Covariates in Time Series Models

FISH 550 – Applied Time Series Analysis

Eli Holmes

18 Apr 2023

Topics for today

Covariates

- Why include covariates?

- Multivariate linear regression on time series data

- Covariates in MARSS models

- Seasonality in MARSS models

- Missing covariates

Why include covariates in a model?

- You want to forecast something using covariates

- We are often interested in knowing the cause of variation

- Covariates can explain the process that generated the patterns

- Covariates can help deal with problematic observation errors

- You are using covariates to model a changing system

- You want to get rid of trends or cycles

Lake WA plankton and covariates
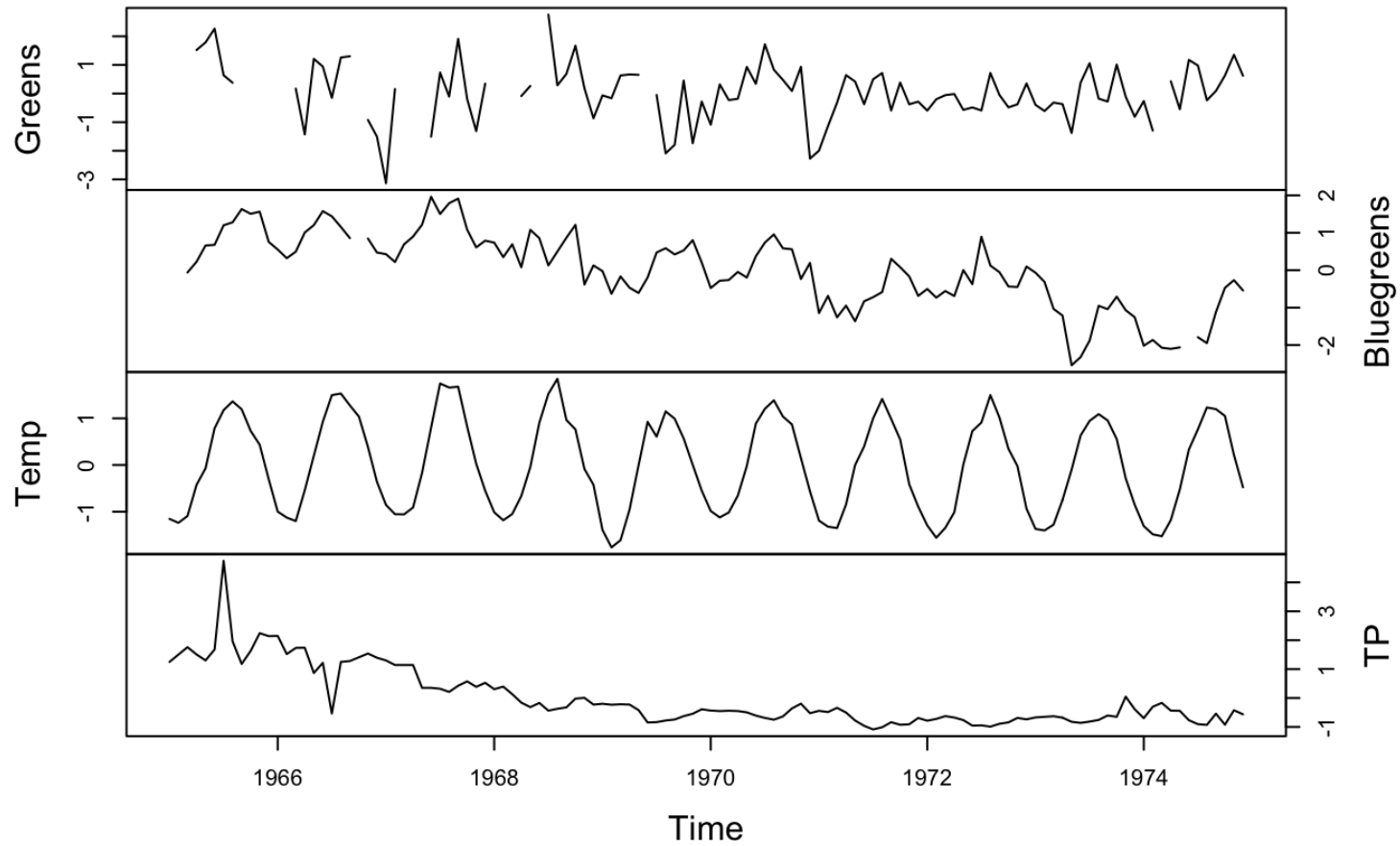
Covariates in time series models

- Multivariate linear regression for time series data

- Linear regression with ARMA errors

- ARMAX - process errors driven by covariates

- MARSS models with covariates = process and observation errors affected by covariates

    - aka Vector Autoregressive Models with covariates and observation error

- Covariates in general state-space models

Multivariate linear regression for time series data

Can you do a linear regression with time series data (response and predictors)? Yes, but you need to be careful. Read Chapter 5 in Hyndman and Athanasopoulos 2018

- Diagnostics that need to be satisfied
    - Residuals are temporally uncorrelated
    - Residuals are not correlated with the predictor variables
- Be careful regarding spurious correlation if both response and predictor variables have trends

Autocorrelated response and predictor variables

Both response and predictors have a seasonal trend

both covariate and response are seasonal
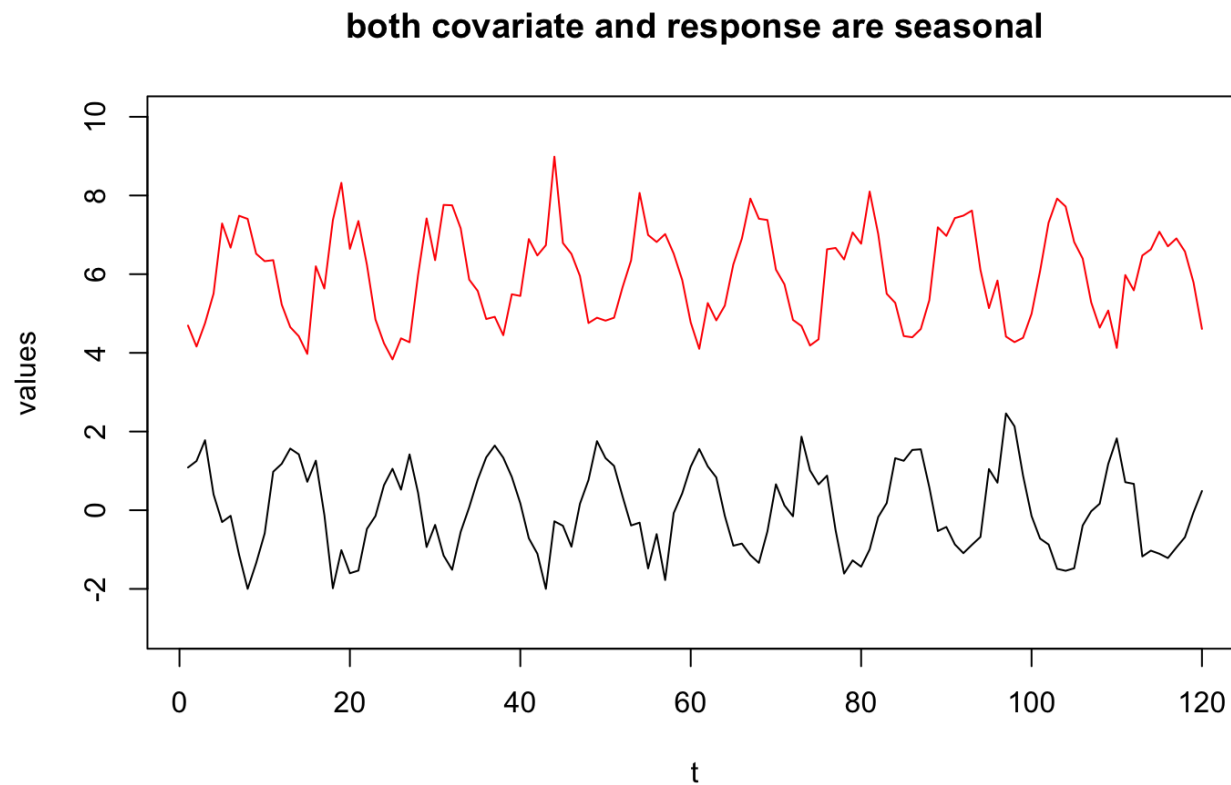
Linear regression with autocorrelated errors

The `xreg` argument in `Arima()` and `arima()` allows you to fit linear regressions with autocorrelated errors. Read Chapter 9 in Hyndman and Athanasopoulos 2018 on Dynamic Regression.

A linear regression with autocorrelated errors is for example:

$$y_t = \alpha + \mathbf{D}\mathbf{d}_t + \nu_t$$

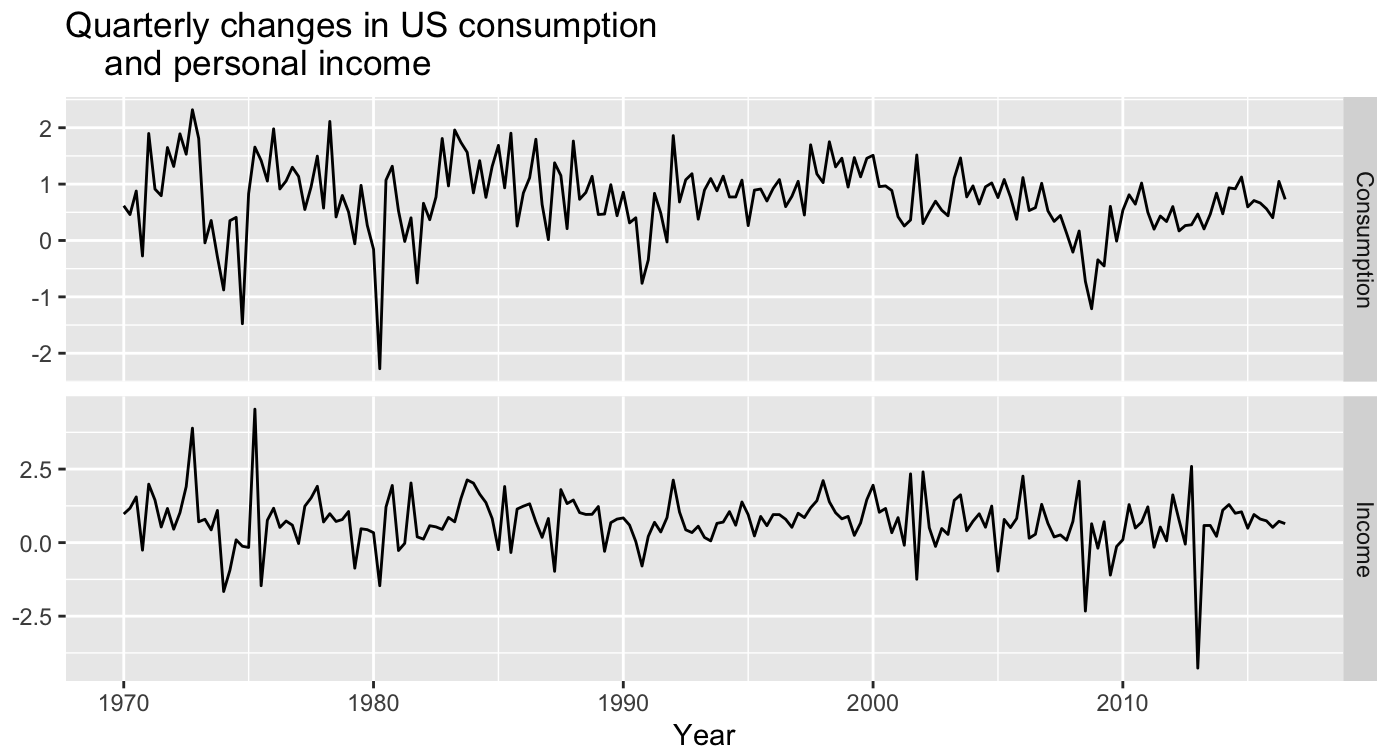$$\nu_t = \theta_1 \nu_{t-1} + \theta_2 \nu_{t-2} + e_t$$

Fitting in R

`Arima()`

```
fit <- Arima(y, xreg=d, order=c(1,1,0))
```

`auto.arima()`
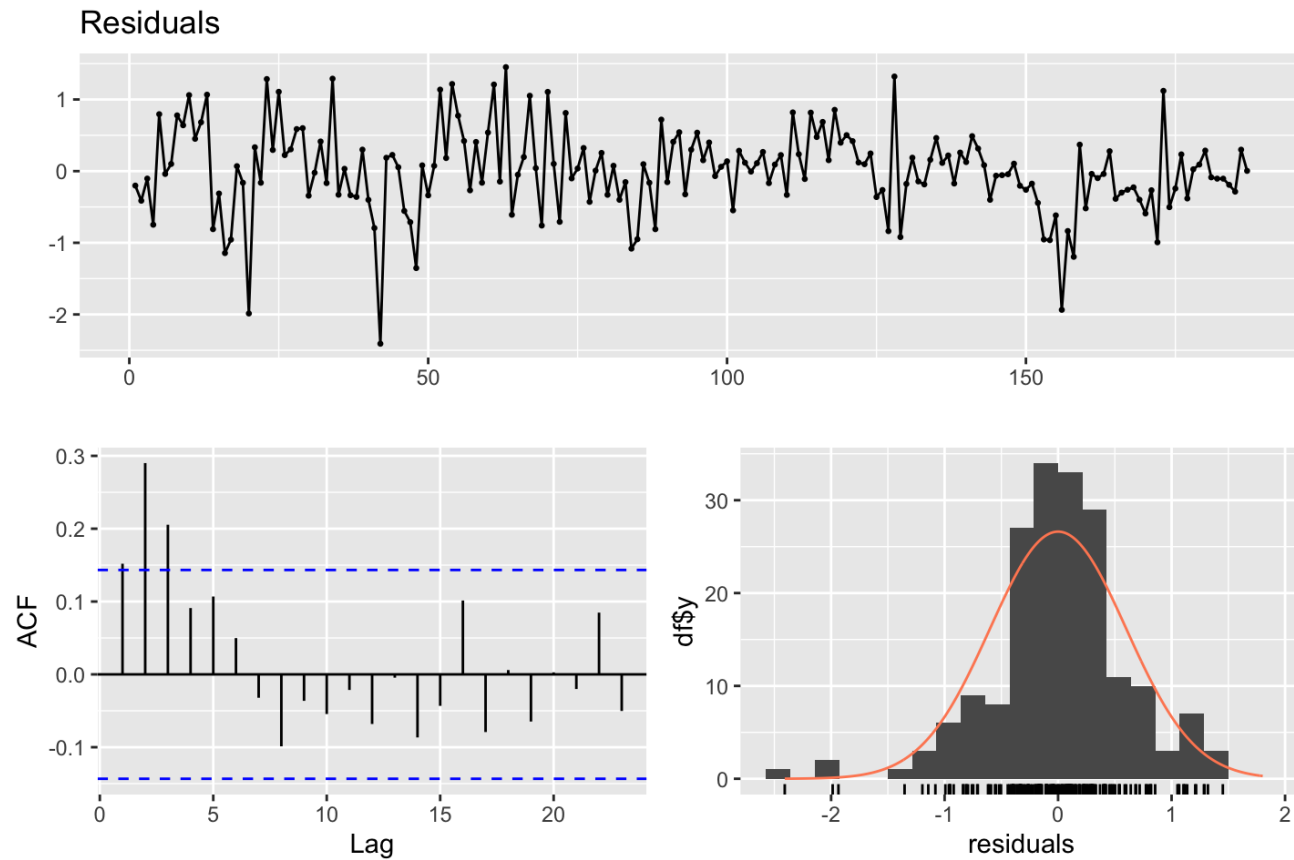
```
fit <- auto.arima(y, xreg=x)
```

LOTS of packages have options for auto-correlated errors

Example from Hyndman and Athanasopoulos 2018

Quarterly changes in US consumption and personal income

A simple regression has problems

```r
y <- uschange[,"Consumption"]; d <- uschange[,"Income"]
fit <- lm(y~d)
checkresiduals(fit)
```
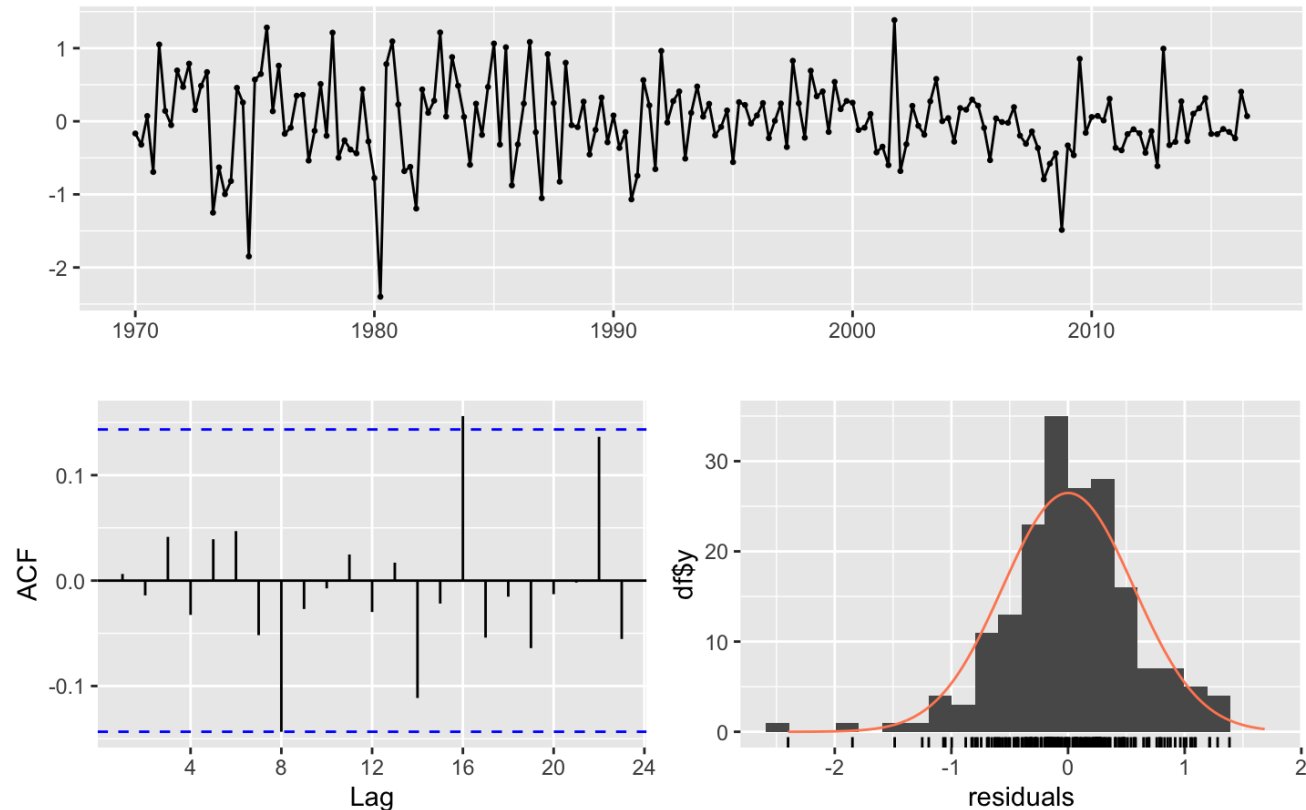


```
##
##  Breusch-Godfrey test for serial correlation of order up to 10
##
## data:  Residuals
## LM test = 27.584, df = 10, p-value = 0.002104
```

Let `auto.arima()` find best model

```
fit <- auto.arima(y, xreg=d) # It finds a ARMA(1,0,2) is best.
checkresiduals(fit)
```


Residuals from Regression with ARIMA(1,0,2) errors

```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,0,2) errors
## Q* = 5.8916, df = 5, p-value = 0.3169
##
```

Collinearity

This a big issue. If you are thinking about stepwise variable selection, do a literature search on the issue. Read the chapter in Holmes 2018: Chap 6 on catch forecasting models using multivariate regression for a discussion of

- Stepwise variable regression in R
- Cross-validation for regression models
- Penalized regression in R
  - Lasso
  - Ridge
  - Elastic Net
- Diagnostics

ARMAX

ARMAX models are different. In this case, the covariates affect the amount the auto-regressive process changes each time step.

You can think of this as

- covariates affect the process errors (good bad years)

$$x_t = bx_{t-1} + \underbrace{\boxed{\mathbf{Cc}_t + w_t}}_{\text{process error}}$$

- covariates drive the drift, aka **u**

$$x_t = bx_{t-1} + \underbrace{\boxed{\mathbf{Cc}_t}}_{\text{drift "u"}} + w_t$$

Covariates in MARSS models

This is a state-space model that allows you to have the covariate affects process error and affect observation errors.

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{C}\mathbf{c}_t + \mathbf{w}_t$$

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

Example - univariate state-space models

$$x_t = bx_{t-1} + u + \mathbf{C}\mathbf{c}_t + w_t$$

$$y_t = x_t + \mathbf{D}\mathbf{d}_t + v_t$$

Now we can model how covariates affect the hidden process.

Example - univariate state-space models

$$x_t = x_{t-1} + u + \boxed{\mathbf{Cc}_t} + w_t$$

$$y_t = x_t + v_t$$

Random walk with drift. How does covariate affect the drift term?

Example. You have tag data on movement of animals in the ocean. How does water temperature affect the speed (jump length) of the movement.

Example - univariate state-space models

$$x_t = x_{t-1} + u + w_t$$

$$y_t = x_t + \boxed{\mathbf{Dd}_t + v_t}$$

How does covariate affect observation error relative to our stochastic trend.

Example. You are tracking population size using stream surveys. Turbidity affects your observation error.

Multivariate Example - Covariates in state process

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{u} + \boxed{\mathbf{C}\mathbf{c}_t} + \mathbf{w}_t$$

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

$\mathbf{C}\mathbf{c}_t$ The covariate is in $\mathbf{c}_t$ and the effect is in matrix $\mathbf{C}$.

Example. lat/lon movement data so $\mathbf{x}$ and $\mathbf{y}$ are 2 dimensional (our lat and lon values).

Multivariate Example - Covariates in state process

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_t = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{t-1} + \begin{bmatrix} C_a & C_b \\ C_a & C_b \end{bmatrix} \begin{bmatrix} temp \\ TP \end{bmatrix}_t + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}_t$$

The model for x_t in site 1 (or species 1) is:

$$x_{1,t} = x_{1,t-1} + C_a \times temp_t + C_b \times TP_t + w_{1,t}$$

There is an effect of the prior x_t and an effect of temperature and phosporous.

The structure of $\mathbf{C}$

The structure of $\mathbf{C}$ can model different effect structures

Effect of temp and TP is the same

$$\begin{bmatrix} C & C \\ C & C \end{bmatrix} \begin{bmatrix} temp \\ TP \end{bmatrix}_t$$

Effect of temperature and TP is different but the same across sites, species, whatever the row in x is

$$\begin{bmatrix} C_a & C_b \\ C_a & C_b \end{bmatrix} \begin{bmatrix} temp \\ TP \end{bmatrix}_t$$

Effect of temperature and TP is different across sites or whatever the row in x is

$$\begin{bmatrix} C_{a1} & C_{b1} \\ C_{a2} & C_{b2} \end{bmatrix} \begin{bmatrix} temp \\ TP \end{bmatrix}_t$$

Effect of temperature is the same across sites but TP is not

$$\begin{bmatrix} C_a & C_{b1} \\ C_a & C_{b2} \end{bmatrix} \begin{bmatrix} temp \\ TP \end{bmatrix}_t$$

Multivariate Example - Covariates in observation process

eg, vegetation obscures individuals, temperature affects behavior making animals more or less visible

$$\mathbf{y}_t = \underbrace{\boxed{\mathbf{Zx}_t}}_{\text{hidden state}} + \underbrace{\mathbf{a}}_{\text{scaling term}} + \underbrace{\boxed{\mathbf{Dd}_t + \mathbf{v}_t}}_{\text{observation error}}$$

Covariates in the observation process

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}_t = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_t + \begin{bmatrix} 0 \\ a_2 \\ 0 \end{bmatrix} + \begin{bmatrix} D_a & D_b \\ D_a & D_b \\ D_a & D_b \end{bmatrix} \begin{bmatrix} temp \\ wind \end{bmatrix}_t + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}_t$$

In this case the covariate does not affect the *state x*. It affects the observation of the state.

The model for y_t in site 1 is:

$$y_{1,t} = x_{1,t} + D_a \times temp_t + D_b \times wind_t + v_{1,t}$$

The structure of $\mathbf{D}$

The structure of $\mathbf{D}$ can model many different structures of the effects.

Effect of temp and wind is the same across sites 1 & 2 but different for site 3. In site 3, temp has an effect but wind does not

$$\begin{bmatrix} D_a & D_b \\ D_a & D_b \\ D_c & 0 \end{bmatrix} \begin{bmatrix} temp \\ wind \end{bmatrix}_t$$

Why include covariates in the process?

- We want to understand how covariates drive the hidden process.

- We want to test hypotheses for what caused a perturbation or change in the dynamics.

- We want to forecast using covariates.

- We want to model the *autocorrelation in the process errors* using the known driver.

- We want to remove seasonality or cycles

Why include covariates in a observation?

Auto-correlated observation errors

- Model your v_t as a AR-1 process. hard numerically with a large multivariate state-space model

- If know what is causing the auto-correlation, include that as a covariate. Easier.

Correlated observation errors across sites or species (y rows)

- Use a $\mathbf{R}$ matrix with off-diagonal terms. really hard numerically

- If you know or suspect what is causing the correlation, include that as a covariate. Easier.

We want to remove seasonality or cycles

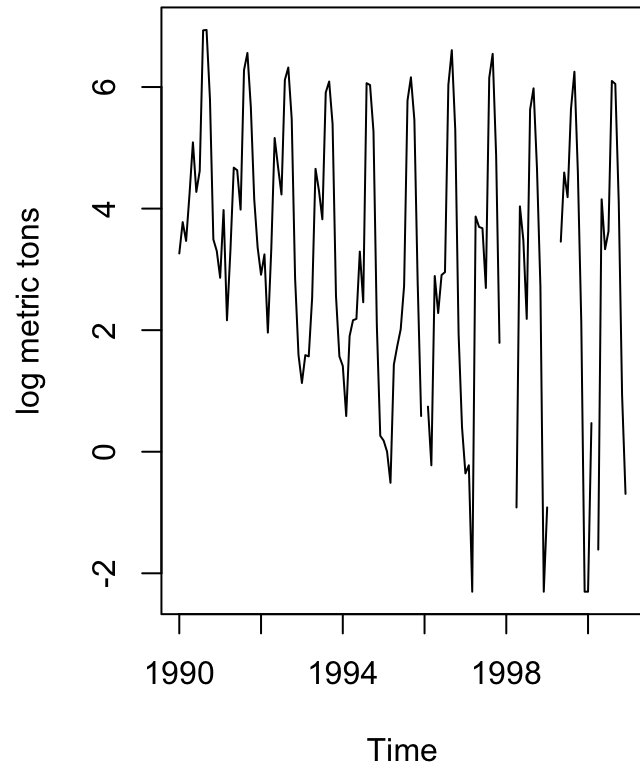"hard numerically" = you need a lot of data

Let's work through an example
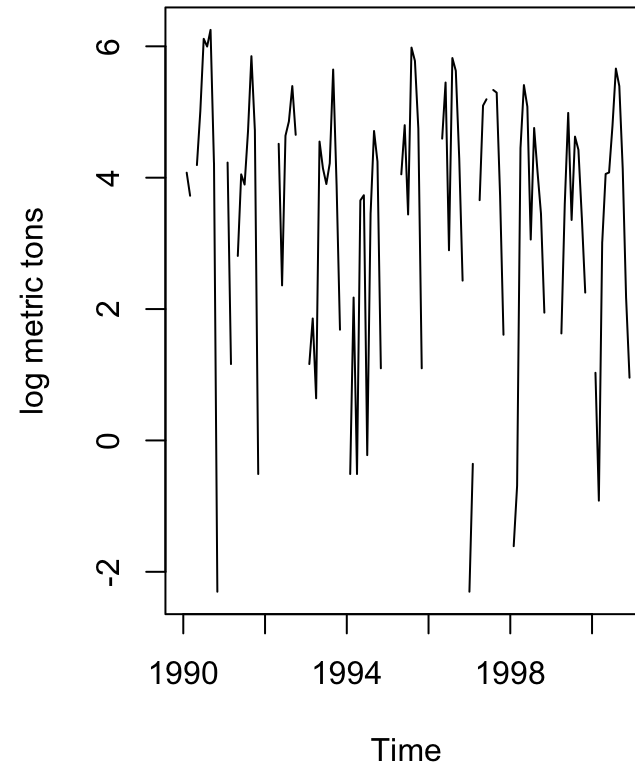
`lec_07_covariates.R` in the Fish550 repo

Follows Chapter 8 in the ATSA lab book

Seasonality

Seasonality

- Different approaches to modeling seasonality

    - Factors

    - Polynomials

    - Sines and cosines (Fourier series)

Monthly factors in $\mathbf{x}$

Introduce 12 covariates: January, February, etc. If t is in January the January covariate is 1 otherwise it is 0.

$$x_t = x_{t-1} + \underbrace{\mathbf{Cc}_t}_{\text{drift}} + w_t$$

$$\mathbf{Cc}_t = \alpha_m$$

where m is the month at time step t, so we have a different drift term each month.

Monthly factors

$$
\begin{array}{c}
\text{month effects} \\
\text{a 2 by 12 matrix}
\end{array}
$$

$$
\begin{array}{c}
\text{WA} \\
\text{OR}
\end{array}
\begin{bmatrix}
C_1 & C_2 & C_3 & \ldots & C_{12} \\
C_1 & C_2 & C_3 & \ldots & C_{12}
\end{bmatrix}
$$

$$
\begin{array}{c}
\text{covariates} \\
\text{a 12 by T matrix}
\end{array}
$$

$$
\begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ldots & 0 \\
0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}
\begin{array}{l}
\text{Jan} \\
\text{Feb} \\
\text{Mar} \\
\vdots \\
\text{Dec}
\end{array}
$$

$$
\begin{array}{ccccc}
1 & 2 & 3 & \ldots & T
\end{array} \quad \text{month}
$$

```
TT <- nrow(chinook.month)/2
covariate <- matrix(0, 12, TT)
monrow <- match(chinook.month$Month, month.abb)[1:TT]
covariate[cbind(monrow,1:TT)] <- 1
covariate[,1:12]
```

```
##        [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
##  [1,]    1    0    0    0    0    0    0    0    0     0     0     0
##  [2,]    0    1    0    0    0    0    0    0    0     0     0     0
##  [3,]    0    0    1    0    0    0    0    0    0     0     0     0
##  [4,]    0    0    0    1    0    0    0    0    0     0     0     0
##  [5,]    0    0    0    0    1    0    0    0    0     0     0     0
##  [6,]    0    0    0    0    0    1    0    0    0     0     0     0
##  [7,]    0    0    0    0    0    0    1    0    0     0     0     0
##  [8,]    0    0    0    0    0    0    0    1    0     0     0     0
##  [9,]    0    0    0    0    0    0    0    0    1     0     0     0
## [10,]    0    0    0    0    0    0    0    0    0     1     0     0
## [11,]    0    0    0    0    0    0    0    0    0     0     1     0
## [12,]    0    0    0    0    0    0    0    0    0     0     0     1
```

WA and OR have different month effects.

```
C <- matrix(paste0(month.abb,rep(1:2,each=12)), 2, 12, byrow = TRUE)
C[,1:6]
```

```
##      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
## [1,] "Jan1" "Feb1" "Mar1" "Apr1" "May1" "Jun1"
## [2,] "Jan2" "Feb2" "Mar2" "Apr2" "May2" "Jun2"
```

WA and OR have same month effects.

```
C <- matrix(month.abb, 2, 12, byrow = TRUE)
C[,1:6]
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
## [1,] "Jan" "Feb" "Mar" "Apr" "May" "Jun"
## [2,] "Jan" "Feb" "Mar" "Apr" "May" "Jun"
```

Season as a 3rd order polynomial

Introduce 3 covariates: m, m^2 and m^3 where m is month (1 to 12).

$$x_t = x_{t-1} + \mathbf{C}\mathbf{c}_t + w_t$$

$$\mathbf{C}\mathbf{c}_t = \beta_1 m + \beta_2 m^2 + \beta_3 m^3$$

where m is month at time t.

Season as polynomial

$$
\begin{array}{c}
\text{month effects} \\
\text{a 2 by 3 matrix}
\end{array}
\qquad
\begin{array}{c}
\text{covariates} \\
\text{a 3 by T matrix}
\end{array}
$$

$$
\begin{array}{c}
\text{WA} \\
\text{OR}
\end{array}
\begin{bmatrix}
C_1 & C_2 & C_3 \\
C_1 & C_2 & C_3
\end{bmatrix}
\begin{bmatrix}
1 & 2 & 3 & \ldots & 12 \\
1^2 & 2^2 & 3^2 & \ldots & 12^2 \\
1^3 & 3^3 & 3^3 & \ldots & 12^3
\end{bmatrix}
\begin{array}{c}
m \\
m^2 \\
m^3
\end{array}
$$

$$
\begin{array}{ccccc}
1 & 2 & 3 & \ldots & T \qquad \text{month}
\end{array}
$$

```
TT <- nrow(chinook.month)/2
monrow <- match(chinook.month$Month, month.abb)[1:TT]
covariate <- rbind(monrow, monrow^2, monrow^3)
rownames(covariate) <- c("m", "m2", "m3")
covariate[,1:13]
```

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## m     1    2    3    4    5    6    7    8    9    10    11    12     1
## m2    1    4    9   16   25   36   49   64   81   100   121   144     1
## m3    1    8   27   64  125  216  343  512  729  1000  1331  1728     1
```

WA and OR have different seasonal pattern.

```
C <- matrix(paste0(c("m", "m2", "m3"),".",rep(1:2,each=3)), 2, 3,
            byrow = TRUE)
C
```

```
##      [,1]   [,2]    [,3]
## [1,] "m.1"  "m2.1"  "m3.1"
## [2,] "m.2"  "m2.2"  "m3.2"
```

WA and OR have same seasonal pattern.

```
C <- matrix(c("m", "m2", "m3"), 2, 3, byrow = TRUE)
C
```

```
##      [,1] [,2] [,3]
## [1,] "m"  "m2" "m3"
## [2,] "m"  "m2" "m3"
```

Season as a Fourier series

- Fourier series are paired sets of sine and cosine waves
- They are commonly used in time series analysis in the frequency domain

Season as a Fourier series

Introduce 2 covariates: $sin(2\pi t/p)$, $cos(2\pi t/p)$ where p is period (12 for monthly) and t is the time step (1 to T).

$$x_t = x_{t-1} + \mathbf{Cc}_t + w_t$$

$$\mathbf{Cc}_t = \beta_1 sin(2\pi t/p) + \beta_2 cos(2\pi t/p)$$

where p is 12 (for monthly).

covariates sines and cosines
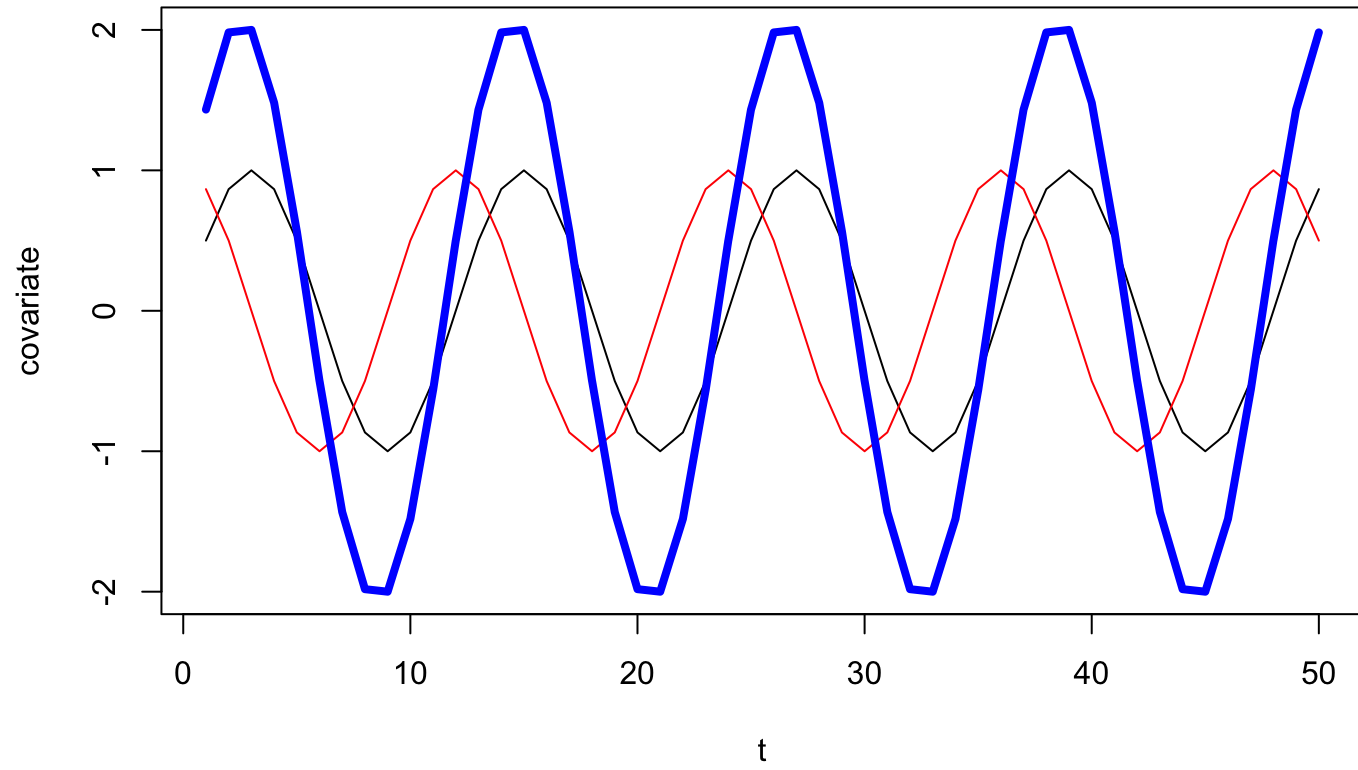
forecast `fourier()` function

You can also use the `fourier()` function in the forecast package to create your covariates.

```
x <- ts(1:TT, freq=12)
covariate <- forecast::fourier(x, K=1) |> t()
```

This makes it easy to create a longer fourier series to model non-symmetric or multi-modal seaonal cycles.

```
x <- ts(1:TT, freq=12)
covariate2 <- forecast::fourier(x, K=2) |> t()
```

K=2 for fourier transform and now we can have 2 peaks
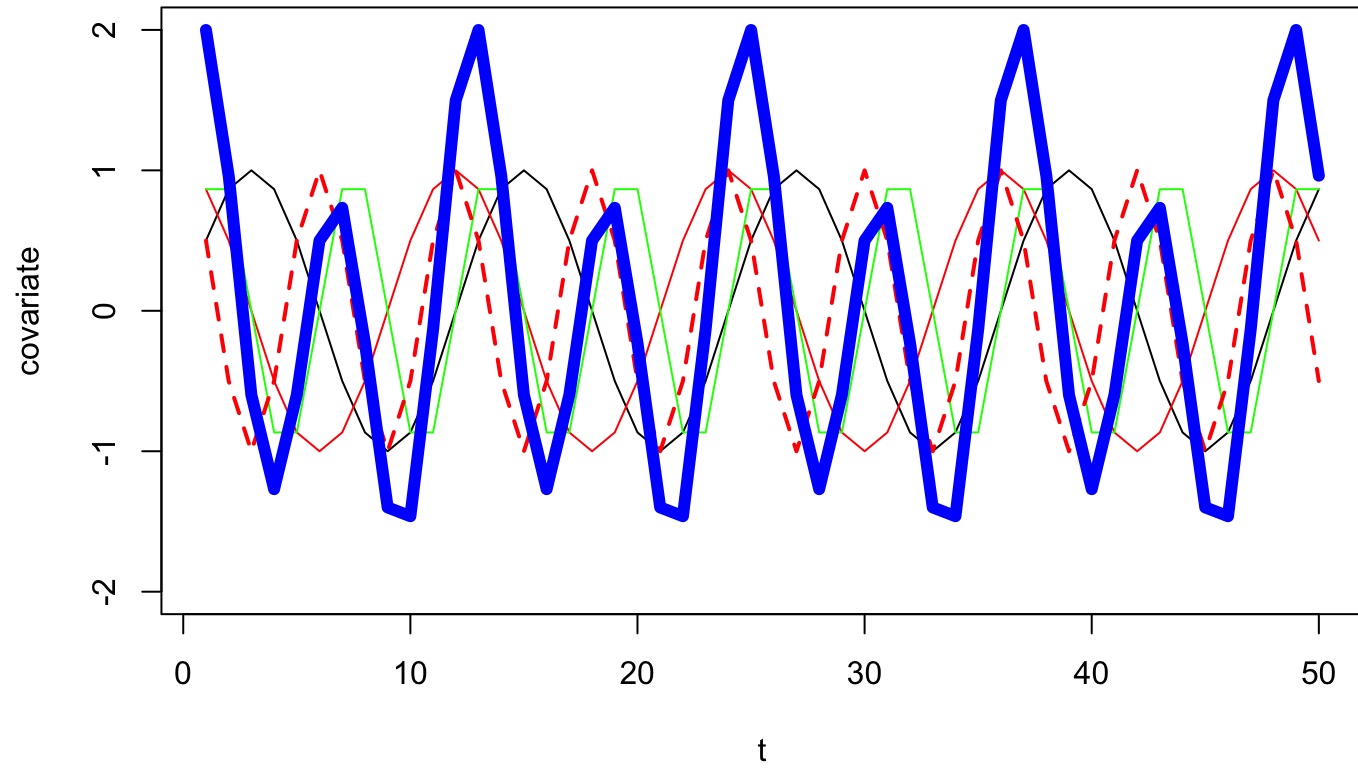
Season as a Fourier series (K=1)

$$
\begin{array}{c}
\text{covariates} \\
\text{a 2 by T matrix}
\end{array}
$$

$$
\begin{array}{c} \text{WA} \\ \text{OR} \end{array}
\begin{bmatrix} C_1 & C_2 \\ C_1 & C_2 \end{bmatrix}
\begin{bmatrix} sin\left(\frac{2\pi 1}{12}\right) & sin\left(\frac{2\pi 2}{12}\right) & \dots & sin\left(\frac{2\pi T}{12}\right) \\ cos\left(\frac{2\pi 1}{12}\right) & cos\left(\frac{2\pi 2}{12}\right) & \dots & cos\left(\frac{2\pi T}{12}\right) \end{bmatrix}
$$

```r
TT <- nrow(chinook.month)/2
covariate <- rbind(sin(2*pi*(1:TT)/12), cos(2*pi*(1:TT)/12))
plot(covariate[1,1:50], type="l")
lines(covariate[2,1:50], col="red")
title("covariates sines and cosines")
```


covariates sines and cosines
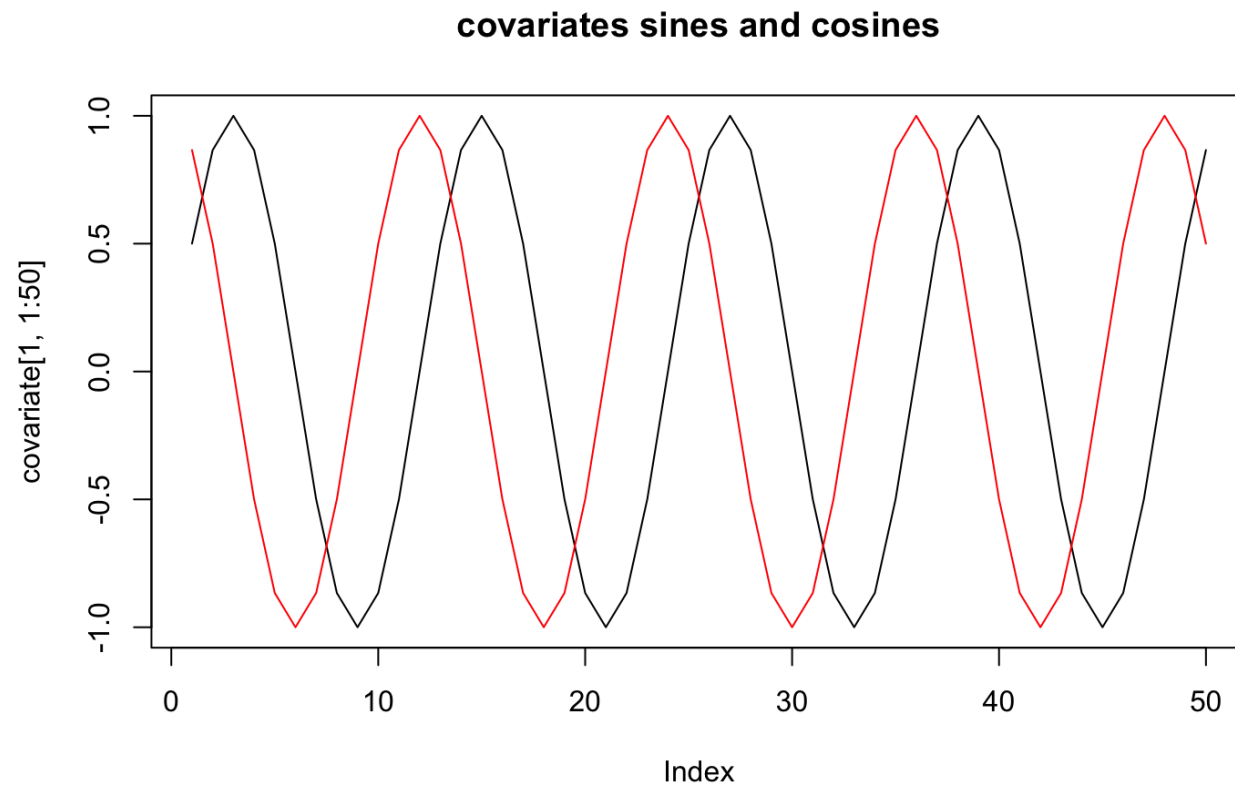
WA and OR have different seasonal pattern.

```
C <- matrix(paste0(c("s", "c"),".",rep(1:2,each=2)), 2, 2, byrow = TRUE)
C
```

```
##      [,1]  [,2]
## [1,] "s.1" "c.1"
## [2,] "s.2" "c.2"
```
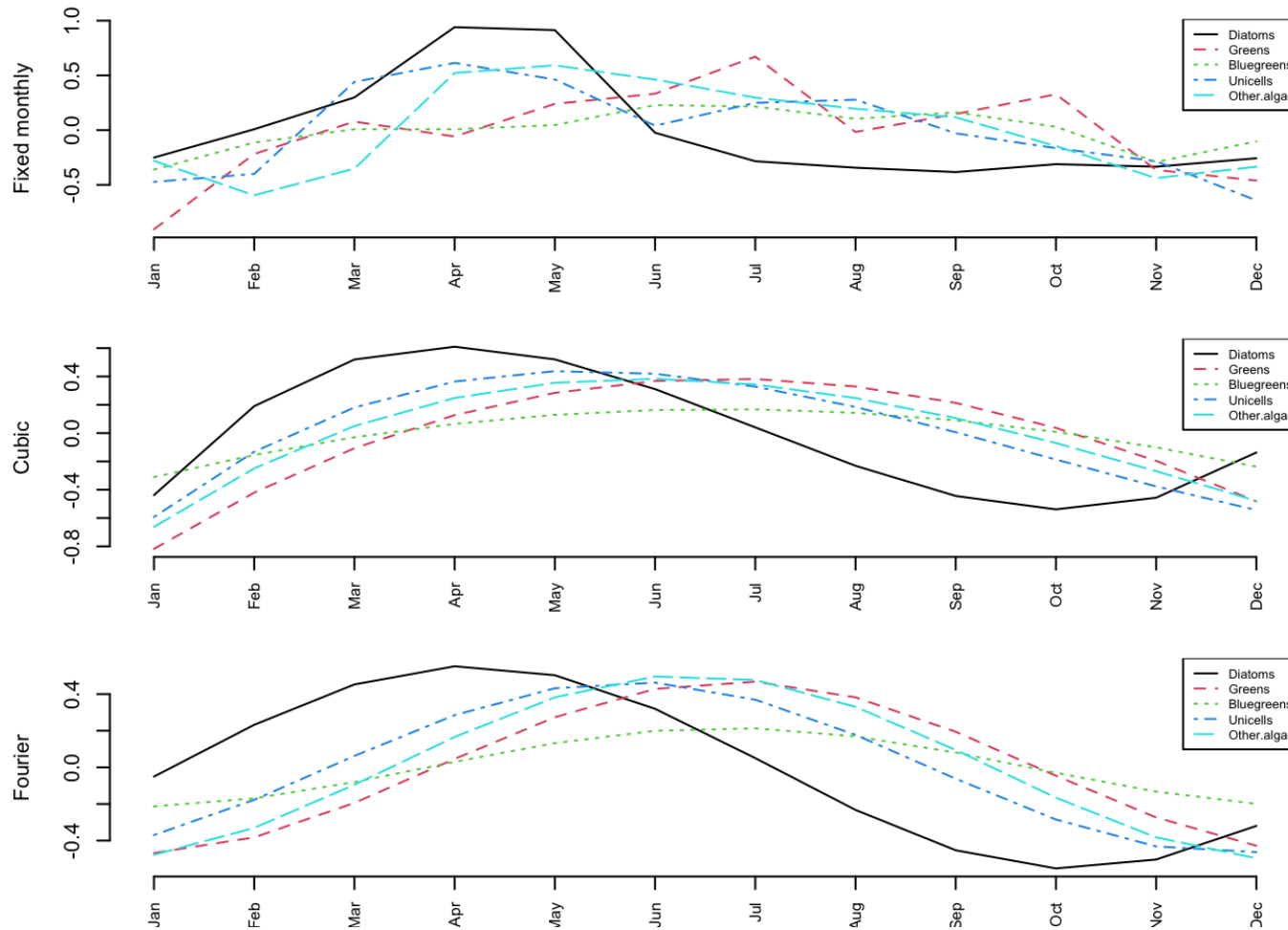
WA and OR have same seasonal pattern.

```
C <- matrix(c("s", "c"), 2, 2, byrow = TRUE)
C
```

```
##      [,1] [,2]
## [1,] "s"  "c"
## [2,] "s"  "c"
```
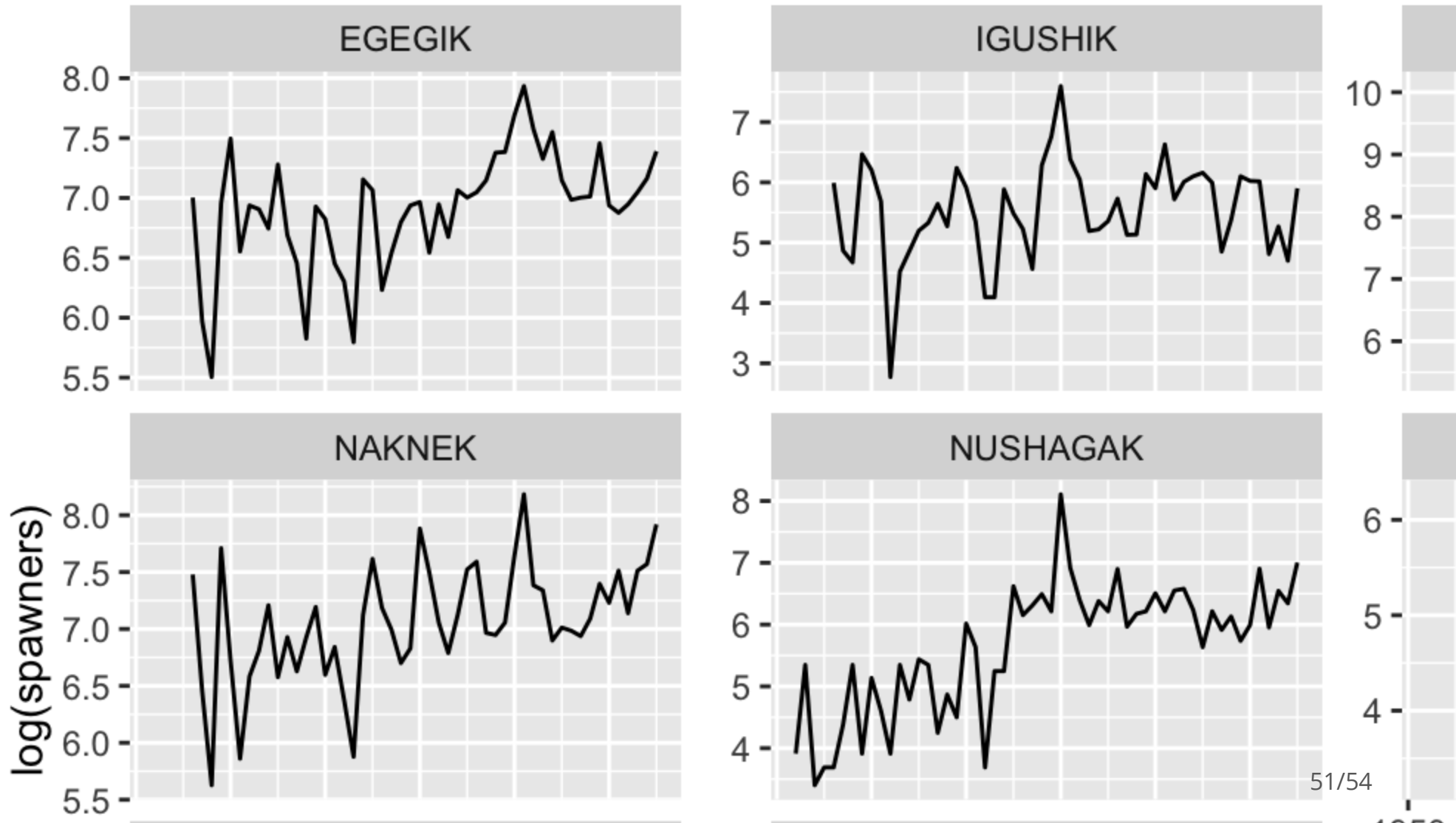
Example Section 8.6

Seasonality of Lake WA plankton

Cyclic salmon

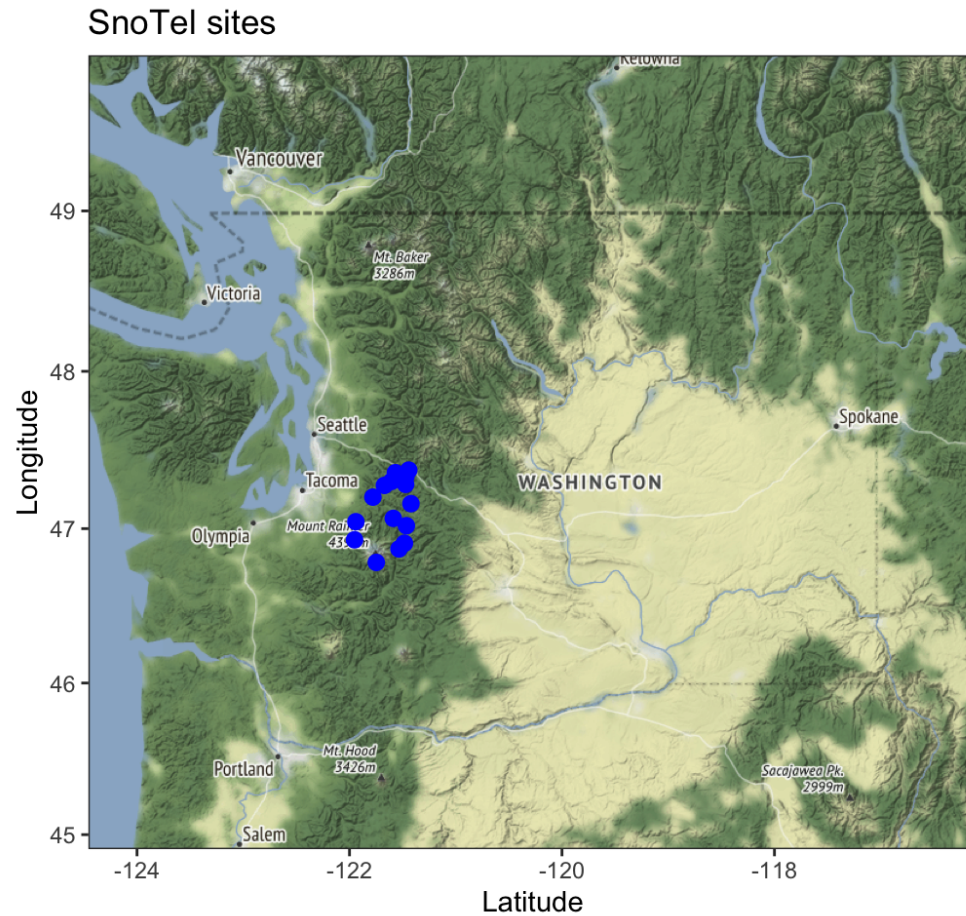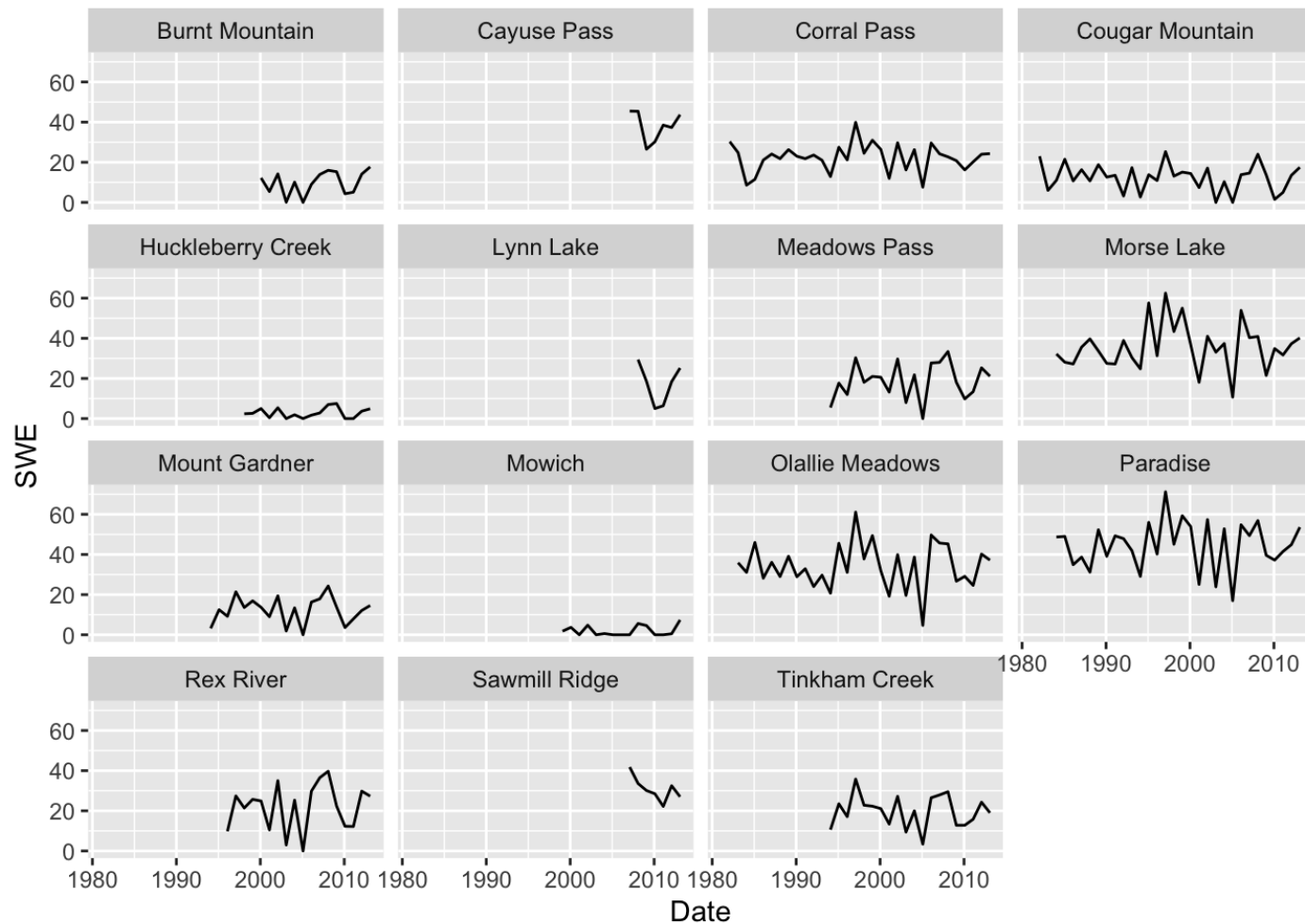## log spawners

Missing covariates

SNOTEL Example Chapter 11

https://atsa-es.github.io/atsa-labs/example-snotel-data.html

SnoTel sites

Snow Water Equivalent (snowpack)

February snowpack estimates

Use MARSS models Chapter 11

- Accounts for correlation across sites and local variability