Ecological forecasting in R

Lecture 1: introduction to time series models

Nicholas Clark

School of Veterinary Science, University of Queensland

0900–1200 CET Monday 27th May, 2024

Welcome

About me

Australian Research Council Early Career Fellow

The University of Queensland

School of Veterinary Science

Located in Gatton, Australia

Expertise in:

Quantitative ecology

Molecular genetics

Multivariate time series modelling

Workflow

Press the "o" key on your keyboard to navigate among slides

Access the tutorial html here

- Download the data objects and exercise `R` script from the html file
- Complete exercises and use Slack to ask questions

Relevant open-source materials include:

- Forecasting Principles and Practice
- Applied Time Series Analysis
- Ecological Forecasting & Dynamics Course
- How to interpret nonlinear effects from GAMs

This lecture's topics

Why forecast?

Why are time series difficult?

Visualizing time series

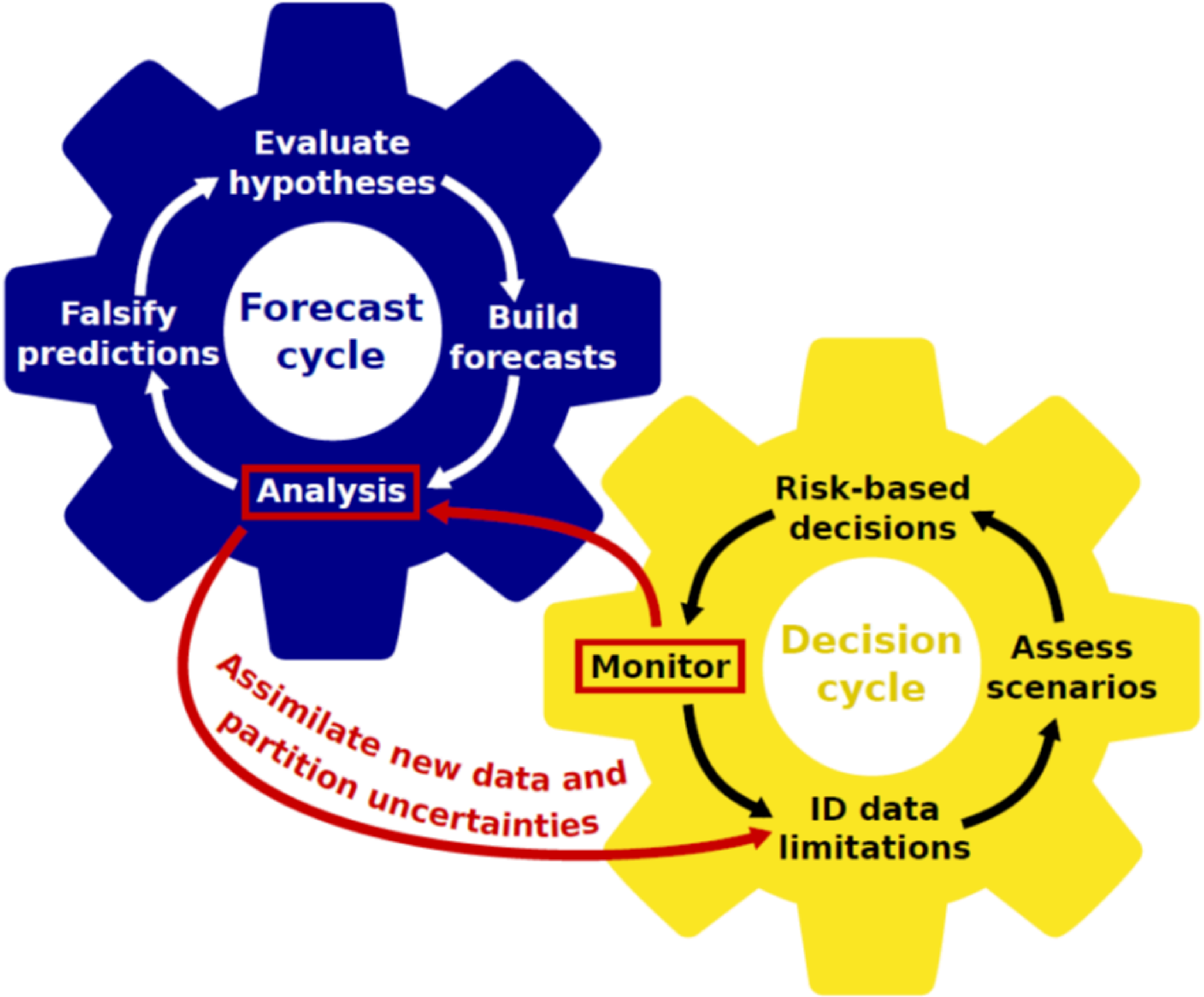
Common time series models

Why they fail in ecology

Why forecast?

“Because all decision making is based on what will happen in the future, either under the status quo or different decision alternatives, decision making ultimately depends on forecasts”

Dietze et al. 2018


Evaluate hypotheses
Build forecasts
Forecast cycle
Falsify predictions
Analysis
Risk-based decisions
Assess scenarios
Decision cycle
Monitor
ID data limitations
Assimilate new data and partition uncertainties

Where is forecasting used?

Fisheries stocks, landings and bycatch risks

Coral bleaching and algal bloom risks

Carbon stocks

Wildlife population dynamics

Many other examples

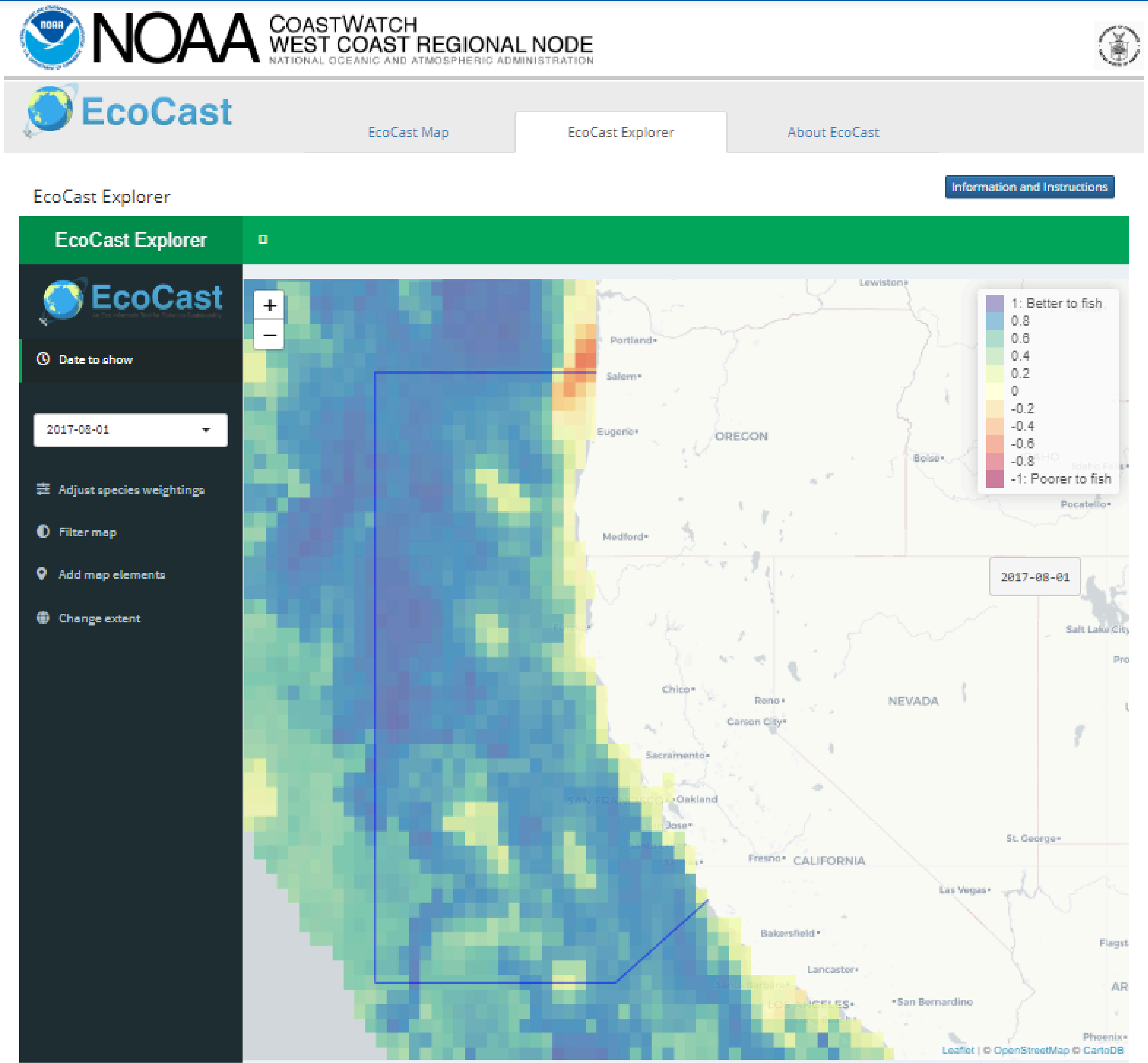
NOAA Coastwatch's EcoCast

Tell fishers where to avoid bycatch

Harnesses up-to-date information for ecological models:

- Fisheries bycatch data
- Satellite observations
- Oceanography products

Builds distribution models and dynamically updates maps


NOAA
COASTWATCH
WEST COAST REGIONAL NODE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
EcoCast
EcoCast Map
EcoCast Explorer
About EcoCast
EcoCast Explorer
Information and Instructions
EcoCast Explorer
EcoCast
Date to show
2017-08-01
Adjust species weightings
Filter map
Add map elements
Change extent
1: Better to fish
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1: Poorer to fish
2017-08-01
Portland
Salem
Eugene
OREGON
Boise
Medford
Pocatello
Salt Lake City
Chico
Reno
Carson City
NEVADA
Sacramento
Oakland
San Jose
Fresno
CALIFORNIA
St. George
Las Vegas
Bakersfield
Lancaster
San Bernardino
Phoenix
Lewiston
Leaflet | © OpenStreetMap © CartoDB

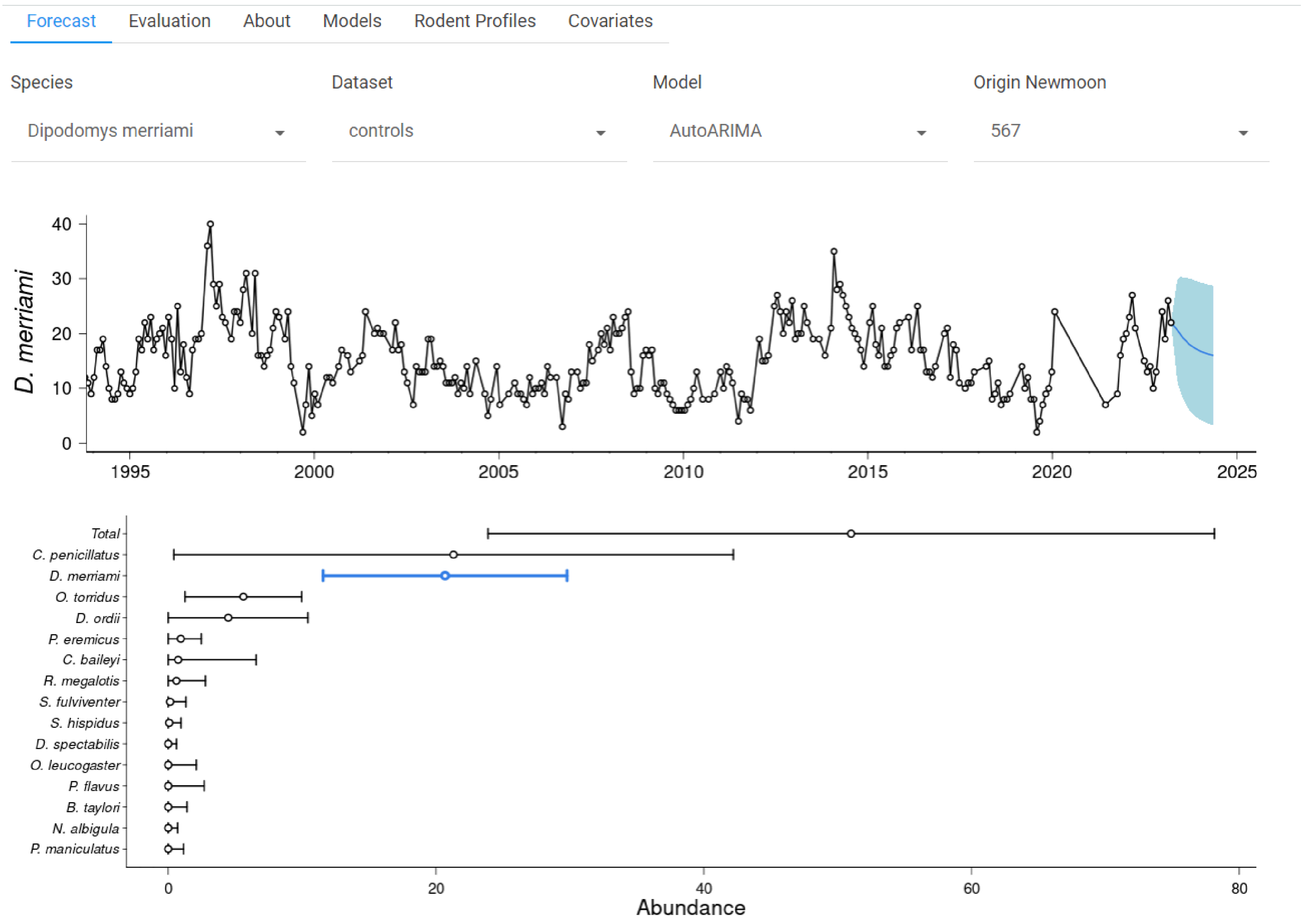
Portal Project's `Portalcast`

Predict rodent abundance up to one year ahead

Harnesses up-to-date information for ecological models:

- Rodent captures from baited traps
- Satellite observations

Builds time series models and dynamically update forecasts



Forecast
Evaluation
About
Models
Rodent Profiles
Covariates
Species
Dipodomys merriami
Dataset
controls
Model
AutoARIMA
Origin Newmoon
567
D. merriami
0
10
20
30
40
1995
2000
2005
2010
2015
2020
2025
Total
C. penicillatus
D. merriami
O. torridus
D. ordii
P. eremicus
C. baileyi
R. megalotis
S. fulviventer
S. hispidus
D. spectabilis
O. leucogaster
P. flavus
B. taylori
N. albigula
P. maniculatus
0
20
40
60
80
Abundance

Why are time series difficult?

Some challenges of time series

Temporal autocorrelation

Lagged effects

Non-Gaussian data and missing observations

Measurement error

Time-varying effects

Nonlinearities

Multi-series clustering

Let's focus on these for now

Temporal autocorrelation

Lagged effects

Non-Gaussian data and missing observations

Measurement error

Time-varying effects

Nonlinearities

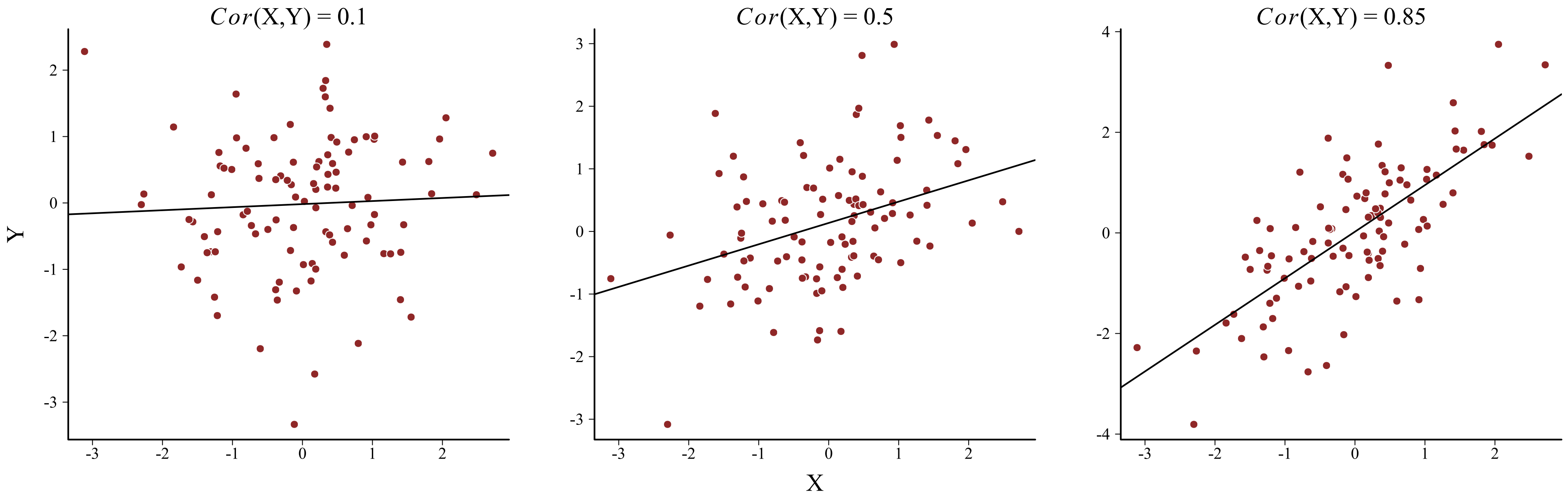
Multi-series clustering

What is temporal autocorrelation?

Values at current time ***correlated with past values***

$$Cor(Y_t, Y_{t-lag}) \neq 0$$

Refresher: what is correlation?



Correlation assumes a ***linear*** relationship among two variables

What is temporal autocorrelation?

Values at current time ***correlated with past values***

$$Cor(Y_t, Y_{t-lag}) \neq 0$$

We can estimate the correlation (β) with linear regression

$$\boldsymbol{Y}_t \sim \text{Normal}(\alpha + \beta * \boldsymbol{Y}_{t-lag}, \sigma)$$

What is temporal autocorrelation?

Values at current time ***correlated with past values***

$$Cor(Y_t, Y_{t-lag}) \neq 0$$

We can estimate the correlation (β) with linear regression

$$\boldsymbol{Y_t} \sim \text{Normal}(\alpha + \beta * \boldsymbol{Y_{t-lag}}, \sigma)$$

Generalize to state that current value of a series (at time t) is ***a function*** of it's own past values (at time $t - lag$)

$$\boldsymbol{Y_t} \sim \text{f}(\boldsymbol{Y_{t-lag}})$$

A *positively* autocorrelated series

Code Model Plot

```
# set seed for reproducibility
set.seed(1111)

# number of timepoints
T <- 100

# use arima.sim to simulate from an AR(1) model
series <- arima.sim(model = list(ar = 0.8), n = T, sd = 1)

# plot the time series as a line
plot(series, type = 'l', bty = 'l', lwd = 2,
     col = 'darkred', ylab = 'Y', xlab = 'Time')
```

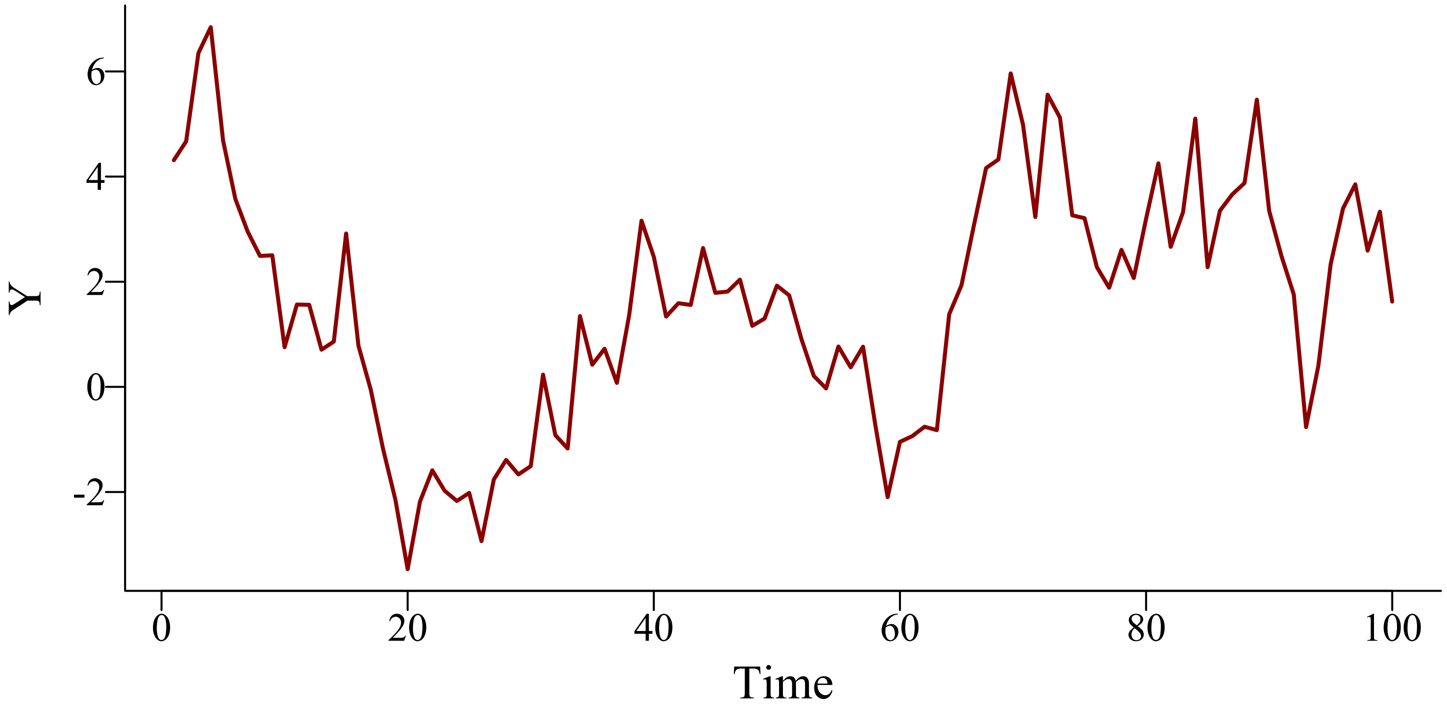
A *positively* autocorrelated series

Code Model Plot

$$\boldsymbol{Y}_t \sim \text{Normal}(0.8 * \boldsymbol{Y}_{t-1}, 1)$$

A *positively* autocorrelated series

Code Model Plot



A *negatively* autocorrelated series

Code Model Plot

```
# set seed for reproducibility
set.seed(1111)

# number of timepoints
T ← 100

# use arima.sim to simulate from an AR(1) model
series ← arima.sim(model = list(ar = -0.8), n = T, sd = 1)

# plot the time series as a line
plot(series, type = 'l', bty = 'l', lwd = 2,
     col = 'darkred', ylab = 'Y', xlab = 'Time')
```

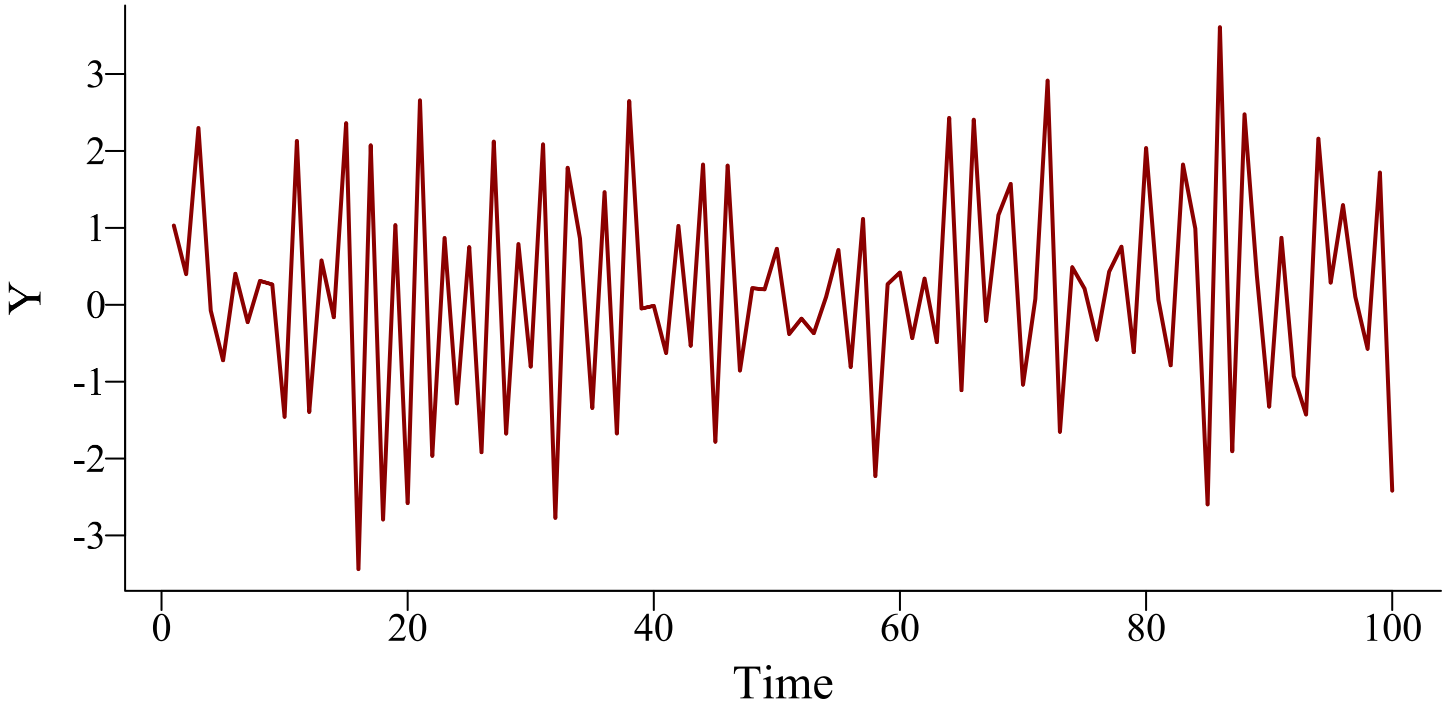
A *negatively* autocorrelated series

Code | Model | Plot

$$\boldsymbol{Y}_t \sim \text{Normal}(-0.8 * \boldsymbol{Y}_{t-1}, 1)$$

A *negatively* autocorrelated series

Code Model Plot



Correlations *over >1 lag*

Can include multiple lags of the same predictor variable (the response in this case)

$$\boldsymbol{Y}_t \sim \mathrm{f}(\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-2}, \boldsymbol{Y}_{t-3})$$

Lagged effects of predictors

External conditions (eg temperature, humidity, landcover) can also influence what happens to a series at later timepoints

$$\boldsymbol{Y}_t \sim \mathrm{f}(\boldsymbol{Y}_{t-lag}, \boldsymbol{X}_{t-lag})$$

Where:

$\boldsymbol{X}_t$ is the matrix of predictor values at time t

Many series show complex correlation structures; they can also show other temporal patterns

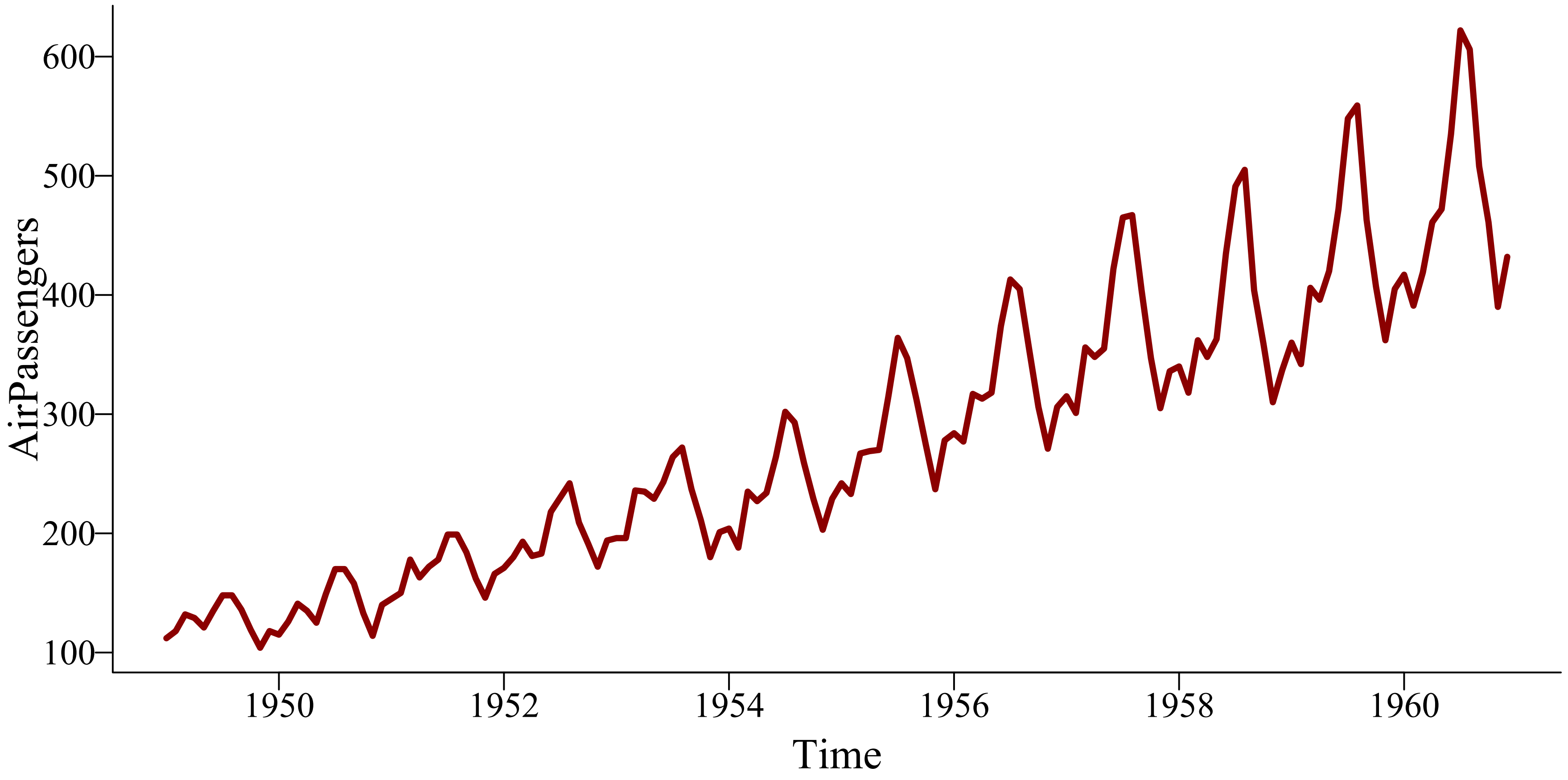
Seasonality

Many time series show
repeated periodic cycles

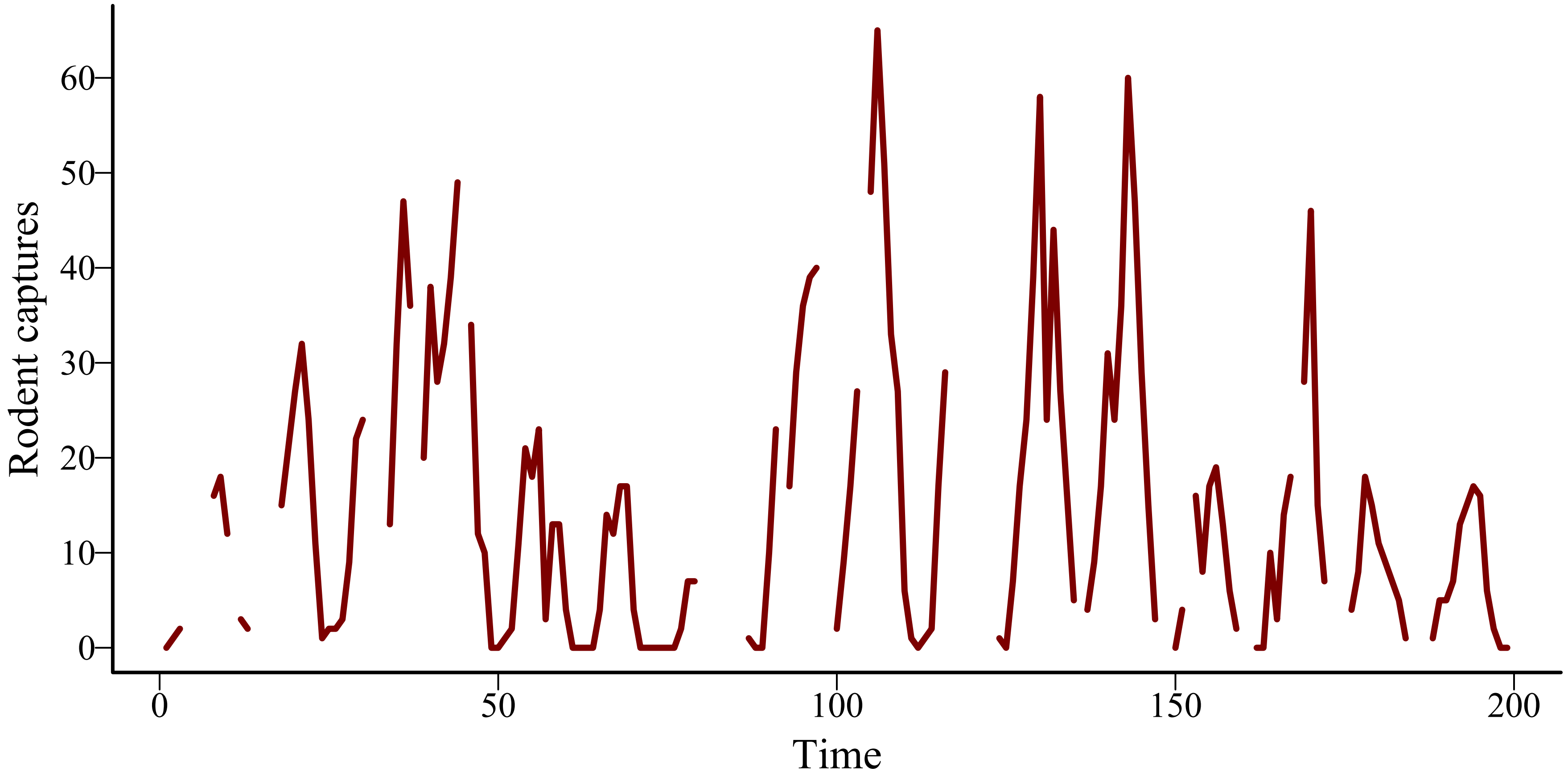
- Breeding seasons
- Migration
- Green-ups / green-downs
- Lunar cycles
- Predator / prey dynamics

Often change slowly over time

Example seasonal series



Another seasonal series



Visualizing time series

Detecting lagged effects

Lag plots

Autocorrelation functions (ACFs)

Partial autocorrelation functions (pACFs)

Cross-correlation functions (CCFs)

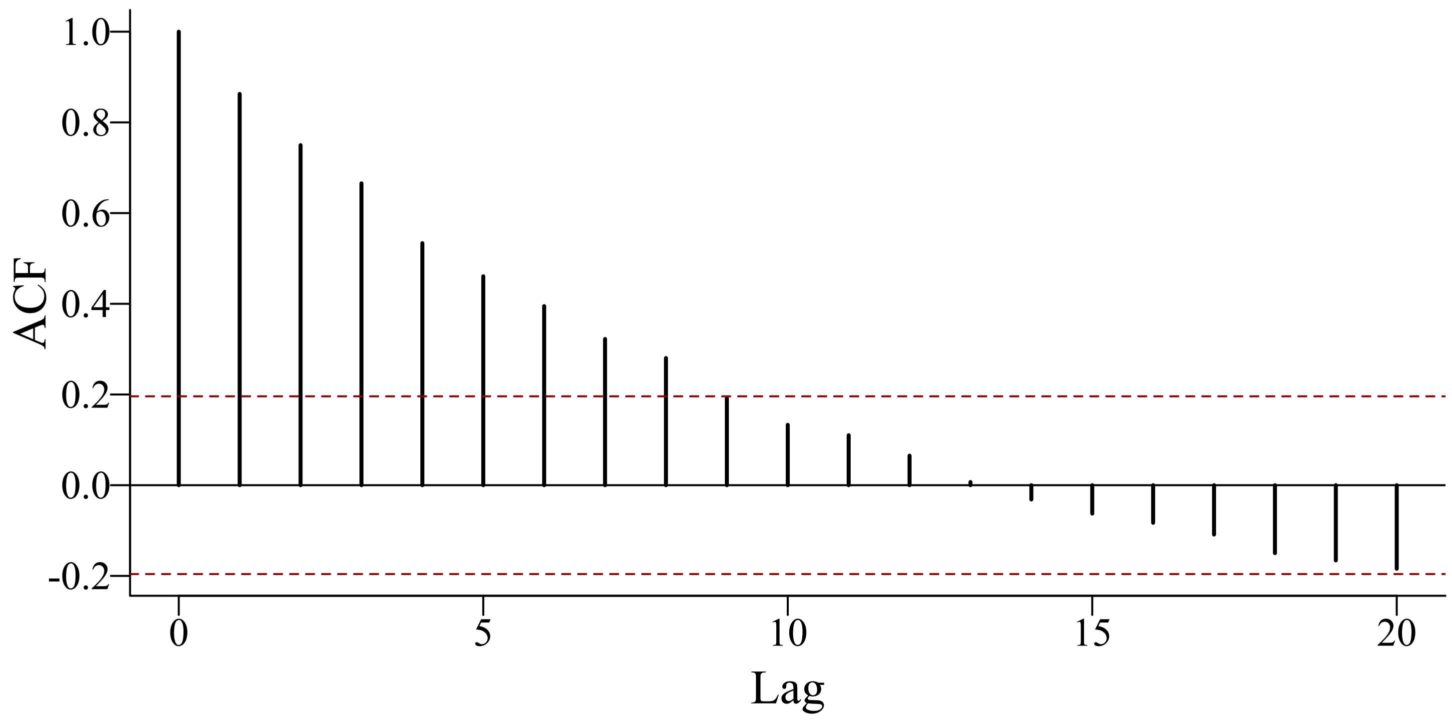
Independent correlations

Code Plot

```r
# set seed for reproducibility
set.seed(1111)

# number of timepoints
T <- 100

# use arima.sim to simulate from an AR(1) model
series <- arima.sim(model = list(ar = c(0.8)), n = T, sd = 1)

# plot the empirical ACF
acf(series, lwd = 2, bty = 'l',
    ci.col = 'darkred', main = '')
```

Independent correlations

Code Plot



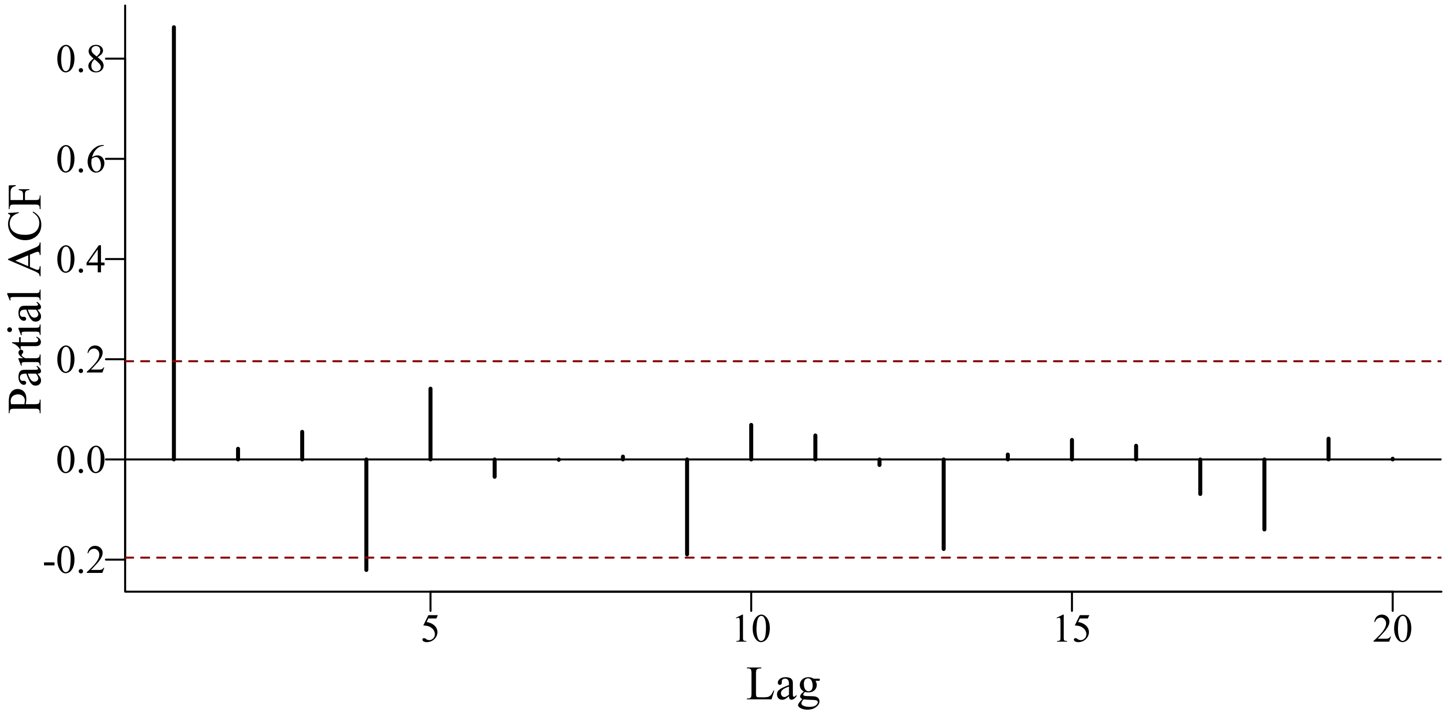
Conditional correlations

Code Plot

```r
# set seed for reproducibility
set.seed(1111)

# number of timepoints
T ← 100

# use arima.sim to simulate from an AR(1) model
series ← arima.sim(model = list(ar = c(0.8)), n = T, sd = 1)

# plot the empirical pACF
pacf(series, lwd = 2, bty = 'l',
     ci.col = 'darkred', main = '')
```

Conditional correlations

Code Plot



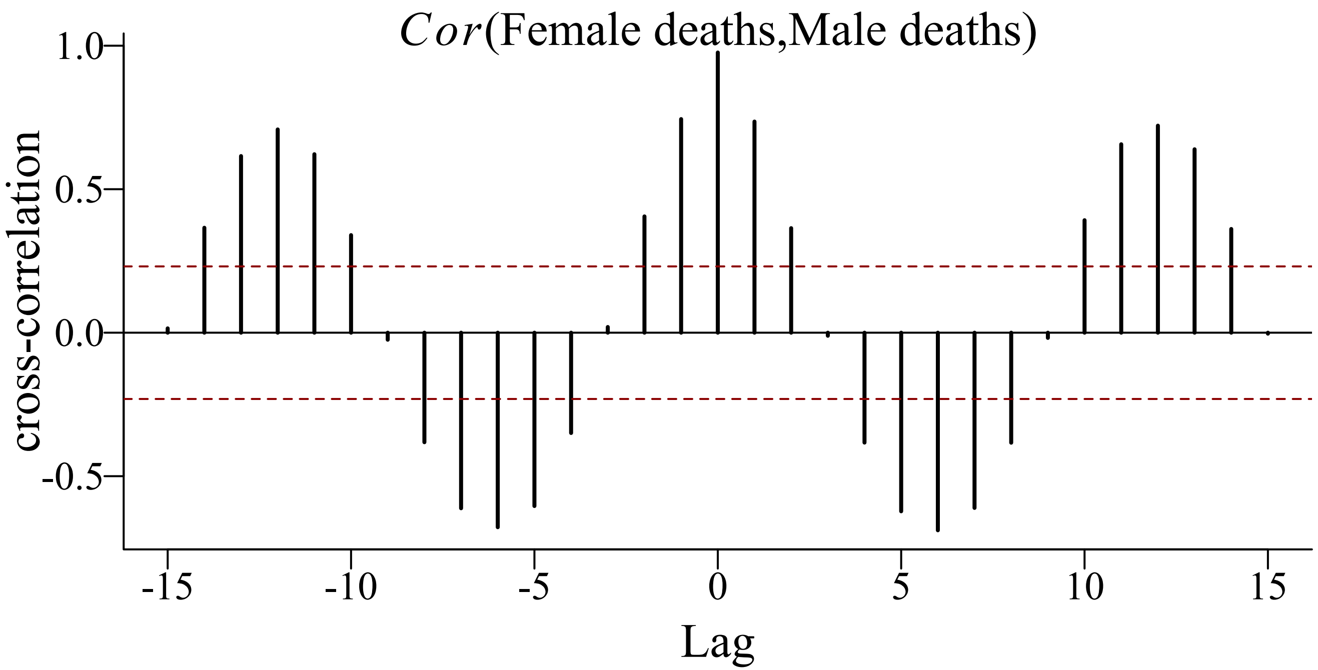
Independent cross-correlations

Code Plot

```
# compute a CCF of the built-in lung cancer dataset
ccf(as.vector(mdeaths), as.vector(fdeaths),
    # compute cross-correlations at each lag
    type = 'correlation',
    bty = 'l', lwd = 2, ci.col = 'darkred',
    ylab = "cross-correlation", main = "")
# add an informative title
title(main = expression(paste(italic(Cor),
                              "(", Female~deaths,
                              ",", Male~deaths, ")")),
          line = 0)
```

Independent cross-correlations

Code Plot



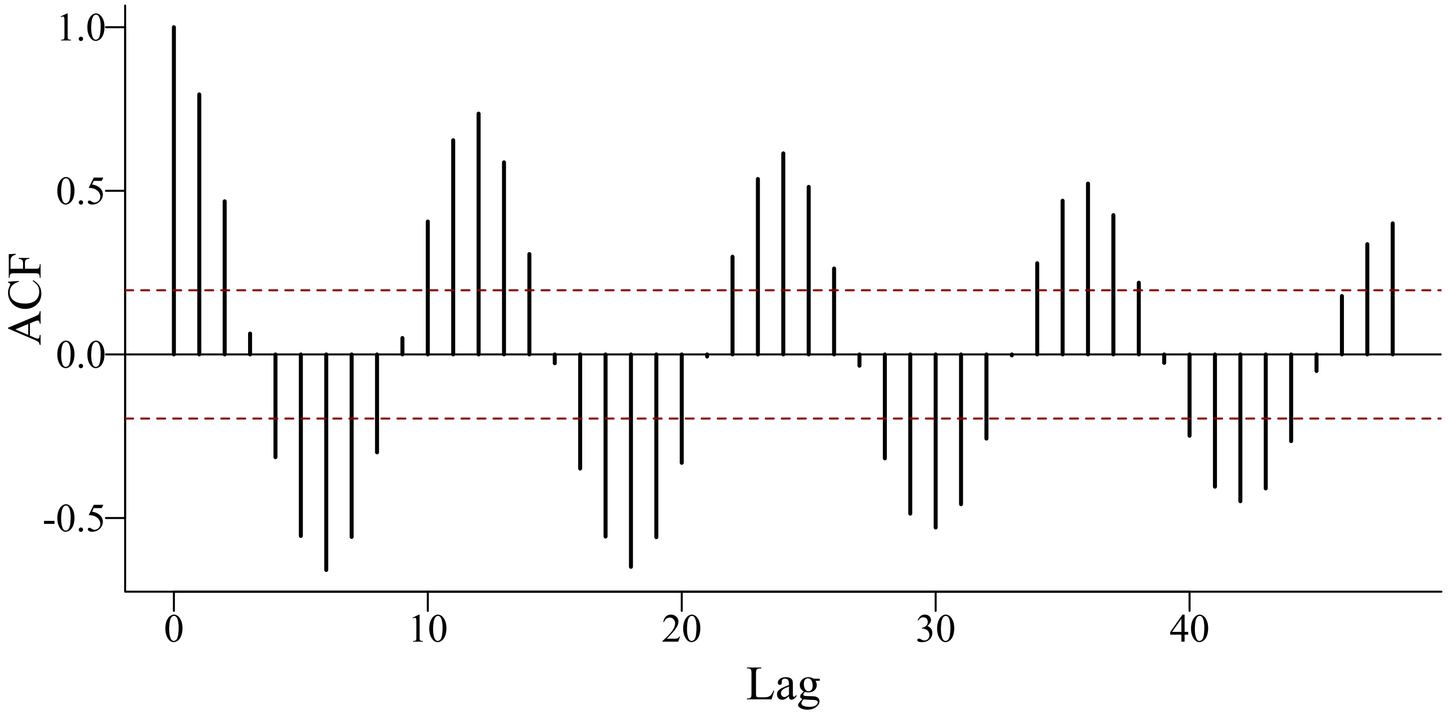
ACFs often detect seasonality

Code Plot

```r
# load the 'gas' dataset from the forecast library
library(forecast)
data(gas)

# subset to the final 100 observations
gas <- gas[377:476]

# plot the empirical ACF over 48 lags
acf(gas, lag.max = 48, lwd = 2, bty = 'l',
    ci.col = 'darkred', main = '')
```

ACFs often detect seasonality

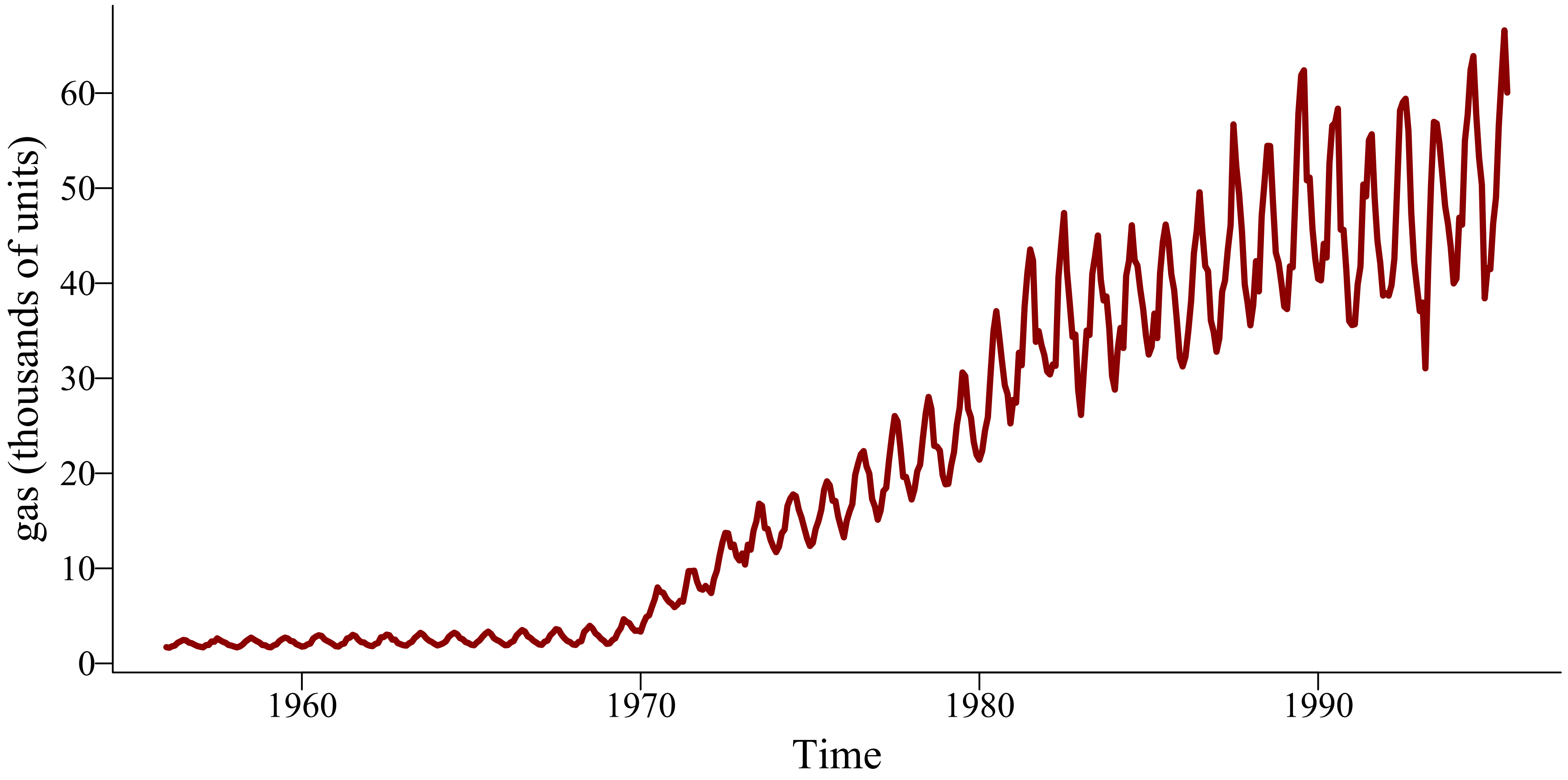
Code Plot



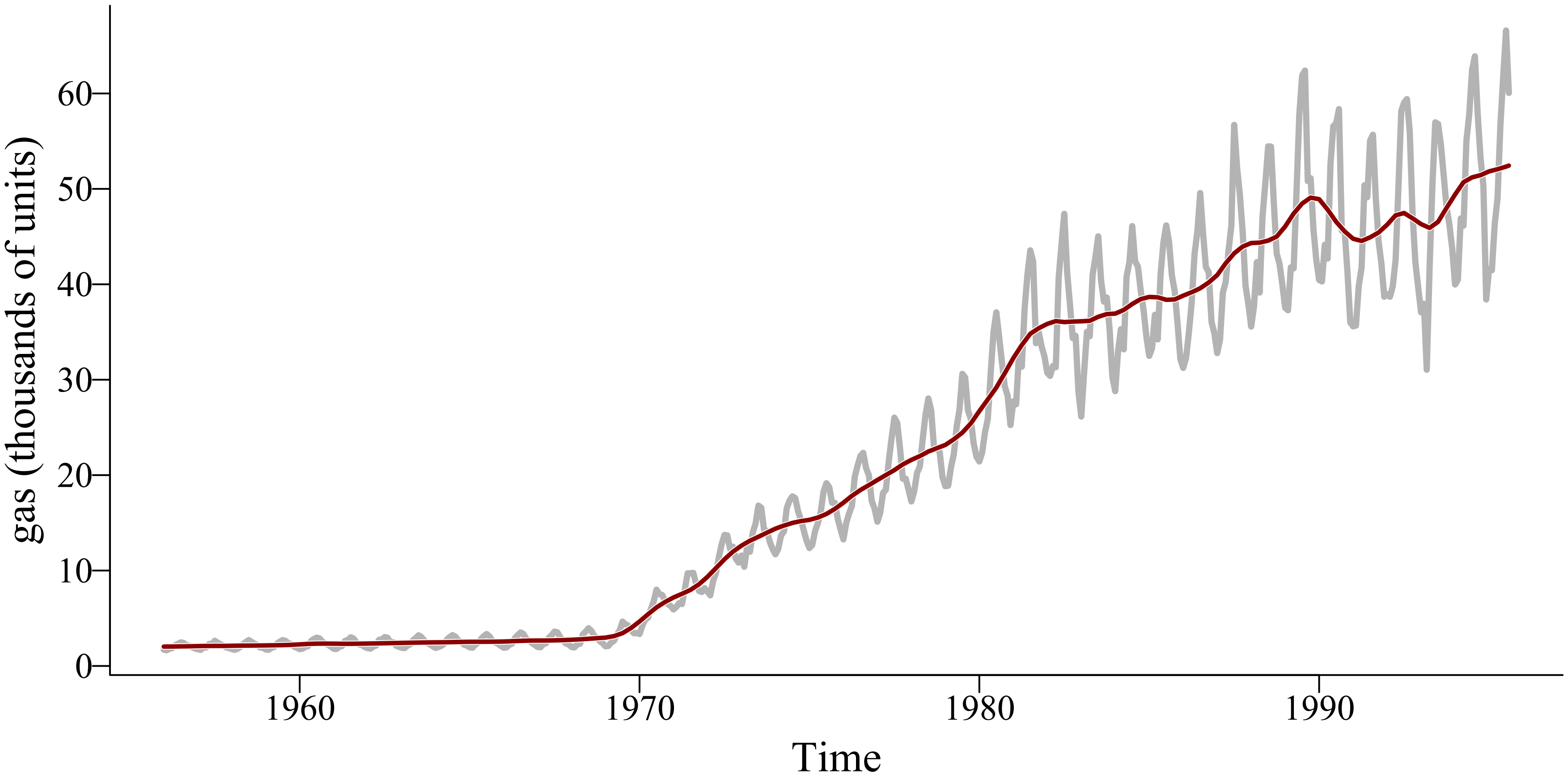
But why did we subset?

```r
# load the 'gas' dataset from the forecast library
library(forecast)
data(gas)

# subset to the final 100 observations
gas <- gas[377:476]

# plot the empirical ACF over 48 lags
acf(gas, lag.max = 48, lwd = 2, bty = 'l',
    ci.col = 'darkred', main = '')
```

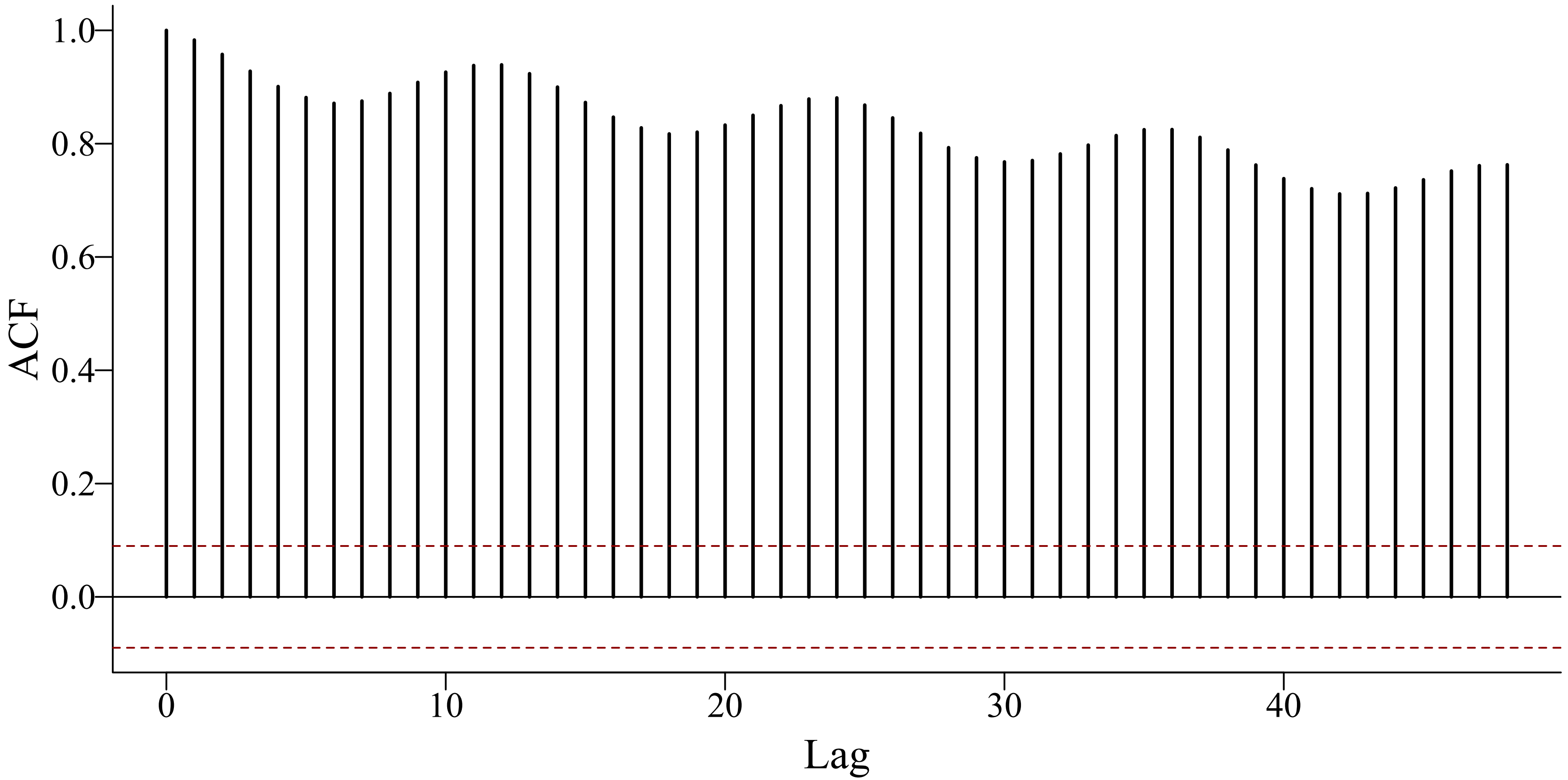
Because gas looks like this ...



... and has a nonlinear *trend*



Raw ACF is misleading



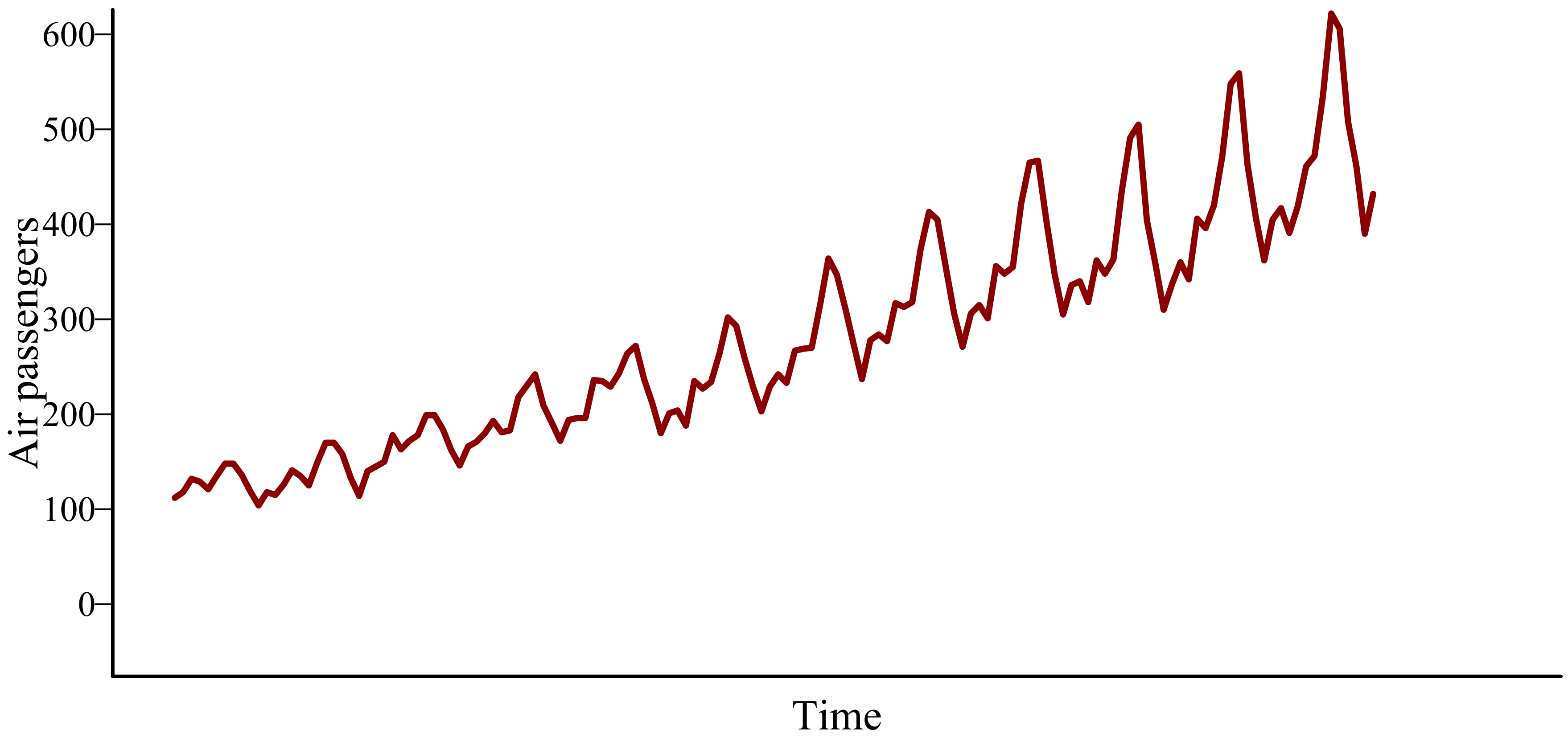
Decompositions

Often it is helpful to split (i.e. decompose) a time series into several sub-components

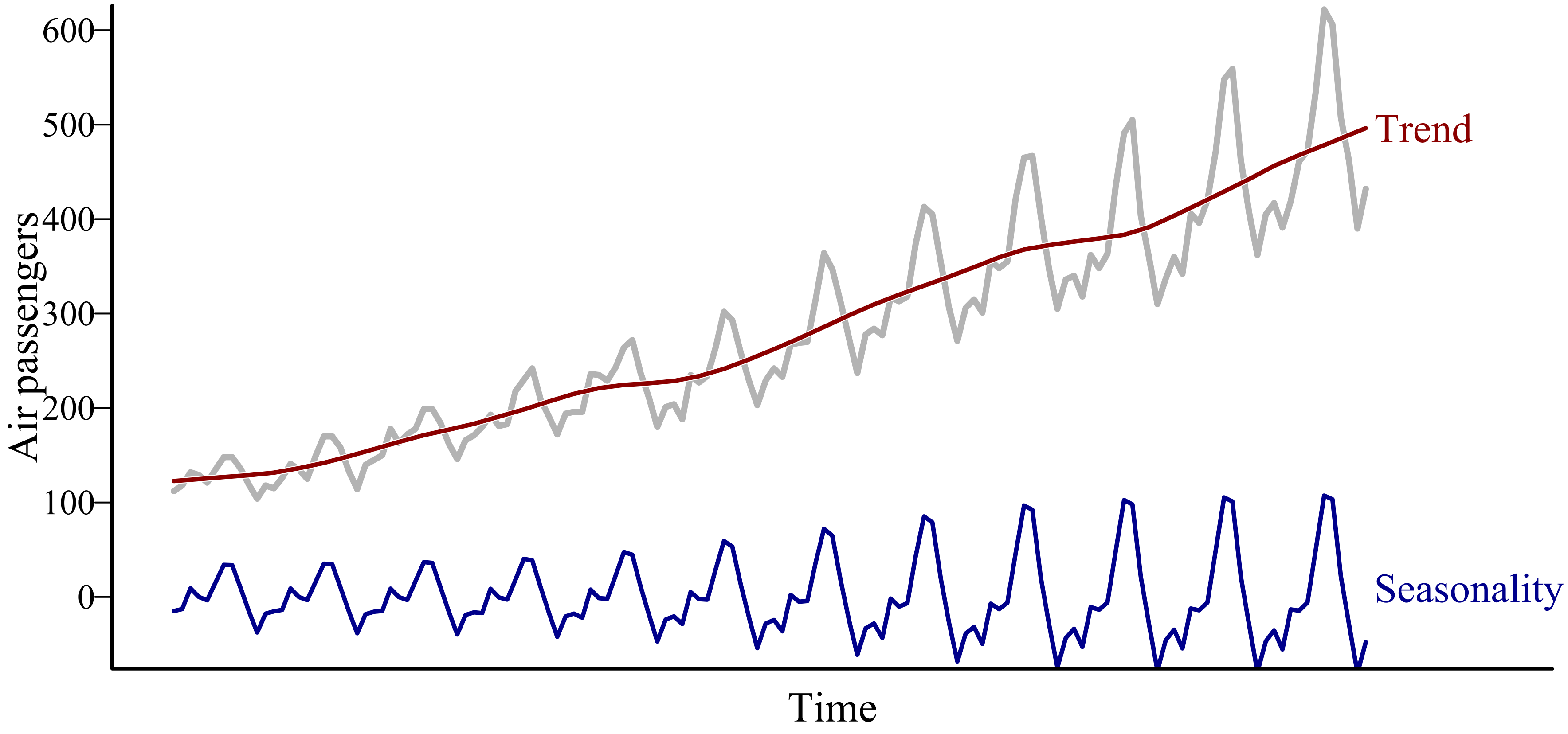
- Long-term trends
- Repeated seasonal patterns
- Remaining non-temporal variation

These components can be summed to give the original series

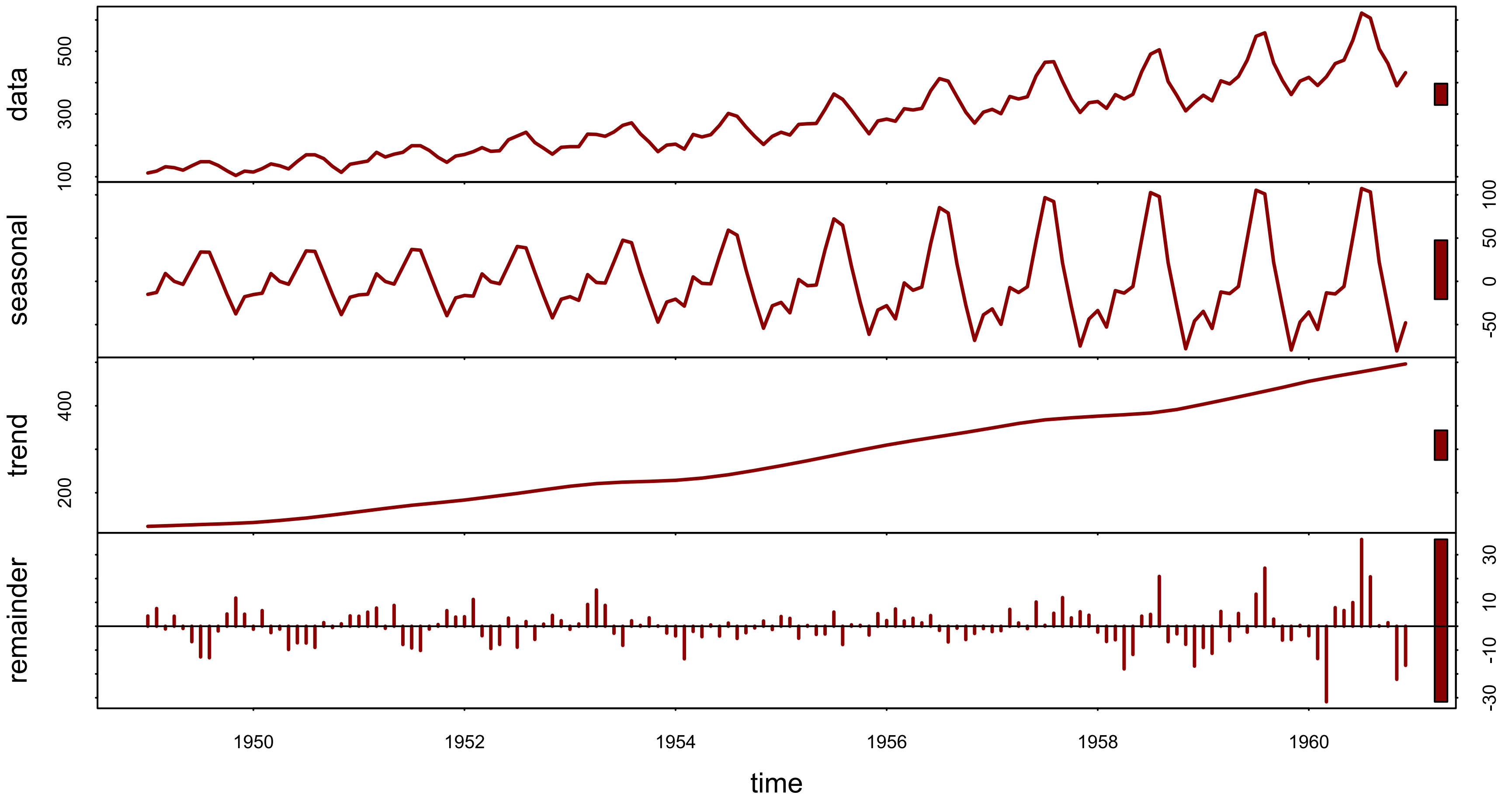
Example: a complex series



Decompose: trend + seasonality



Under the hood



Modelling these multiple components, either additively or multiplicatively, is a major goal of most time series analysis procedures

Common time series models

Common time series models

Random Walk (RW)

Autoregressive (AR)

Autoregressive Integrated Moving Average (ARIMA; require stationarity)

Exponential Smoothing (ETS)

Regression with ARIMA errors

Very easy to apply in R

Hyndman's tools in the `forecast` 📦 are hugely popular and accessible for time series analysis / forecasting

ETS handles many types of seasonality and nonlinear trends

Regression with ARIMA errors includes additive fixed effects of predictors while capturing trends and seasonality

Some of these algorithms can handle missing data

All are extremely fast to fit and forecast

Great! But what about these?

Temporal autocorrelation

Lagged effects

Non-Gaussian data and missing observations

Measurement error

Time-varying effects

Nonlinearities

Multi-series clustering

Time series models fail in ecology

Ecological time series include

Counts of multiple species over time

Presence-absence of species

Repeated captures in multiple plots

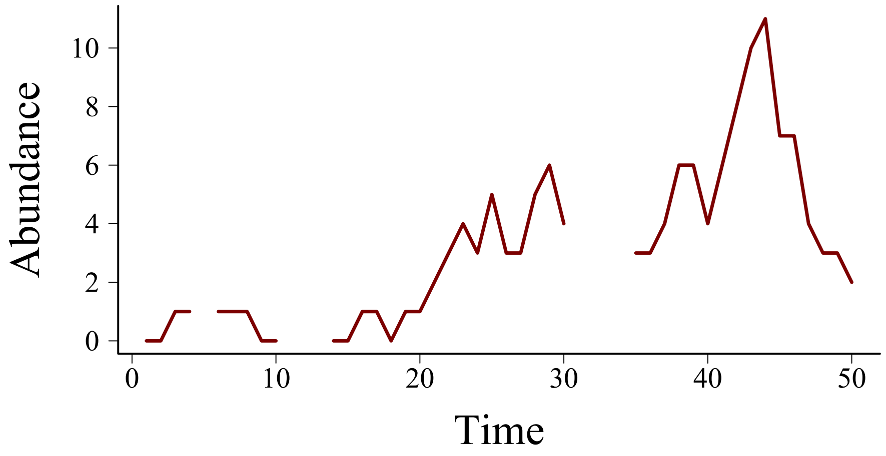
Censored measures (OTUs / pollutants with limits of detection)

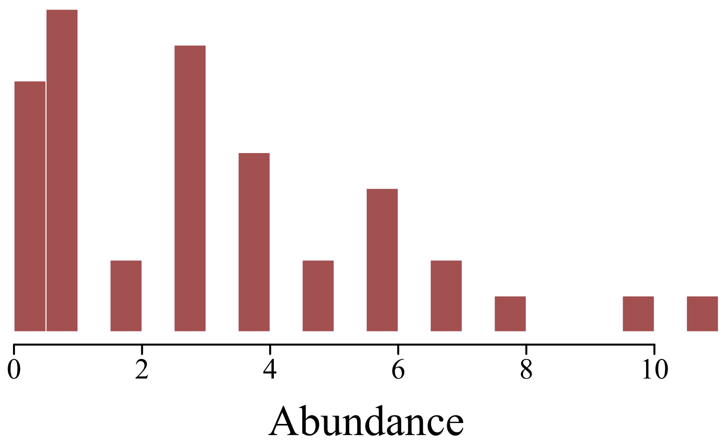
Phenology records

Tree rings

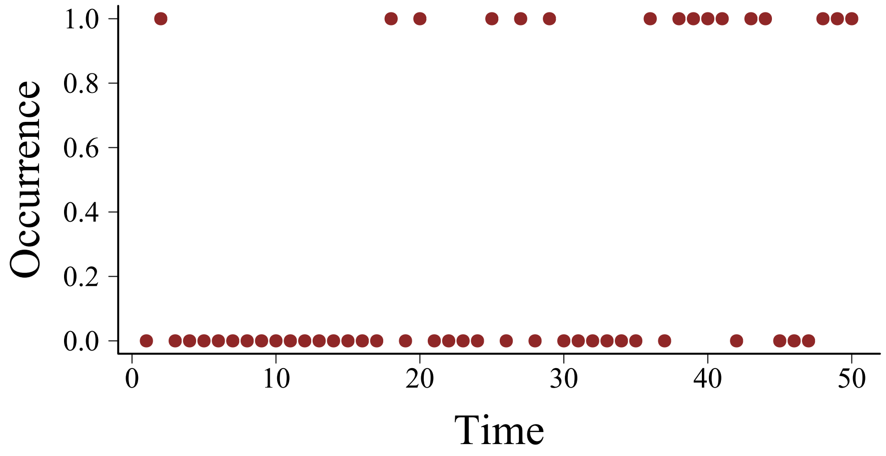
etc...

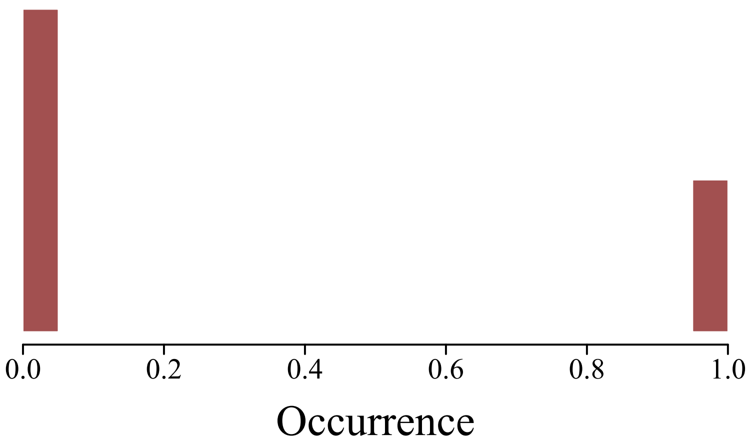
Example ecological time series



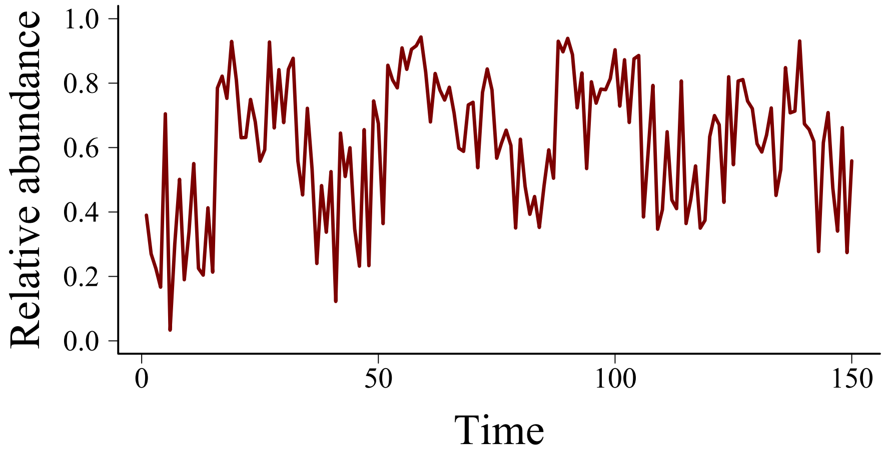


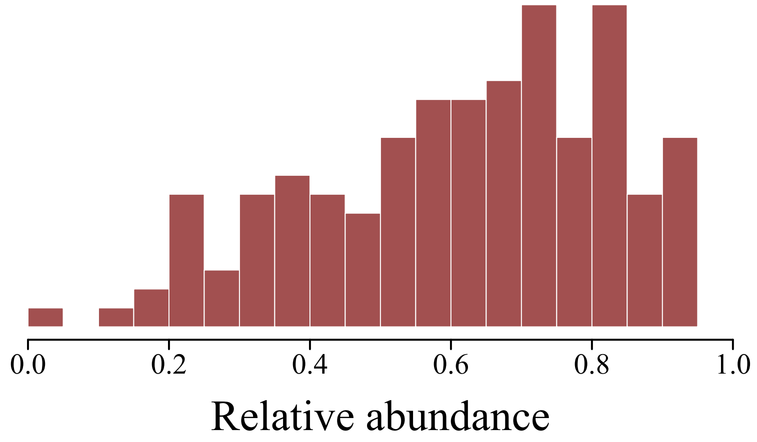
Another ecological time series



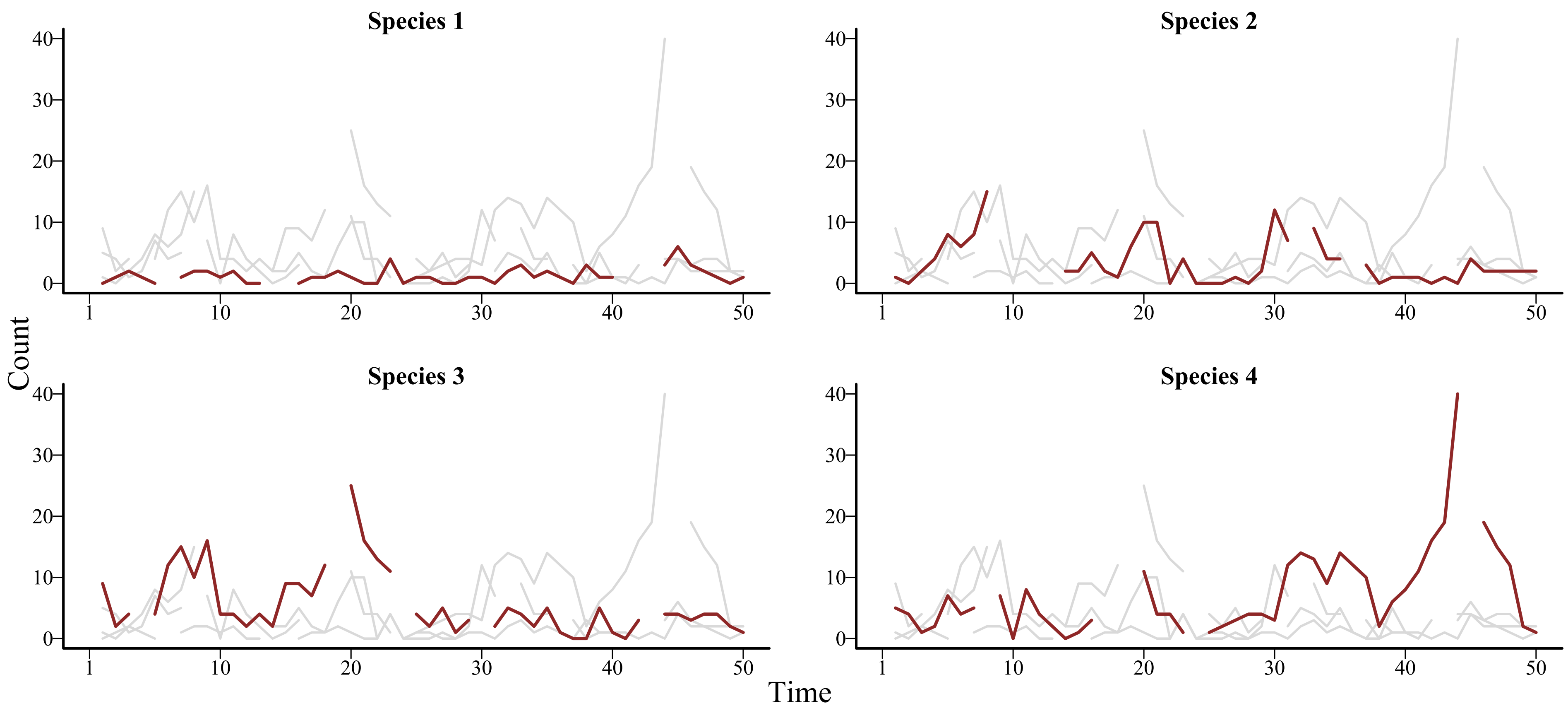


Yet another ecological time series

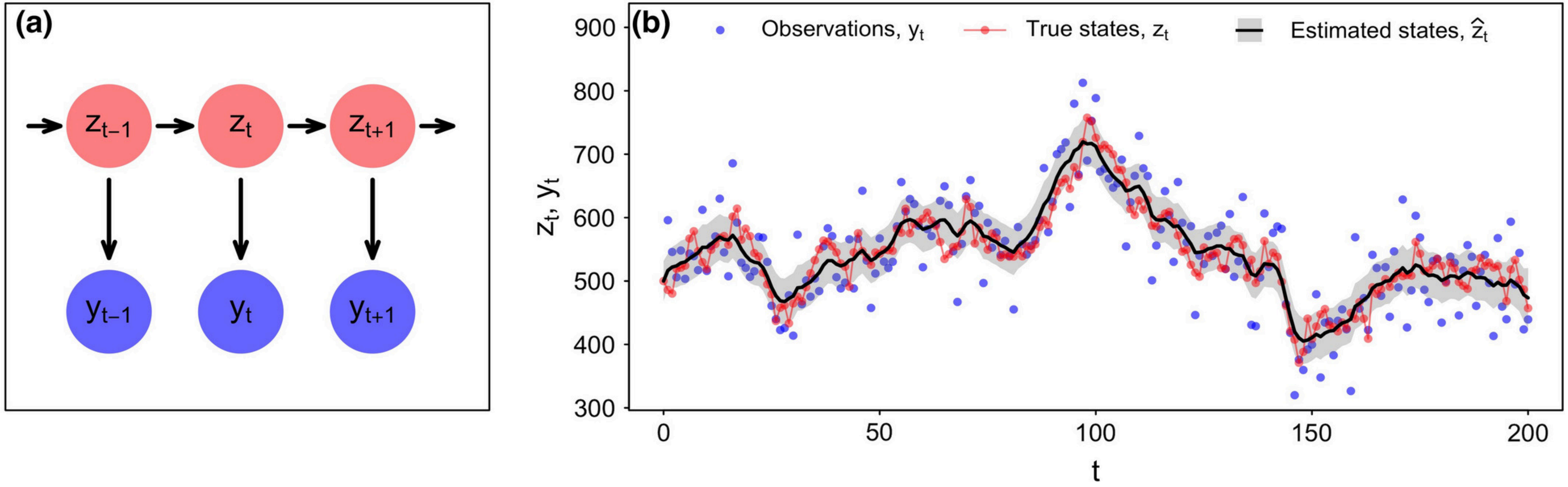




Collections of ecological series



All can have measurement error



[Auger-Methe *et al* 2021]

How can we do better?

In the next lecture, we will cover

Useful probability distributions for ecologists

Generalized Linear and Additive Models

Temporal random effects

Temporal residual correlation structures