

An introduction to Dynamic Factor Analysis

Applied Time Series Analysis for Ecologists

Stockholm, Sweden

24-28 March 2014

Finding common patterns in data

- We have learned several forms of models for analyzing ts
- General idea is to reduce ts to trends, seasonal effects, and stationary remainder
- If many ts involved, we could model each separately
- But, common (environmental) drivers may create common patterns among some ts

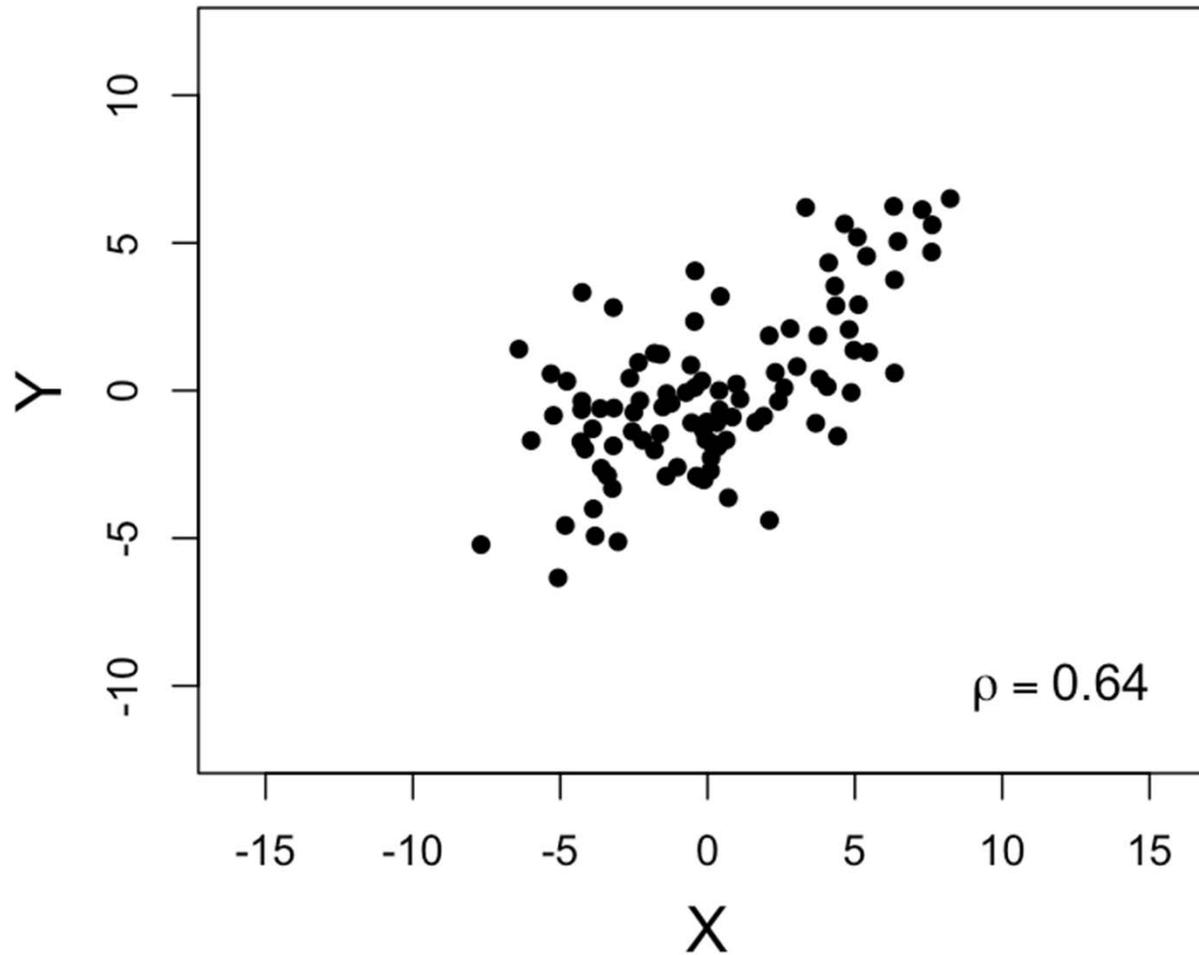
Finding common patterns in data

- If we had N ts, could we use M common trends to describe their temporal variability, such that $N \gg M$?
- Dynamic Factor Analysis (DFA) is an approach to ts modeling that does just that

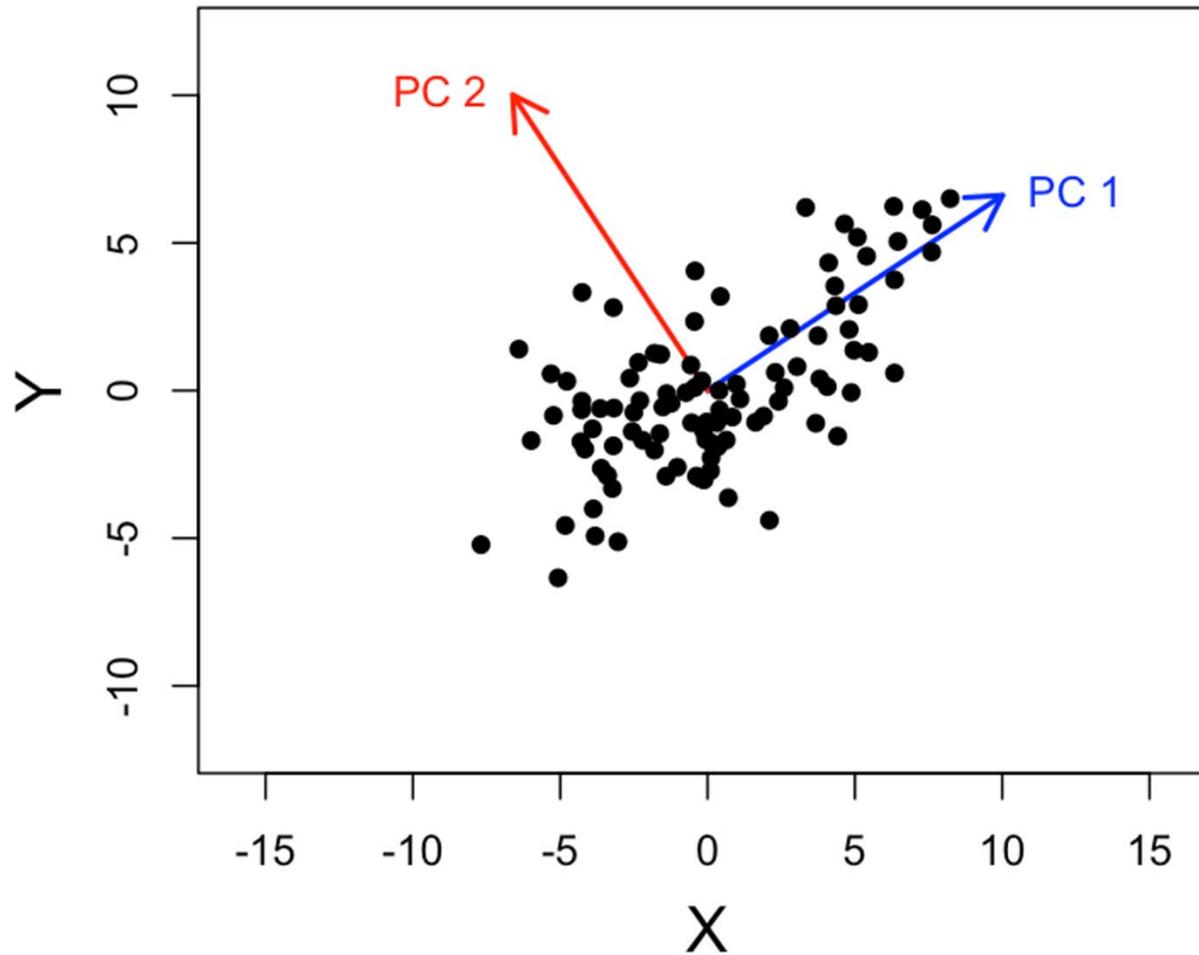
Let's start with PCA

- PCA stands for Principal Component Analysis
- Goal is to reduce some correlated variables to fewer uncorrelated values
- Number of principal components is generally less than the number of original variables

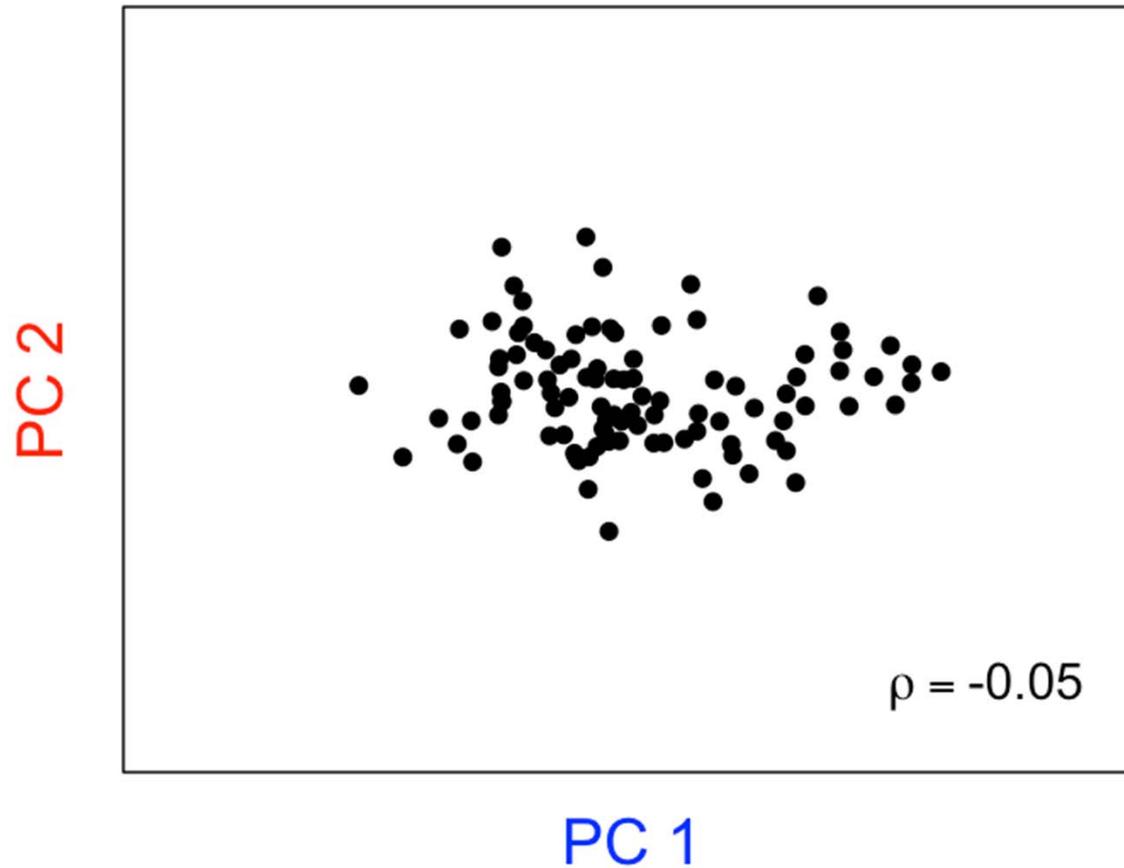
A graphical example



Adding in the first 2 PC's



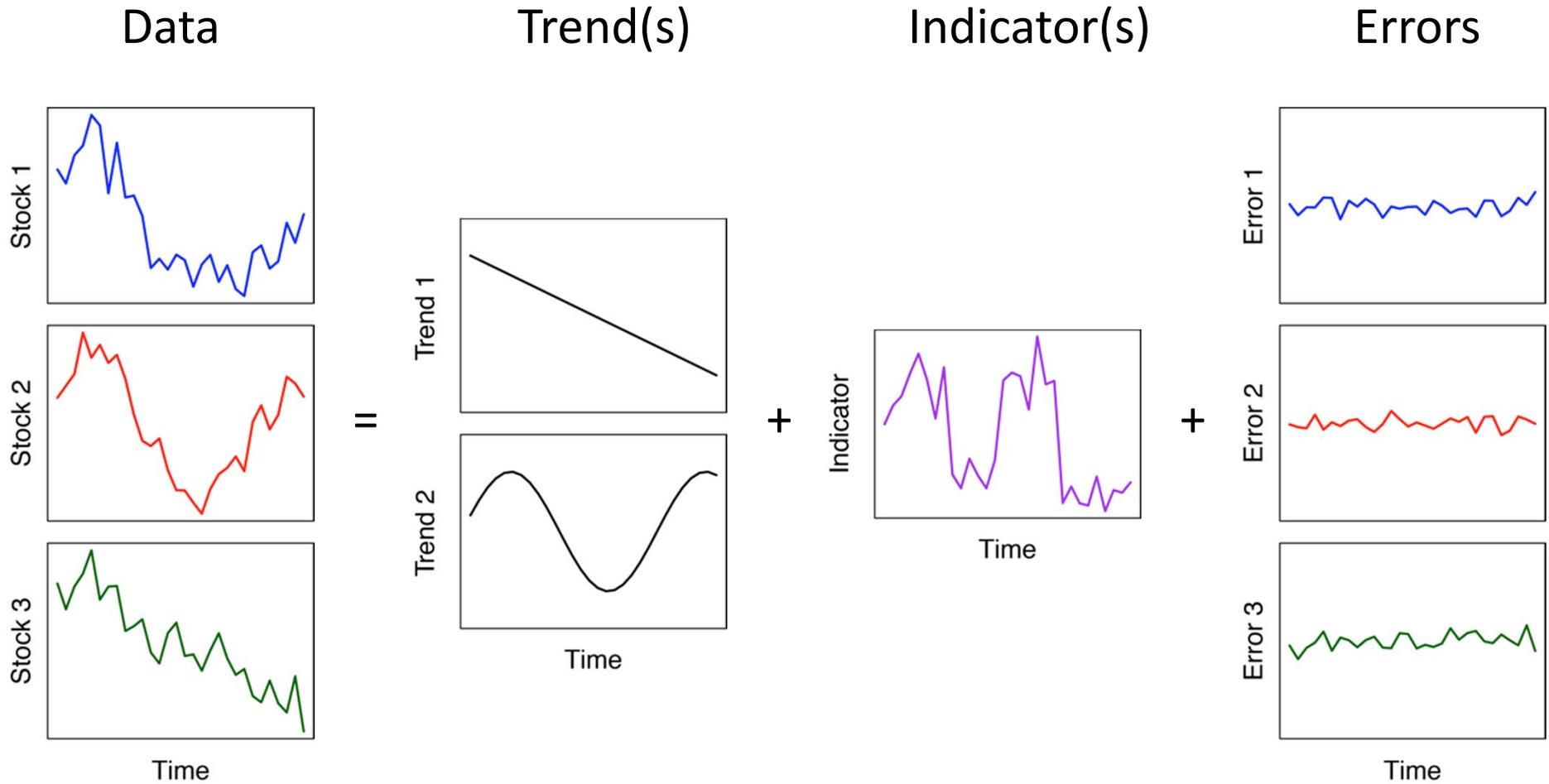
And rotating the basis



What exactly *is* DFA?

- It's like PCA for time series
- DFA can indicate whether there are any:
 - 1) underlying common patterns/trends in the time series,
 - 2) interactions between the response variables, and
 - 3) what the effects of explanatory variables are.
- The mathematics are complex—for details, see Zuur et al. (2003) *Environmetrics*

DFA in common terms



DFA in matrix form

State equation

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{w}_t$$

$$\mathbf{w}_t \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$$

Common pattern(s) over time

Observation equation

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{v}_t$$

$$\mathbf{v}_t \sim \text{MVN}(\mathbf{0}, \mathbf{R})$$

Relate pattern(s) to observations via \mathbf{Z}

DFA with covariates

State equation

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{w}_t$$

$$\mathbf{w}_t \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$$

Common trends over time

Observation equation

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t$$

$$\mathbf{v}_t \sim \text{MVN}(\mathbf{0}, \mathbf{R})$$

Relate trends (\mathbf{x}) to observations (\mathbf{y}) via \mathbf{Z} ,
and covariates (\mathbf{d}) to \mathbf{y} via \mathbf{D}

Relationship between PCA & DFA

- Similarity with PCA can be seen via

$$\text{Cov}(\mathbf{y}_t) = \mathbf{Z}\mathbf{Z}^T + \mathbf{R}$$

- In PCA, however, \mathbf{R} is constrained to be diagonal
- Not so in DFA

Various forms for R

Diagonal & equal

$$\begin{bmatrix} \sigma & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 \\ 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & \sigma \end{bmatrix}$$

Equal variance & covariance

$$\begin{bmatrix} \sigma & \gamma & \gamma & \gamma \\ \gamma & \sigma & \gamma & \gamma \\ \gamma & \gamma & \sigma & \gamma \\ \gamma & \gamma & \gamma & \sigma \end{bmatrix}$$

Diagonal & unequal

$$\begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix}$$

Spp-specific variances & covariances

$$\begin{bmatrix} \sigma_{chin} & 0 & 0 & 0 \\ 0 & \sigma_{coho} & 0 & \gamma_{coho} \\ 0 & 0 & \sigma_{pink} & 0 \\ 0 & \gamma_{coho} & 0 & \sigma_{coho} \end{bmatrix}$$

Some caveats in fitting DFA models

State equation

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{w}_t$$

$$\mathbf{w}_t \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$$

1) Set $\mathbf{Q} = \text{Identity}$

Observation equation

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t$$

$$\mathbf{v}_t \sim \text{MVN}(\mathbf{0}, \mathbf{R})$$

2) Consider constraints of \mathbf{Z} and \mathbf{a}

Constraining the \mathbf{a} vector

Observation equation

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t \quad \mathbf{v}_t \sim \text{MVN}(\mathbf{0}, \mathbf{R})$$

Constraining portions of \mathbf{a} (eg, $n = 5$; $m = 3$)

$$\mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ a_i \\ a_i \end{bmatrix}$$

in first m rows of \mathbf{a} , $a_i = 0$

Note: This approach causes the EM algorithm to take a very long time to converge

Soln: We will demean our data and set $\mathbf{a} = \mathbf{0}$

Constraining the \mathbf{Z} matrix

Observation equation

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t \quad \mathbf{v}_t \sim \text{MVN}(\mathbf{0}, \mathbf{R})$$

Constraining portions of \mathbf{Z} (eg, $n = 5; m = 3$)

$$\mathbf{Z} = \begin{bmatrix} z_{ij} & \mathbf{0}_{ij} & \mathbf{0}_{ij} \\ z_{ij} & z_{ij} & \mathbf{0}_{ij} \\ z_{ij} & z_{ij} & z_{ij} \\ z_{ij} & z_{ij} & z_{ij} \\ z_{ij} & z_{ij} & z_{ij} \end{bmatrix} \quad \text{in } m-1 \text{ rows of } \mathbf{Z}, z_{ij} = 0 \text{ if } j > i$$

Rotation matrix \mathbf{H}

If \mathbf{H} is arbitrarily constrained to obtain just itself as sequential

$$(1) \quad \begin{aligned} \mathbf{x}_t &= \mathbf{x}_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &= \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t \end{aligned}$$

We need to choose appropriate \mathbf{H} —we'll use “varimax”

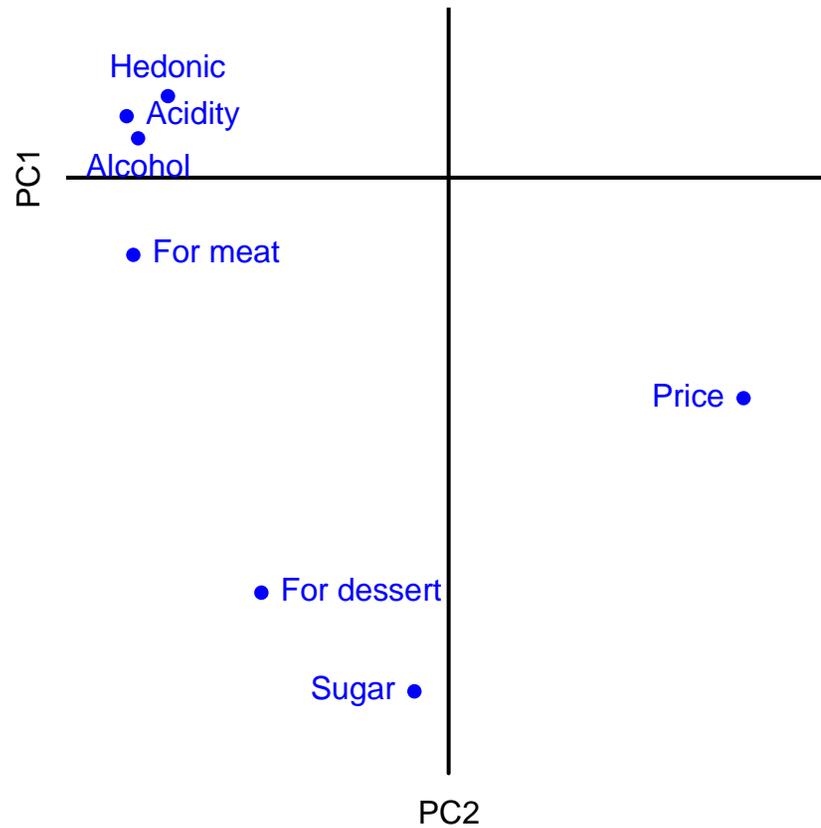
$$(2) \quad \begin{aligned} \mathbf{H}\mathbf{x}_t &= \mathbf{H}\mathbf{x}_{t-1} + \mathbf{H}\mathbf{w}_t \\ \mathbf{y}_t &= \mathbf{Z}\mathbf{H}^{-1}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t \end{aligned}$$

Varimax rotation for H

- A “simple” solution means each factor has a small number of large loadings, and a large number of (near) zero loadings
- After varimax rotation, each original variable tends to be associated with one (or a small number) of factors
- Varimax searches for a linear combination of original factors that *maximizes* the variance of the loadings

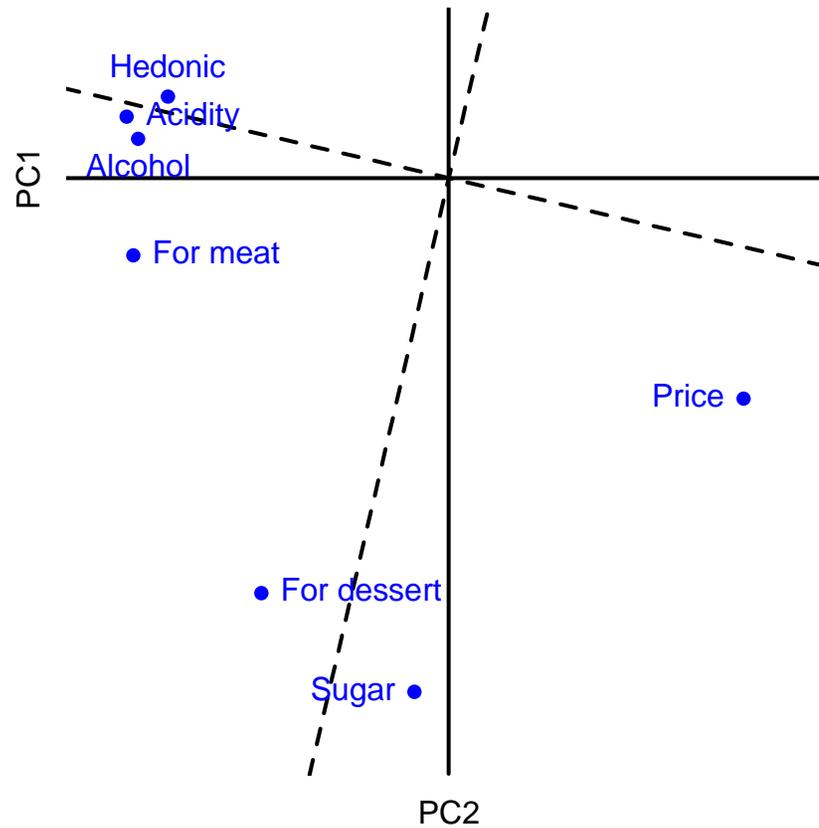
$$\max \sum (z_{ij}^2 - \bar{z}_{ij}^2)^2$$

Varimax rotation for H



Varimax rotation for H

After varimax rotation

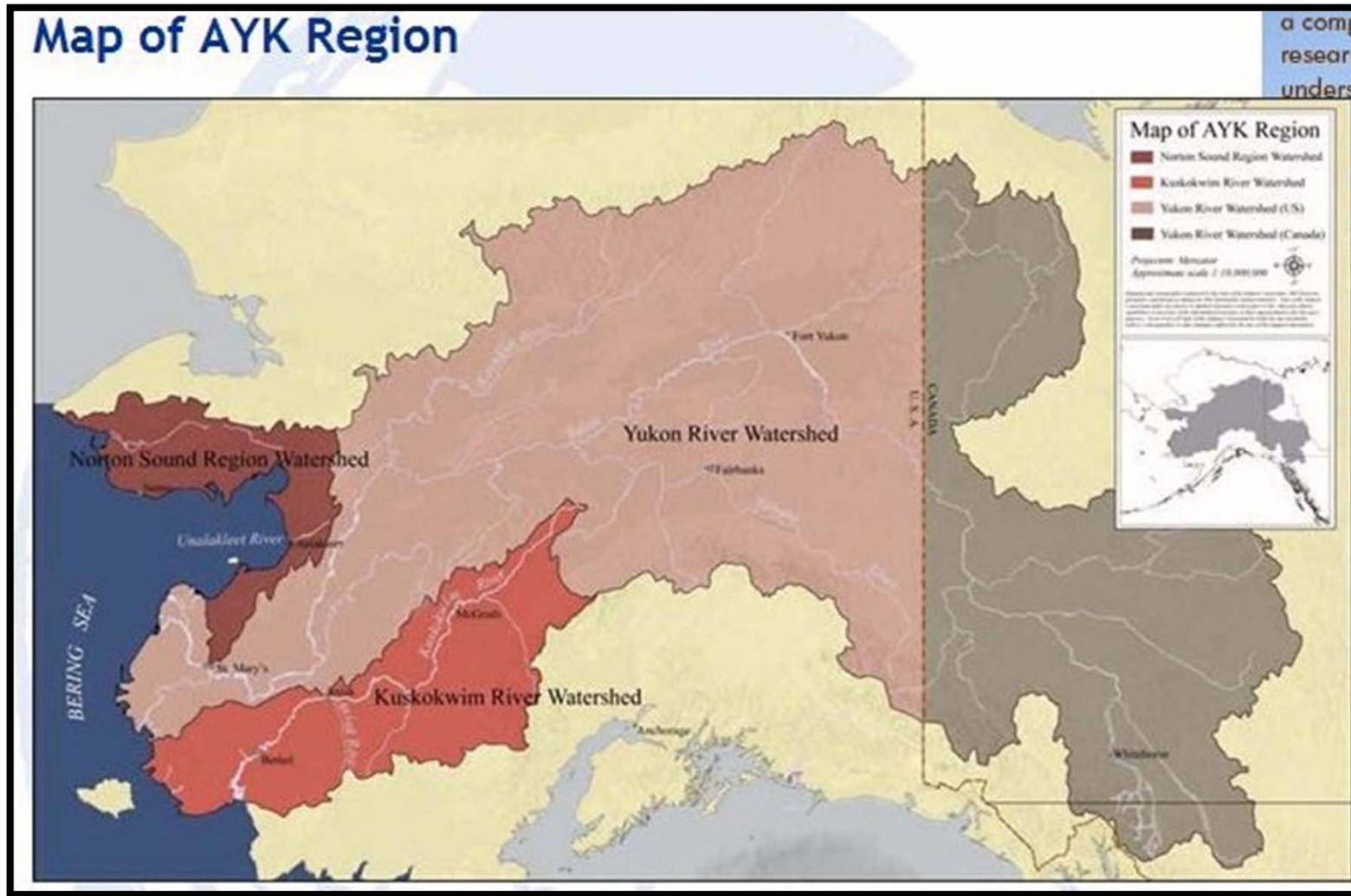




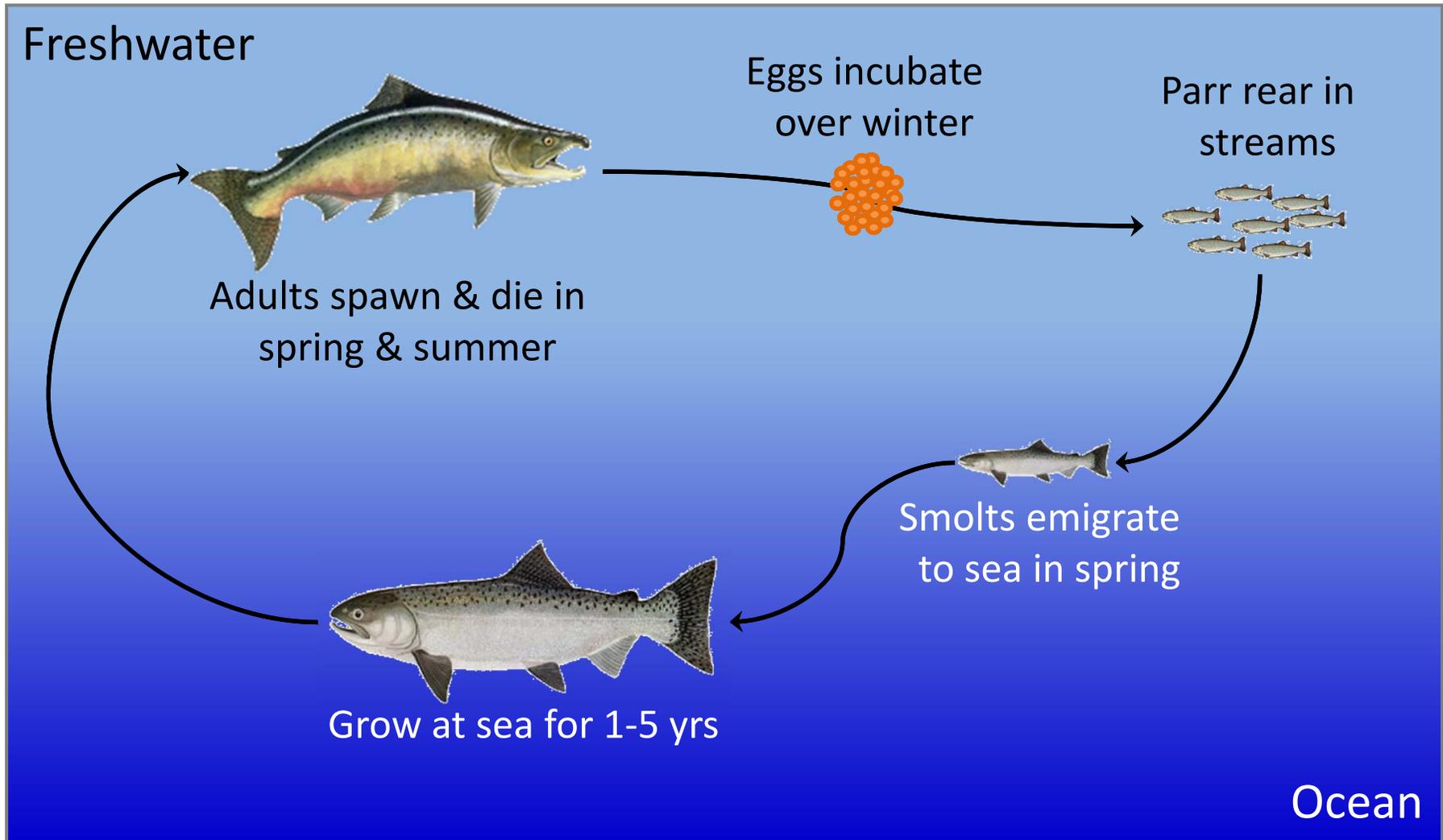
Arctic-Yukon-Kuskokwim salmon

- 1) Poor returns of Chinook & chum in AYK region over past decade have led to severe restrictions on commercial & subsistence harvest
- 2) This has also led to repeated disaster declarations by the state and federal governments (**nobody** fished in 2012!).
- 3) In response, native regional organizations, state and federal agencies formed an innovative partnership to cooperatively address problems (AYK SSI)

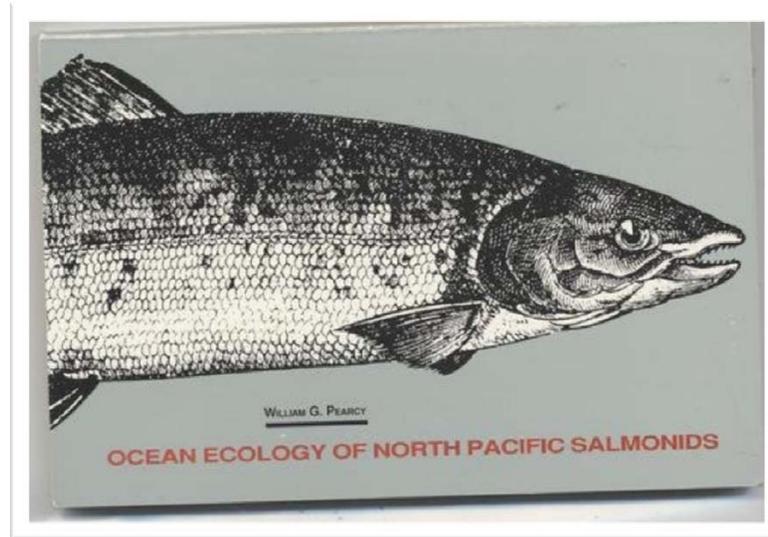
AYK region



Salmon life cycle



Background & motivation



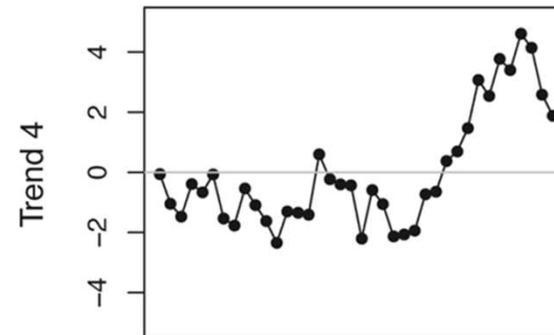
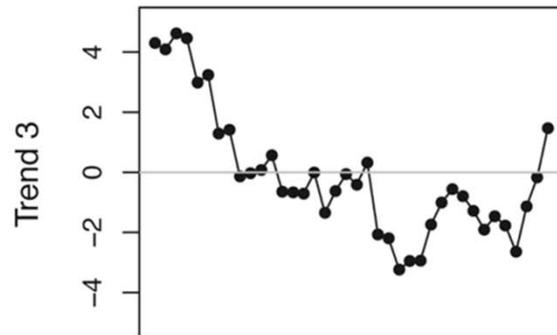
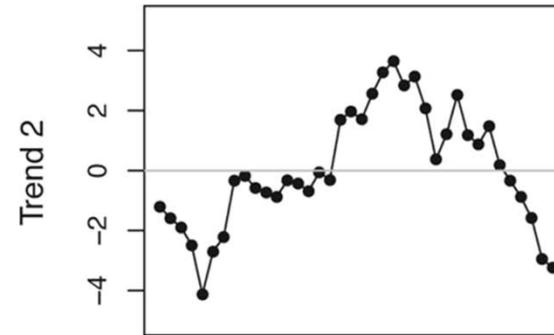
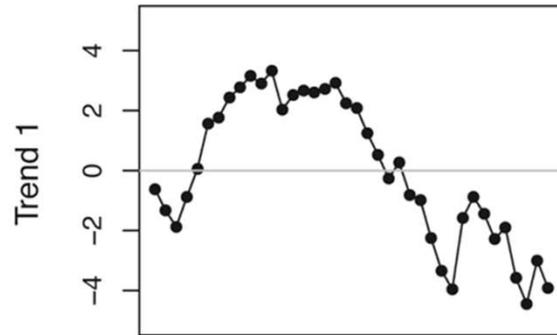
Ocean-climate variability linked to marine survival

- Fraser R sockeye & ENSO (Mysak 1986)
- AK pink & W Coast coho (Sibley & Francis 1991)
- Various spp & Aleutian Low (Beamish & Bouillon 1993)

The question

What evidence exists to support an ocean-climate hypothesis based on a *time series* examination in a *model selection framework*?

Possible trends in the data?



The data

- 5 “stocks” of AYK Chinook
 - Chena & Salcha
 - Goodnews
 - Kuskokwim
 - Unalakleet
 - Yukon (Canadian side)
- Brood years 1981-2005
- Covariates lagged by 1-5 years

Used 3 estimates of “productivity”

- 1) Natural logarithm of recruits per spawner: $\ln(R/S)$
- 2) Residuals from stock-recruit (Ricker) model
- 3) Density-independent parameter from Ricker model

Today's focus

- 1) Natural logarithm of recruits per spawners: $\ln(R/S)$
- 2) Residuals from stock-recruit (Ricker) model
- 3) Density-independent parameter from Ricker model

Environmental indicators

1) Pacific Decadal Oscillation

- Mean annual (Jan-Dec)
- Mean annual (May-Apr)
- Mean winter (Oct-Mar)

2) Arctic Oscillation Index

3) Aleutian Low Pressure Index

4) North Pacific Index

- Winter
- Spring

5) Sea level pressure

- Winter
- Spring

6) Strong winds index

7) Sea Surface Temperature

- Winter
- Spring
- May
- July

8) Ice-out date

9) Forage Fish index

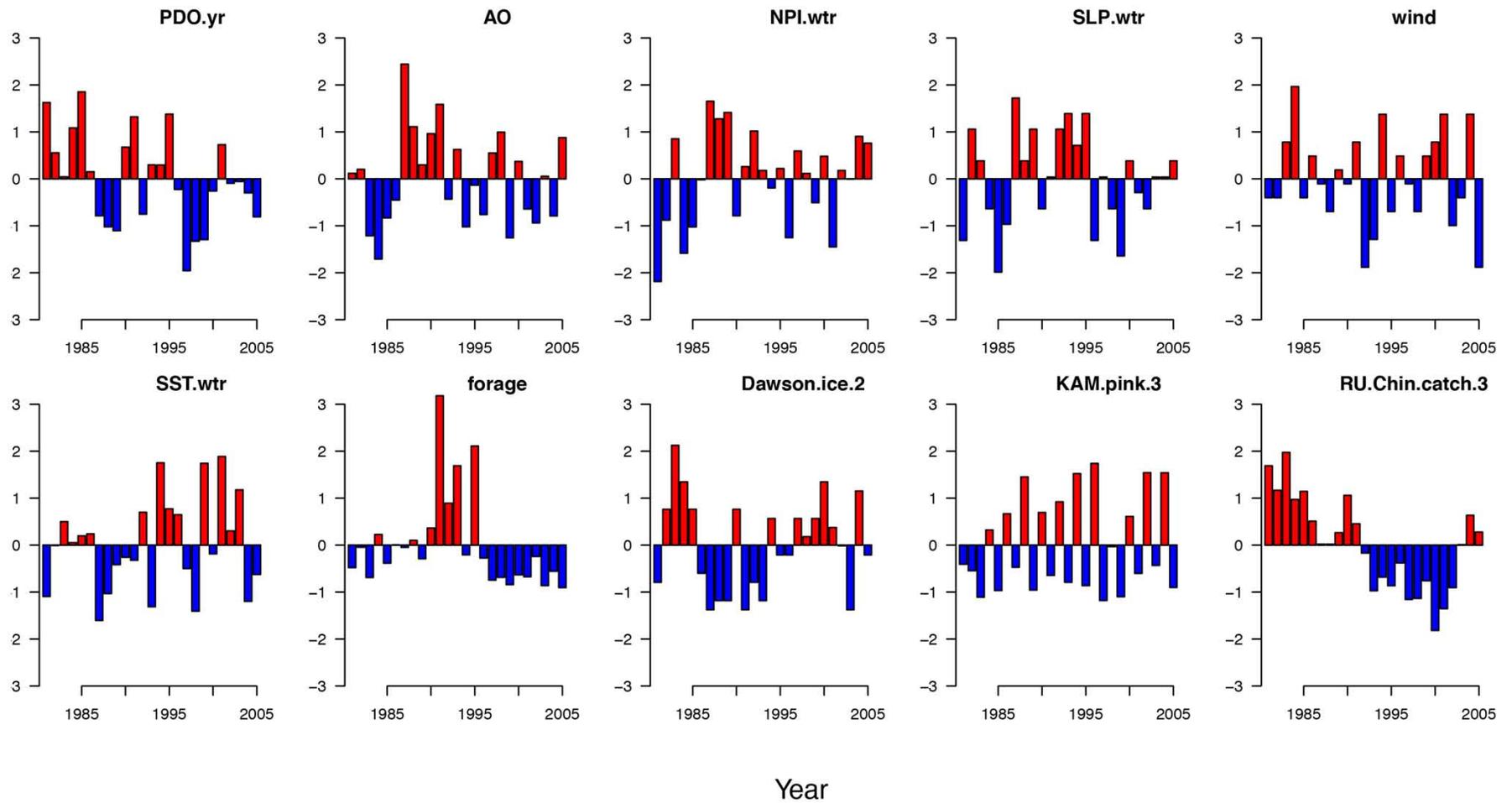
10) Pollock biomass

11) Kamchatka pinks

12) BSAI Chinook bycatch

13) Russian Chinook catch

Examples of some indicators



The analysis

- Varied the number of states/trends from 1-9
- Varied forms of R to try:
 - 1) Diagonal and equal,
 - 2) Diagonal and unequal,
 - 3) Equal variances and covariances.
- Used AICc to select “best” model

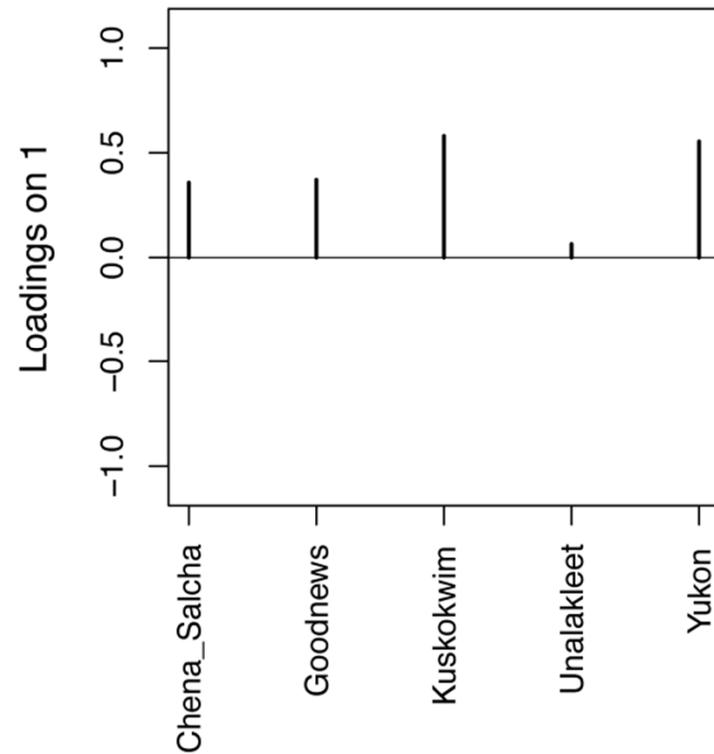
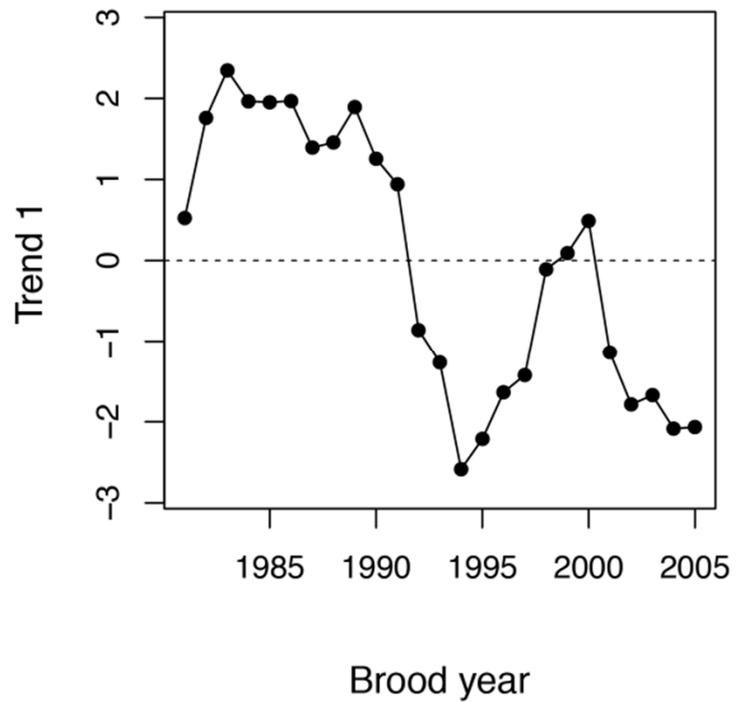
AYK - model selection results

The most parsimonious model had 1 common trend & 2 indicators:

- 1) Timing of ice-out at Dawson in year smolts go to sea
- 2) Russian catches of Chinook during 2nd year at sea

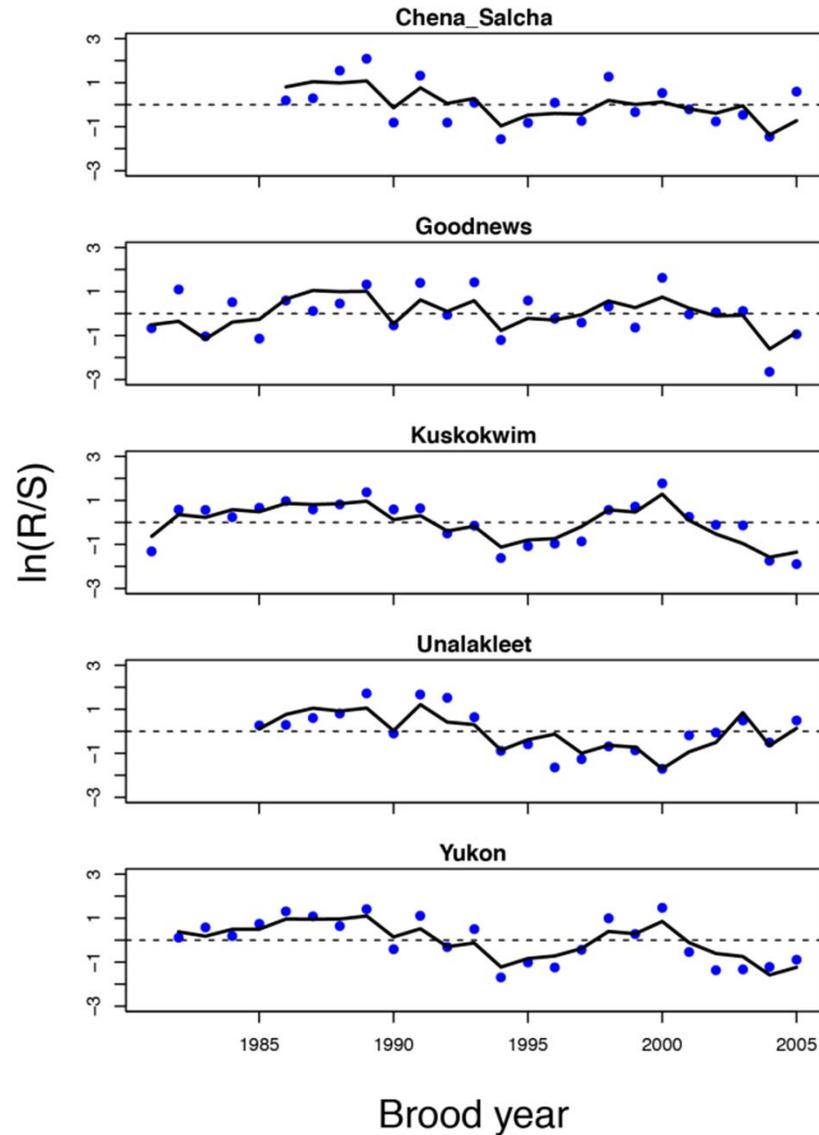
AYK – common trend in $\ln(R/S)$

With Dawson ice-out & RUS catches

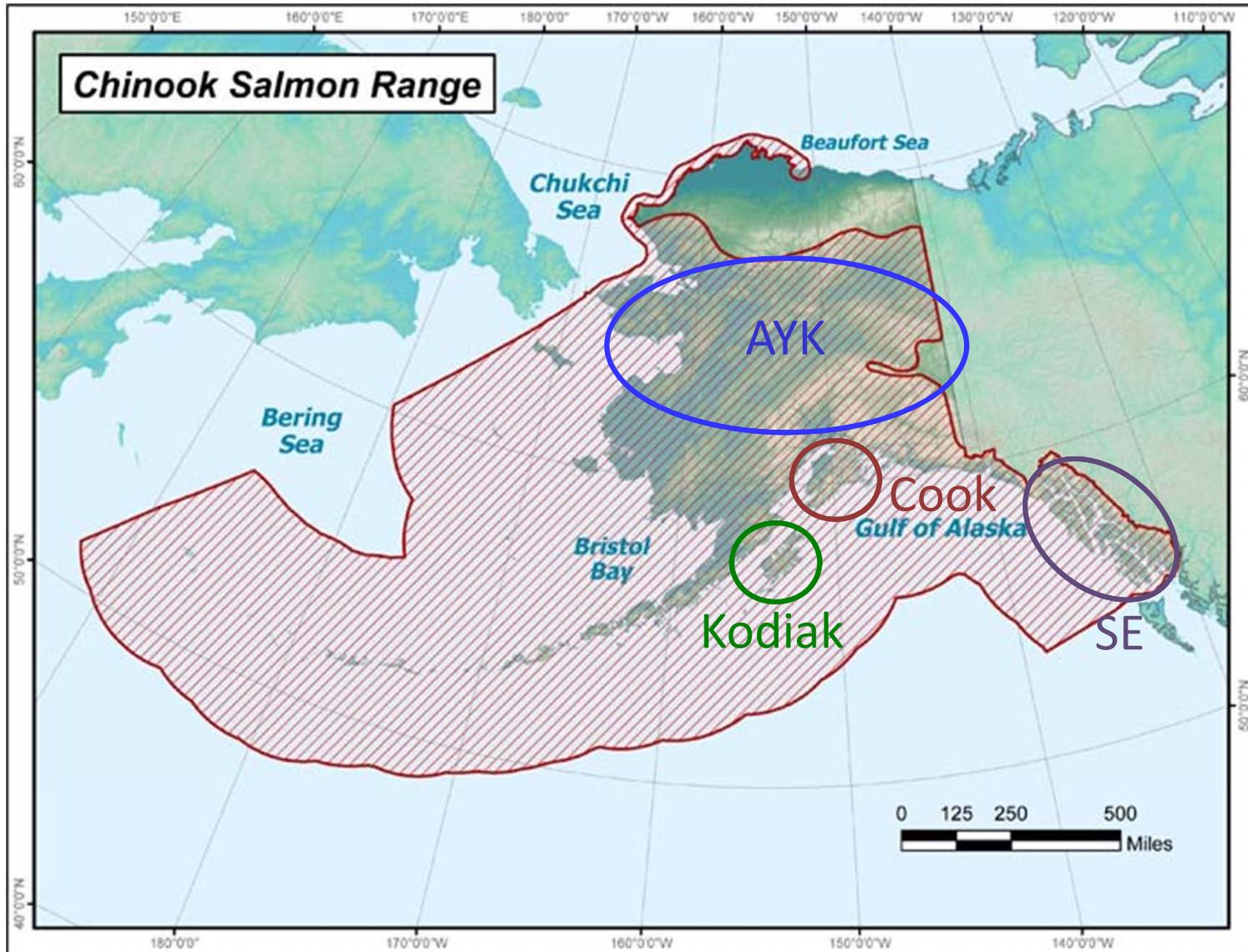


AYK – model fits to data

With Dawson ice-out & Russian catches



Alaska Chinook salmon



Expanding the analysis to statewide

- 10 additional statewide indicator stocks
 - Anchor (Cook)
 - Dershka (Cook)
 - Ayakulik (Kodiak)
 - Karluk (Kodiak)
 - Nelson (Kodiak)
 - Alsek (SE)
 - Blossom (SE)
 - Situk (SE)
 - Stikine (SE)
 - Taku (SE)
- Brood years 1976-2005

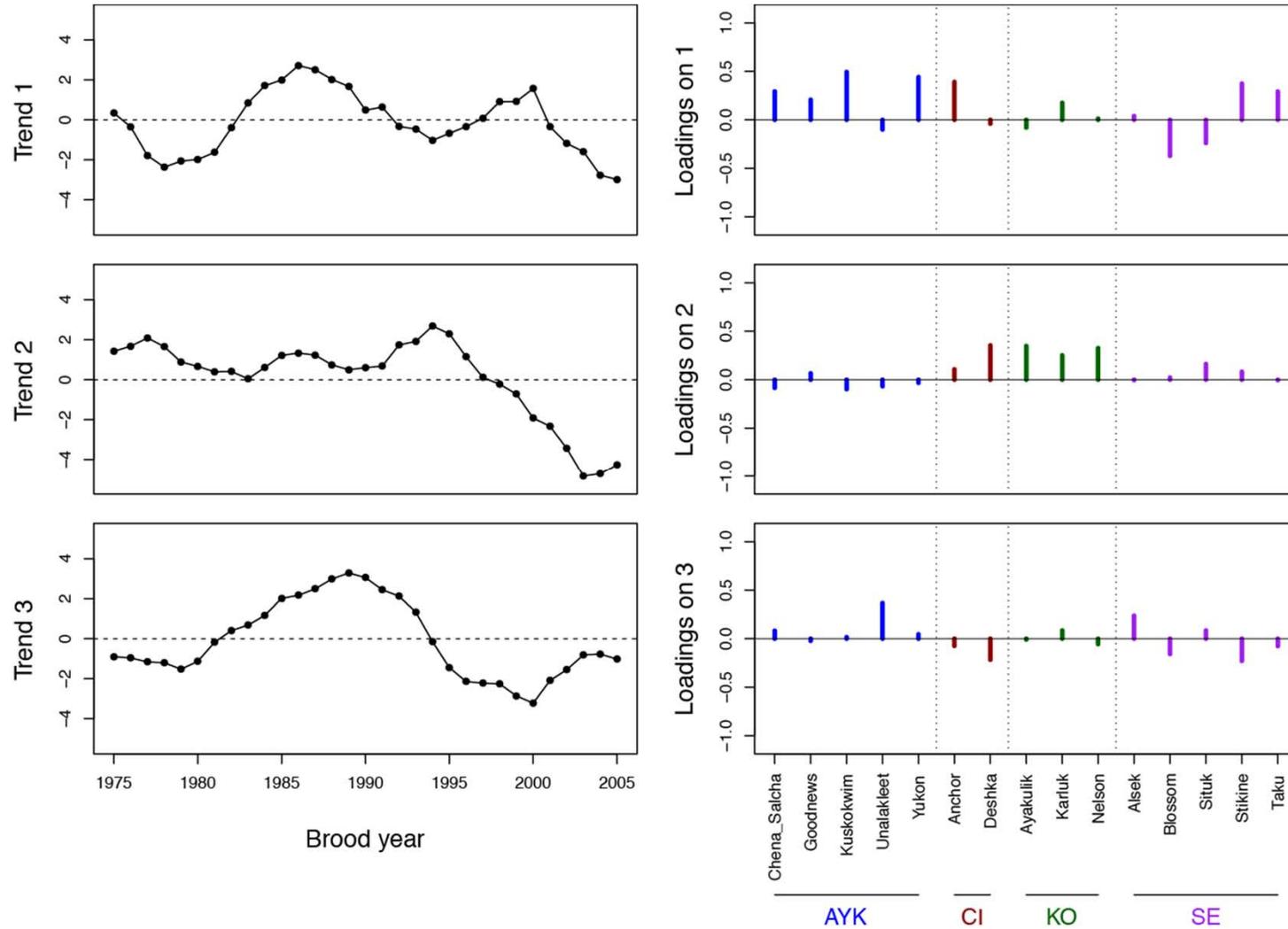
Expanding the analysis to statewide

- Used limited set of “wide-spread” indicators
 - 1) PDO
 - Mean annual (Jan-Dec)
 - Mean annual (May-Apr)
 - Mean winter (Oct-Mar)
 - 2) Arctic Oscillation Index
 - 3) Aleutian Low Pressure Index
 - 4) North Pacific Index
 - Winter
 - Spring

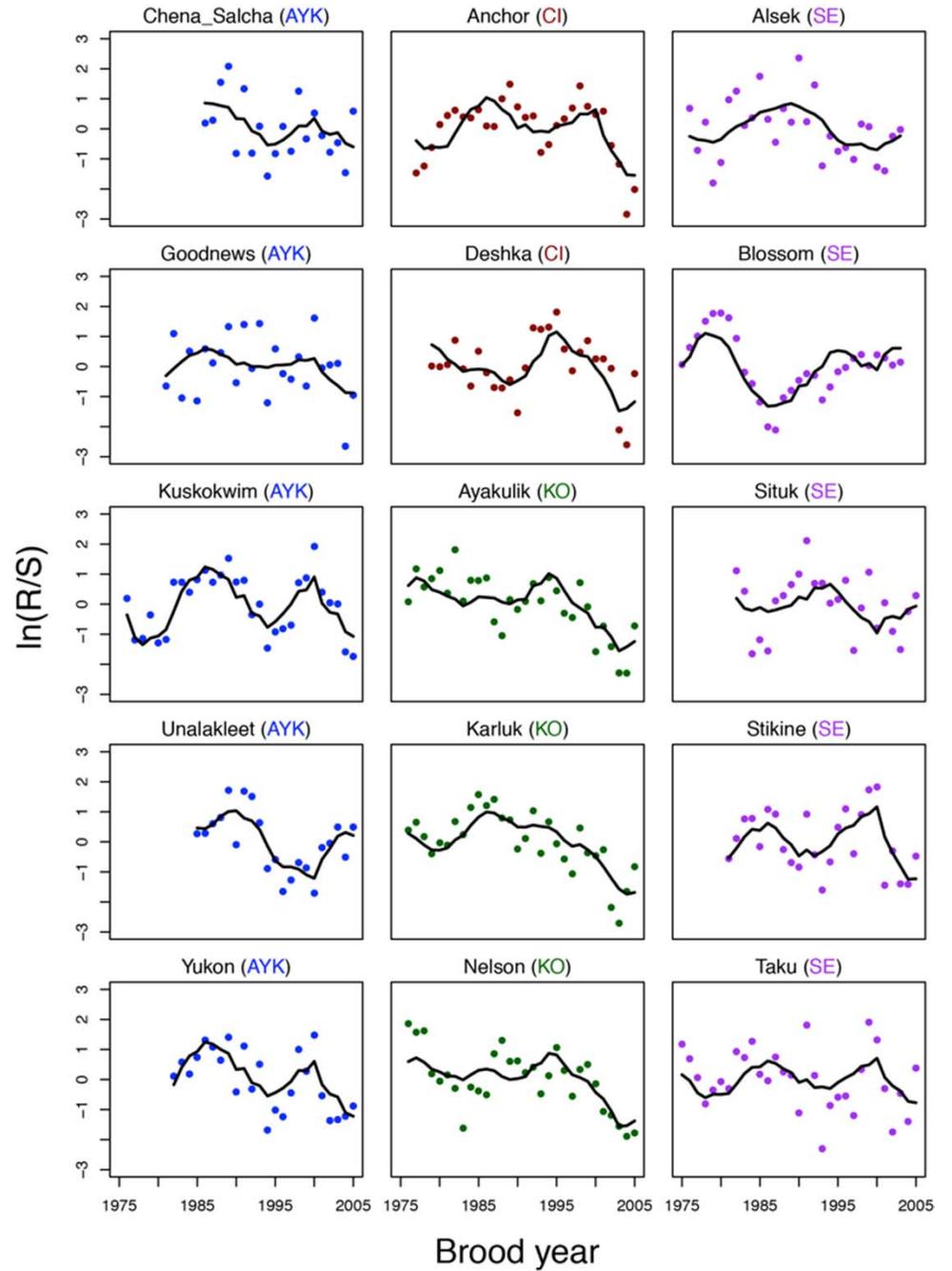
Statewide – model selection results

- The most parsimonious model had 3 common trends & no indicators
- The “best” model with indicator(s) had 3 common trends & 1 indicator:
 - 1) Arctic Oscillation Index

Statewide – common trends in $\ln(R/S)$



Statewide model fits



Topics for lab

- Fitting DFA models without covariates
- Fitting DFA models with covariates
- Doing factor rotations