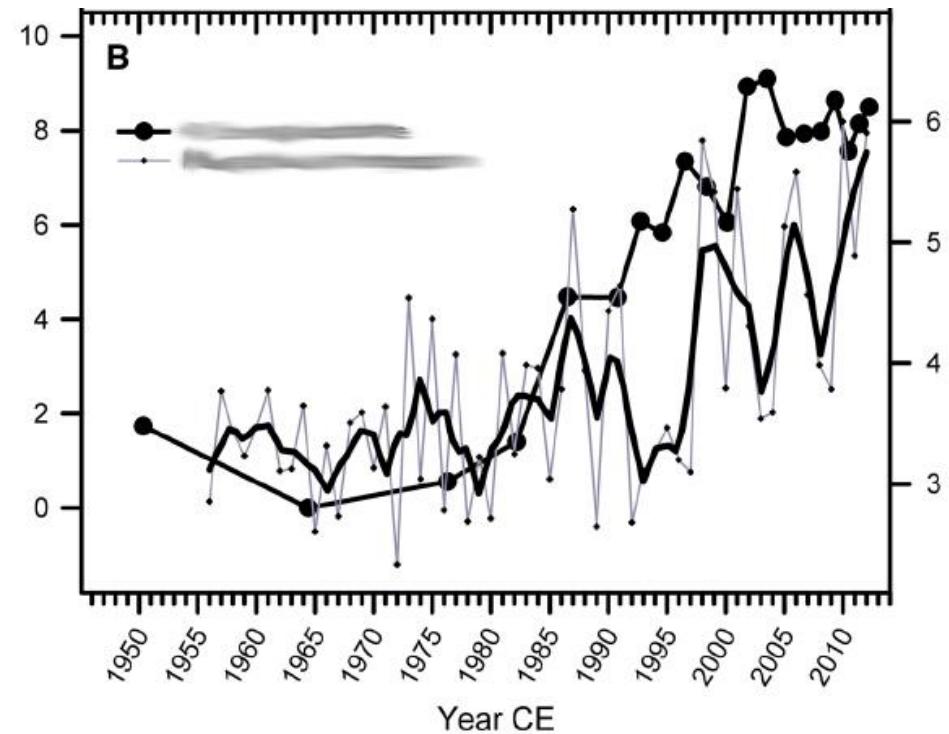
Must choose *span* appropriately

Difficult (impossible?) to do properly...

As a result often chosen subjectively!

Plain wrong – don't do it!

Don't use Loess for inference ... exploratory
data analysis only



What are we to do...?



© Ross Findon

$$\frac{dS}{dt} = \frac{1}{T_{act}} - \varphi(\mu \cdot N_0) (1-\varepsilon S) S + \frac{\nu e}{T_n} - \frac{N}{T_p}$$

$$\frac{dS}{dt} = T_b \varphi(\mu \cdot N_0) (1-\varepsilon S) S + \frac{\nu e N}{T_n} - \frac{S}{T_p}$$

$$\frac{S}{P_t} = \frac{T_p k_0}{T_{act} \eta n c}$$

$$TS < \frac{1}{\varepsilon}$$

$$N = N_0 e^{-\lambda t}$$
$$P_t = (m)$$

Model your data

Many models available for time series but palaeo data are often unhelpful

1. uneven spacing of observations in time (typically)
2. compaction, variable accumulation rates → non-constant variance

Can't use typical statistical time series models

But we can use generalized additive models

Generalized additive model

Linear trend model

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$$

GAM

$$\mathbb{E}(y_i) = \beta_0 + f(x_i)$$

$f(\mathbf{x})$ is the trend – we assume trend is smooth

How do we estimate f ?



WIGGLY
-THINGS-

Splines

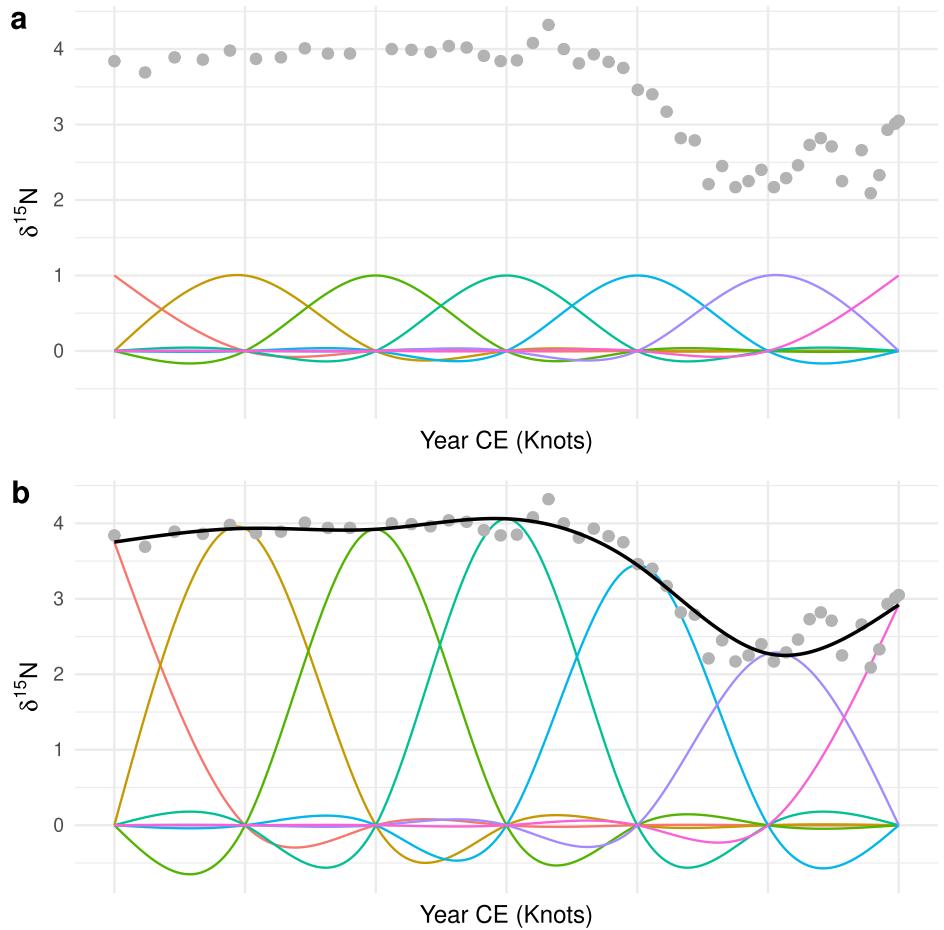
Splines are smooth functions

Made of little basis functions

Estimate β_k for each basis function

Sum the weighted basis functions at each time point

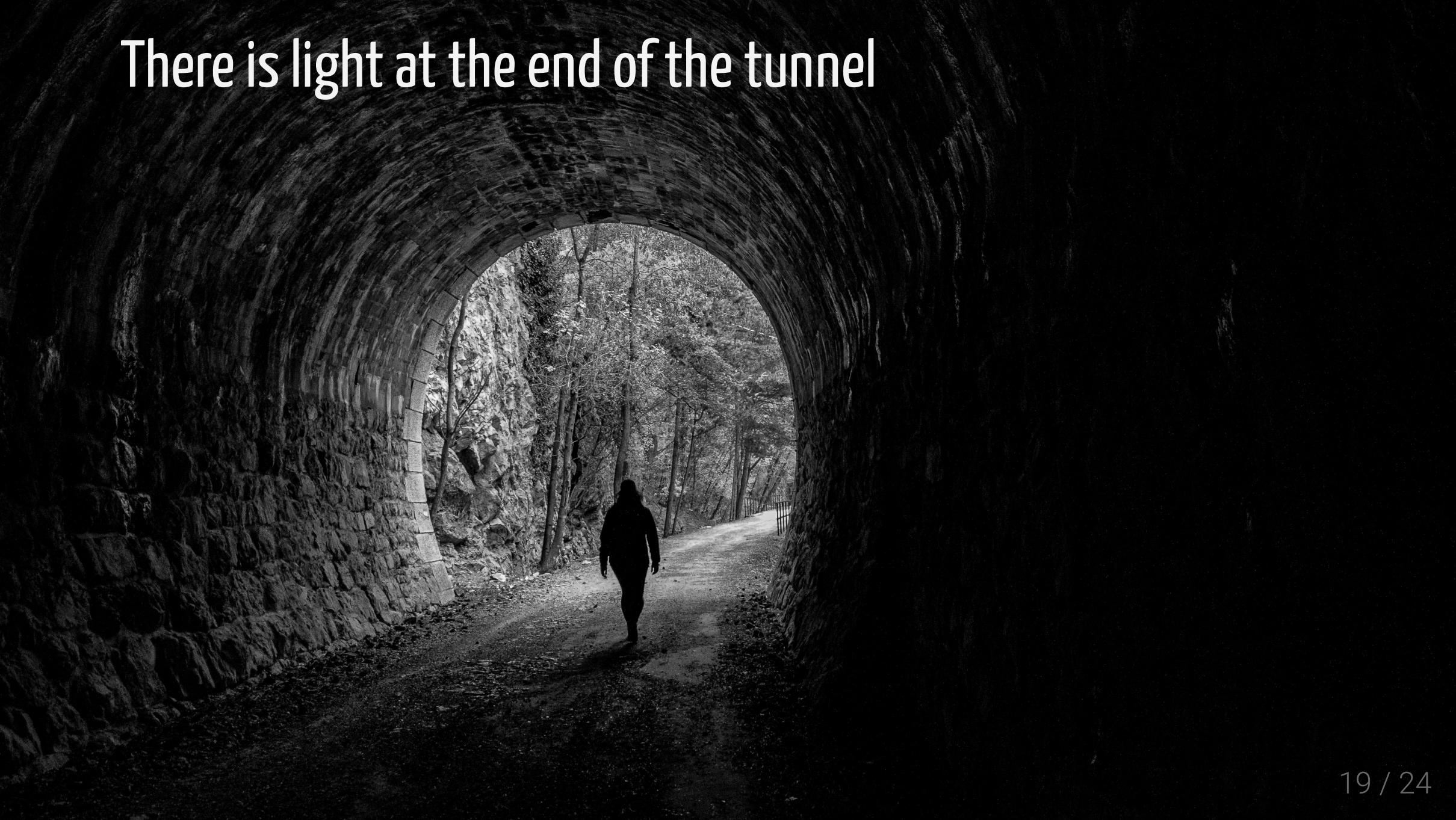
Use a penalty on **wigginess** to avoid overfitting



GAMs are a gateway drug



There is light at the end of the tunnel



Training

Opener: Cylinders & Triangular Prism

Area = $\frac{1}{2} \times b \times h$

Volume = $b \times h \times l$

Surface Area = $b + h + l + \sqrt{b^2 + h^2}$

Example: Find the surface area of a rectangular prism with dimensions $l=100$, $w=100$, and $h=100$.

Solution:

| | |
|-------------------------------------|-----------|
| Area = 100×100 | $= 10000$ |
| Perimeter = $100 + 100 + 100 + 100$ | $= 400$ |
| Surface Area = $10000 + 400$ | $= 10400$ |

Training

Palaeolimnology was a qualitative science

Key figures dragged us into the world of quantitative palaeolimnology

Training and software were critical for this revolution

We have the software — R

A renewed need for training in modern methods

What should statistical palaeolimnology training look like?

Conclusions

Most palaeolimnological data are time series

- we're interested in estimating trends in those data
- we're interested in comparing trends between series

Rarely try to estimate those trends statistically

When we do, we often do it inappropriately

We should be modelling our data using statistical models

GAMs are a (relatively) simple model that we could use to model palaeo time series

Funding



**NSERC
CRSNG**

Want to know more...?

Paper – [doi:10/gfrc4p](https://doi.org/10/gfrc4p)

Blog – www.fromthebottomoftheheap.net

Slides – bit.ly/pals-stats