Intro to Univariate State-Space Models

FISH 550 – Applied Time Series Analysis

Eli Holmes

11 April 2023

Week 3: State-Space Models

We are now starting a 5 lecture block on Gaussian state-space models.

Lectures 1 & 2: building blocks for analysis of multivariate time-series data with observation error, structure, and missing values

Lectures 3-5: Specific applications: covariates, dynamic factor analysis, dynamic linear models

---

- Properties of time series data
- AR and MA models: $x_t = b_1 x_{t-1} + b_2 x_{t-2} + e_t$
- Today: **State-space models (observation error and hidden random walks)**

Univariate linear state-space model

$$x_t = x_{t-1} + u + w_t, \quad w_t \sim N(0, q)$$

$$y_t = x_t + v_t, \quad v_t \sim N(0, r)$$

The x model is the classic "random walk" with drift.

y are the observatons.

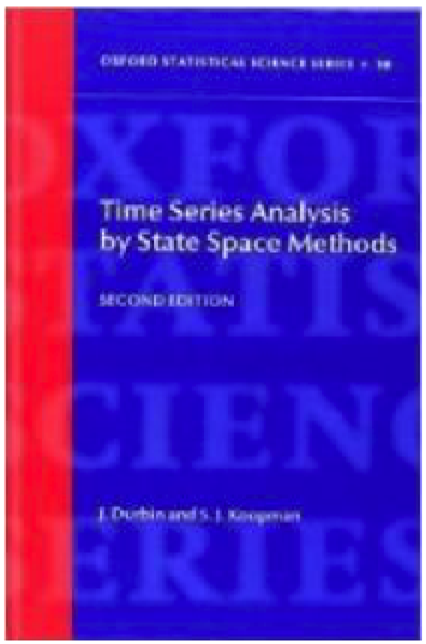
This model is a random walk observed with (Gaussian) error.

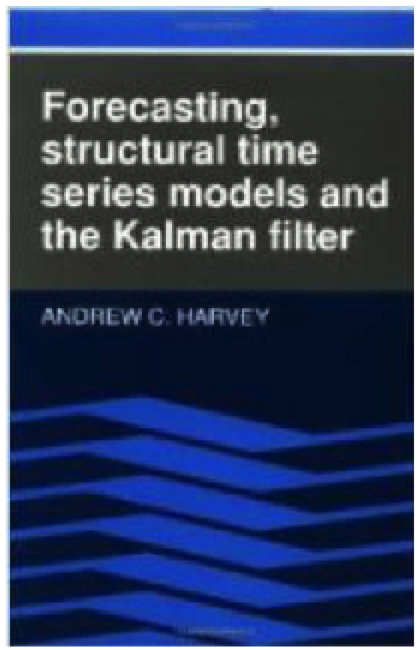
Univariate linear state-space model

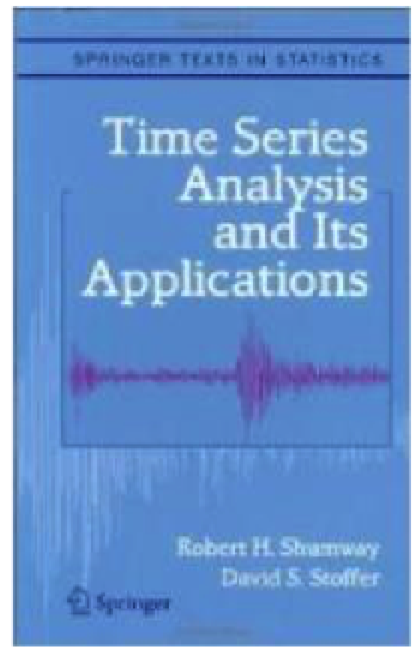
$$x_t = x_{t-1} + u + w_t, \quad w_t \sim N(0, q)$$

$$y_t = x_t + v_t, \quad v_t \sim N(0, r)$$

There are many textbooks on this class of model. It is used in extensively in economics and engineering.







AR(1) or AR lag-1

All of these are examples:

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

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

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

Why is the random walk with drift model so important in analysis of ecological data?

Additive random walks

$$x_t = x_{t-1} + u + w_t, \quad w_t \sim N(0, q)$$

- Movement, changes in gene frequency, somatic growth if growth is by fixed amounts
- Why Gaussian? The average of many small perturbations, regardless of their distribution, is Gaussian.

Multiplicative random walks

$$n_t = \lambda n_{t-1} e_t, \quad log(e_t) \sim N(0, q)$$

- Population growth, somatic growth if growth is by percentage
- Take the log and you get the linear additive model above. log-normal error distribution means that 10% increase is as likely as 10% decrease

Gompertz model

Addition of b with $0 < b < 1$ leads to process model with mean-reversion.

In the ecological literature on density-dependent processes, you may see this in non-log notation:

$$N_t = exp(u + w_t)N_{t-1}^b$$

N_t is population size.

Gompertz model

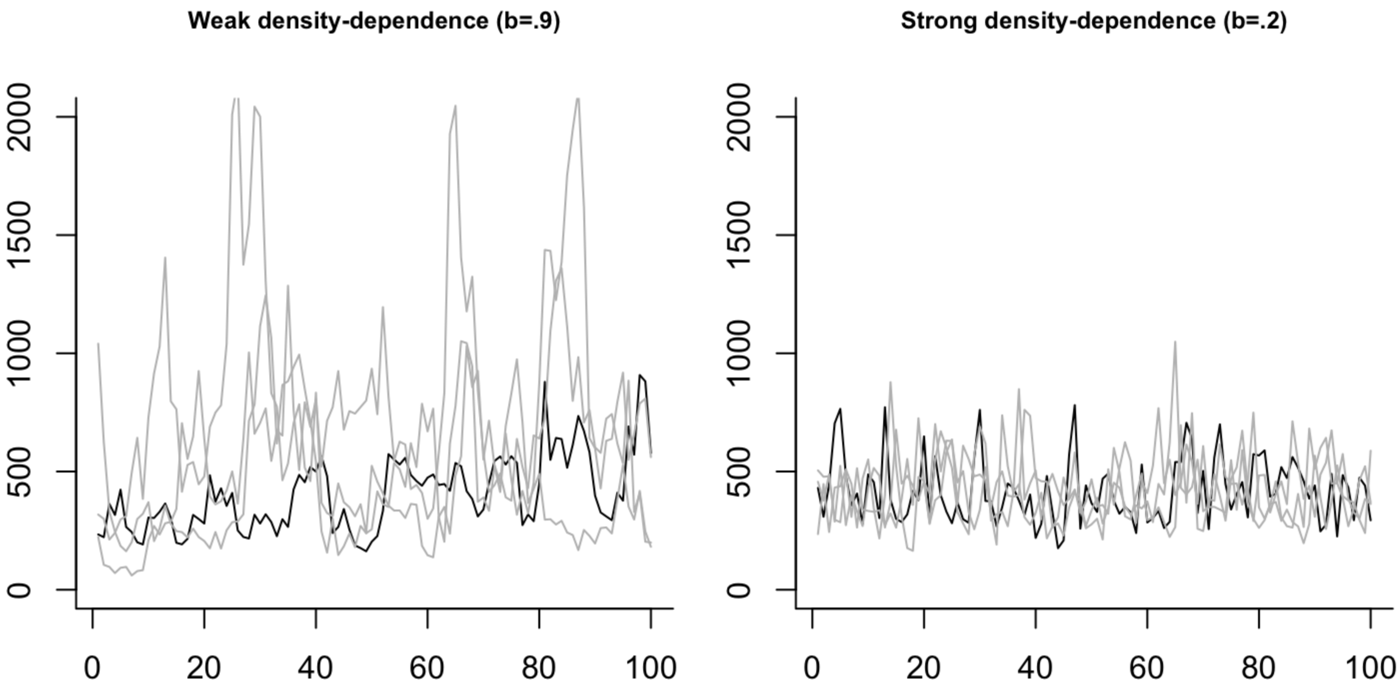
Take the log, and we have

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

$$w_t \sim N(0, q)$$

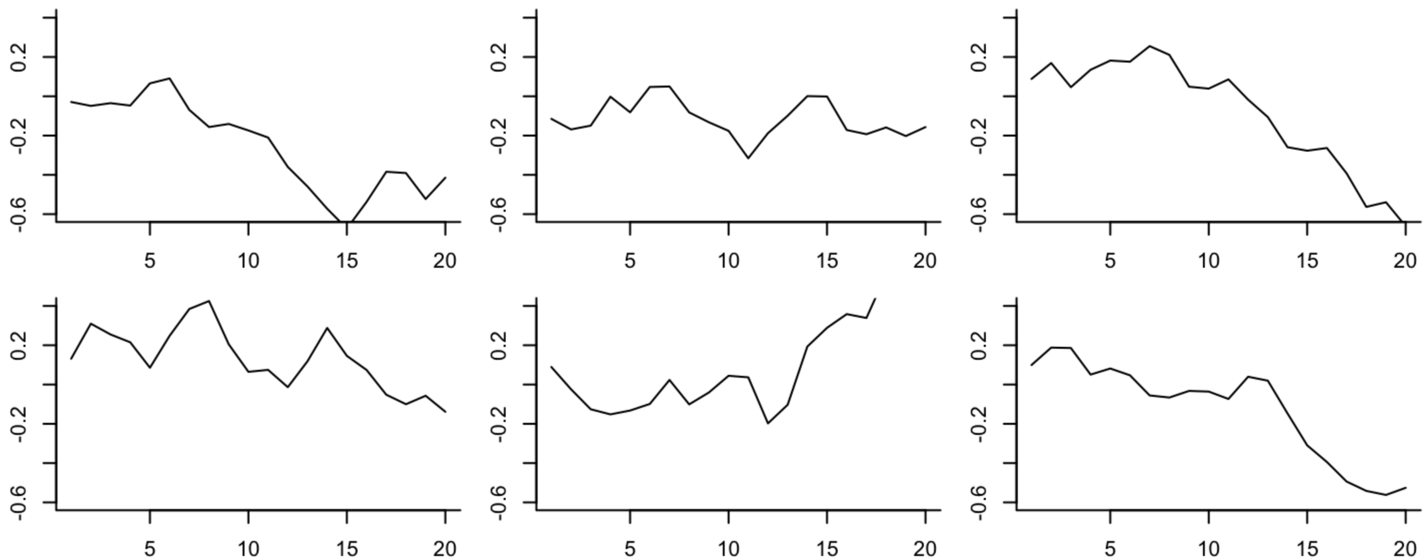
It is not required that w_t is Gaussian but that is a common assumption. Dynamics of processes with non-Gaussian errors, esp long-tailed errors, is a common extension. Autocorrelated errors could be implemented with MA process or covariates.

Gompertz model



Simple model, great flexibility

An random walk can show a wide-range of trajectories, even for the same parameter values. All trajectories below came from the same random walk model: $x_t = x_{t-1} - 0.02 + w_t, w_t \sim N(mean = 0.0, var = 0.01)$.



Definition: state-space

The "state" is a hidden (dynamical) variable. In this class, it will be a hidden random walk or AR(1) process.

Our data are observations of this hidden state.

Often state-space models include inputs (explanatory variables) and the state or the data may be multivariate.

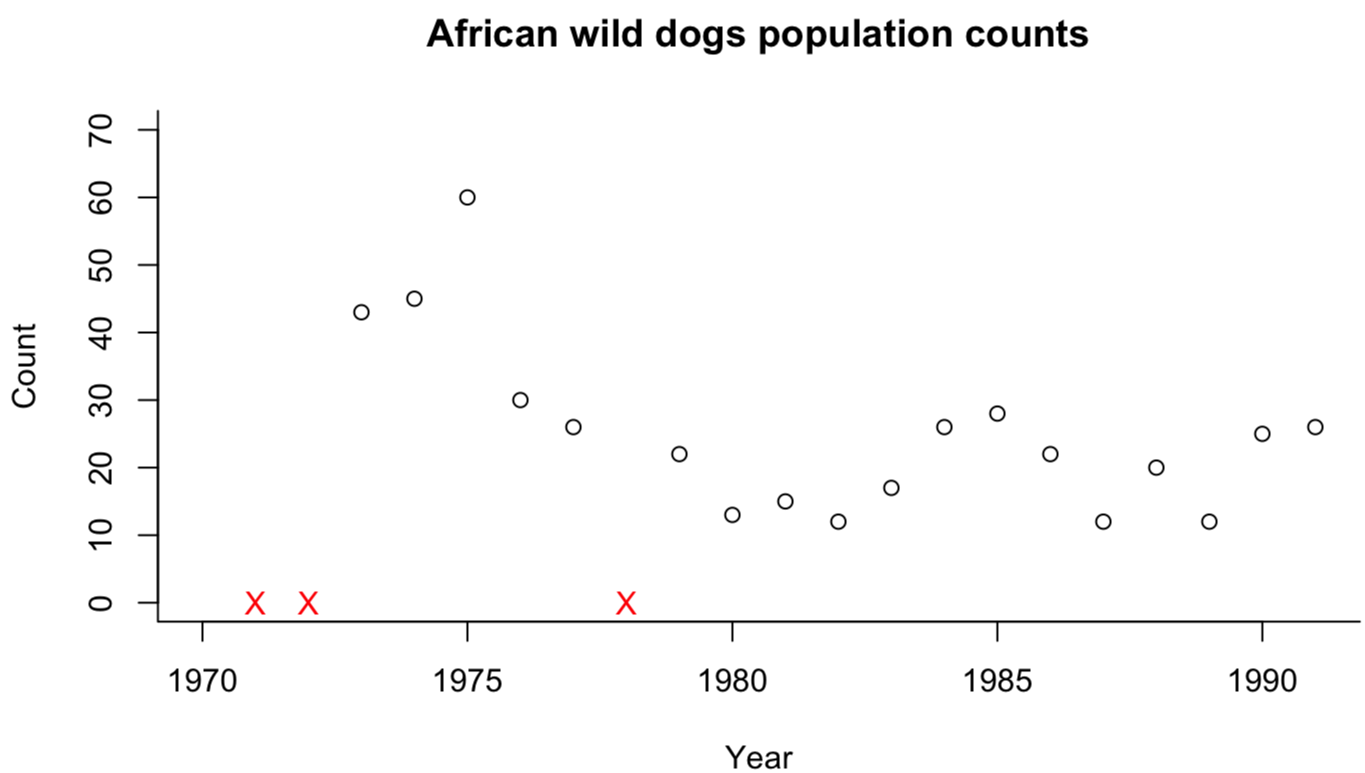
The model you are seeing today is a simple univariate state-space model with no inputs.

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

observation: $y_t = x_t + v_t$

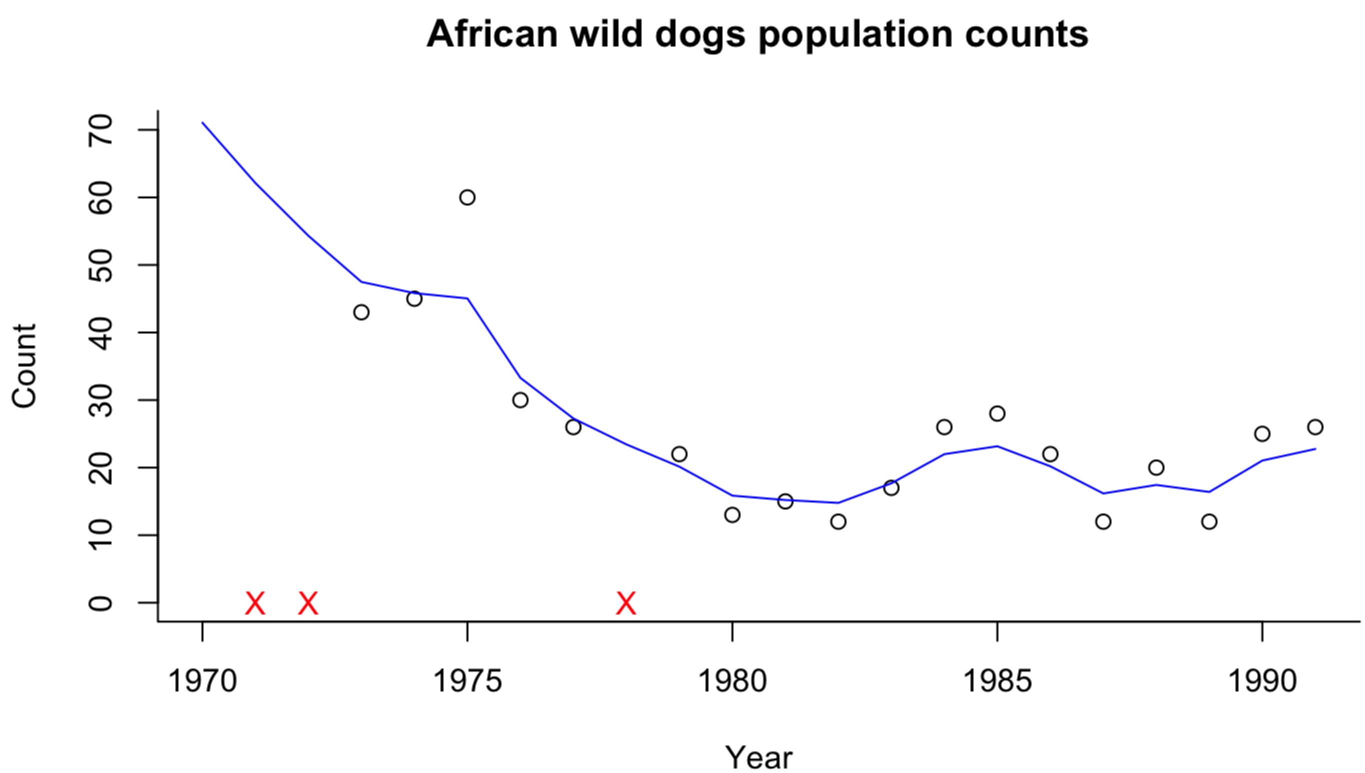
Example: population count data

Yearly, usually, population or subpopulation counts, possibly with missing values.



Example: population count data

The data are observations of a hidden ‘true’ population size. The data are observations of that hidden state and have observation error.



Observation error

This is a survey photograph for Steller sea lions in the Gulf of Alaska. There IS some number of sea lions in our population in year t, but we don’t know that number precisely. It is “hidden”.

The observation error variance is often unknowable in fisheries and ecological analyses

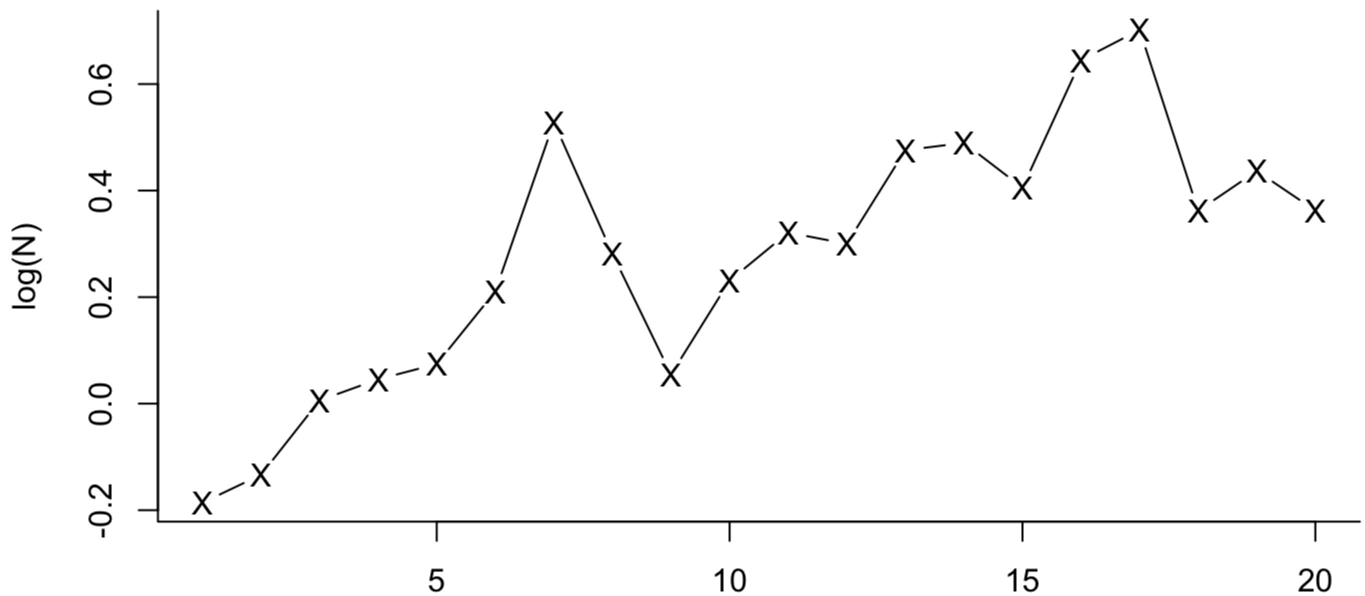
Sightability varies due to factors that may not be fully understood or measureable

- Environmental factors (tides, temperature, etc.)
- Population factors (age structure, sex ratio, etc.)
- Species interactions (prey distribution, prey density, predator distribution or density, etc.)

Sampling variability–due to how you actually count animals–is just one component of observation variance

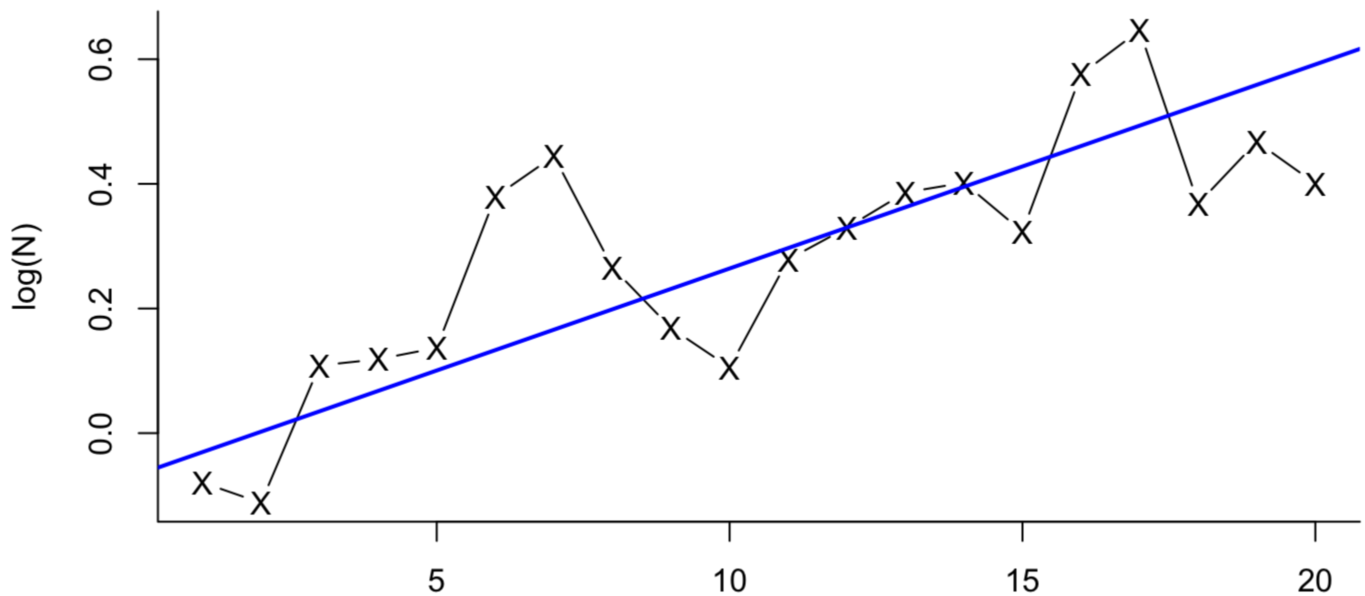
Process versus observaton variability

Suppose we have the following data (say, population density logged)



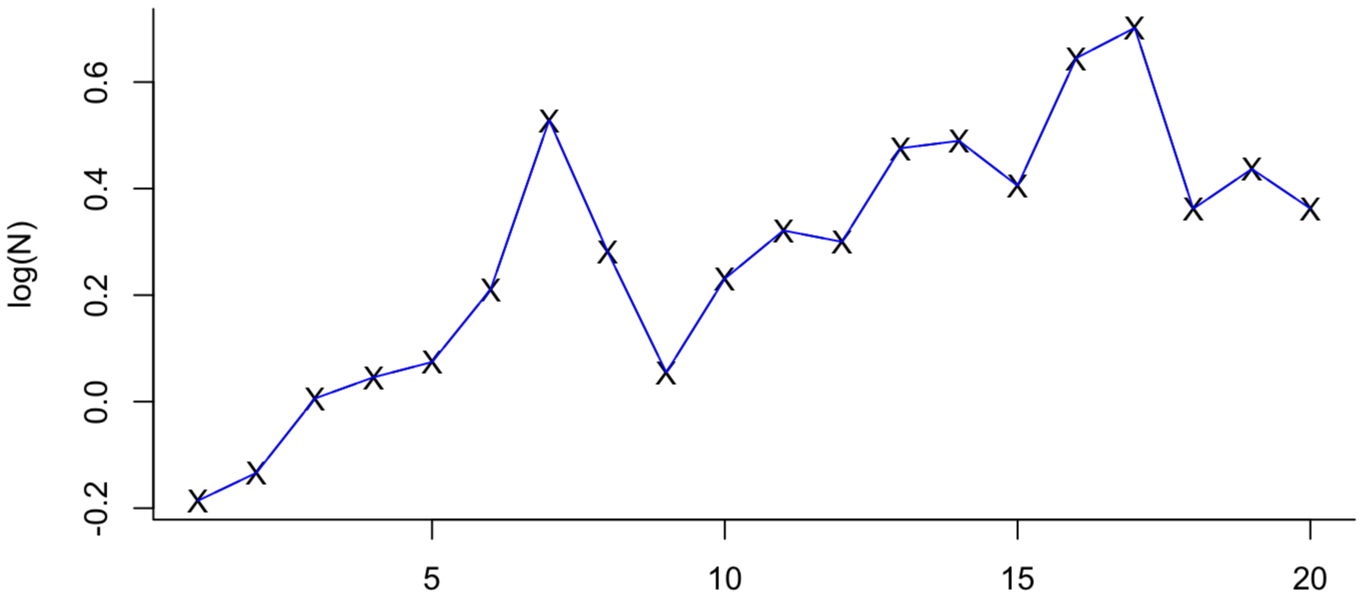
Fit a linear regression

The model of the hidden state in this case is $x_t = \alpha + \beta t$. The observation model is $y_t = x_t + v_t$. All variability = **non-process** or observation variability.



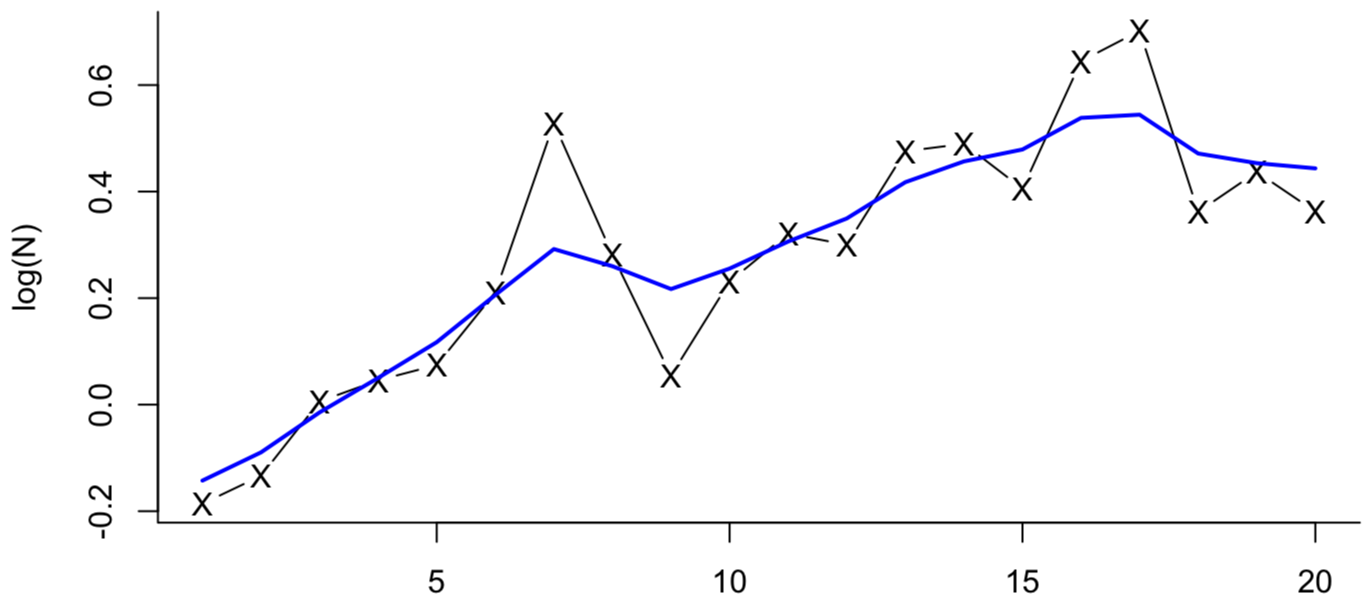
Fit a random walk model

The model of the hidden state in this case is $x_t = \alpha + x_{t-1} + w_t$. The observation model is $y_t = x_t$. All variability = **process** variability.



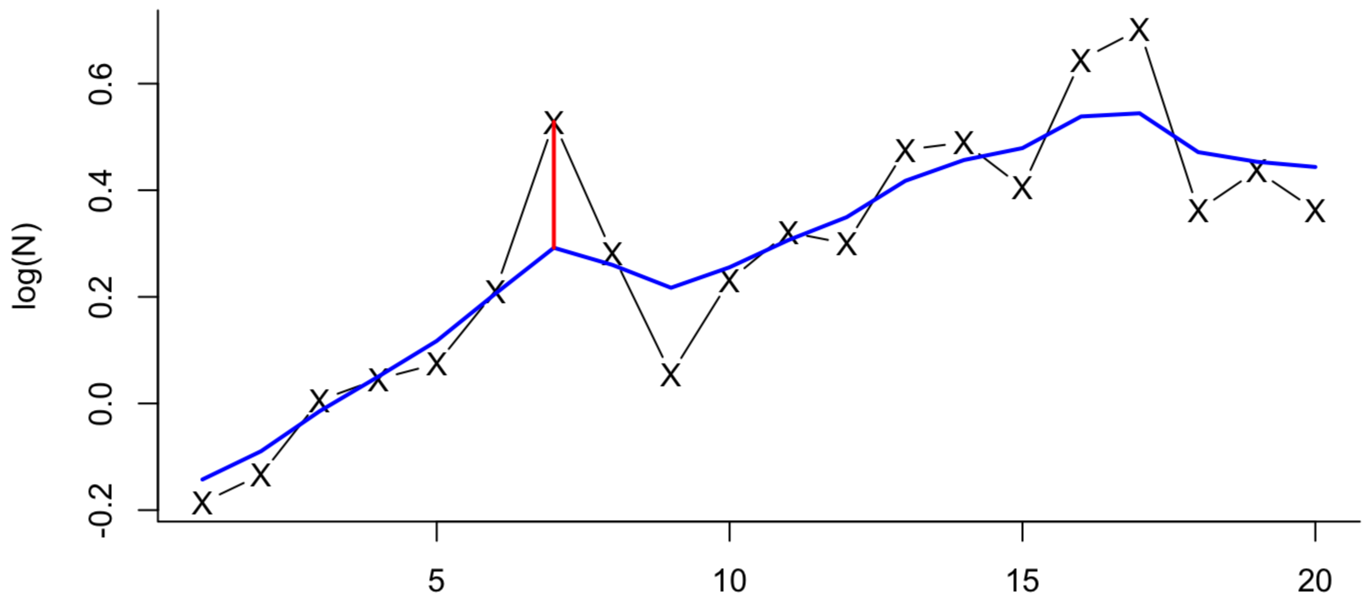
Fit a state-space model

Autoregressive state-space models fit a random walk AR(1) through the data. The variabilty in the data contains both process and non-process (observation) variability.



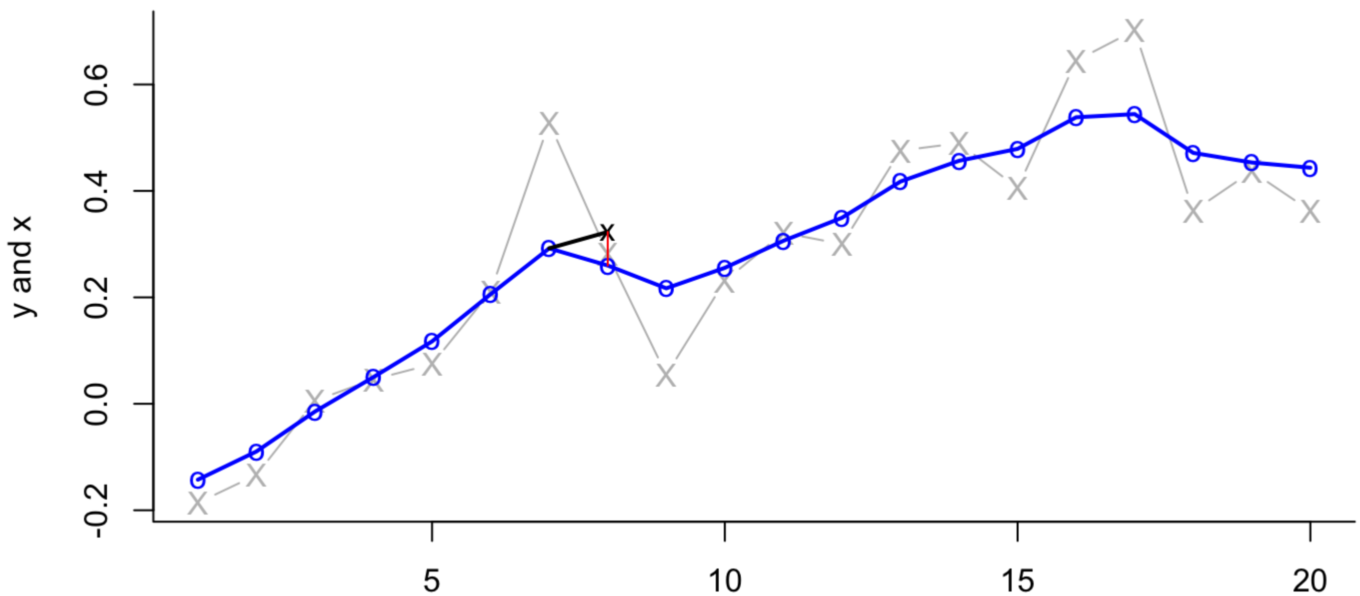
Non-process variability

Observation or “non-process” error is the difference between the hidden state (blue line) and the observation (X).



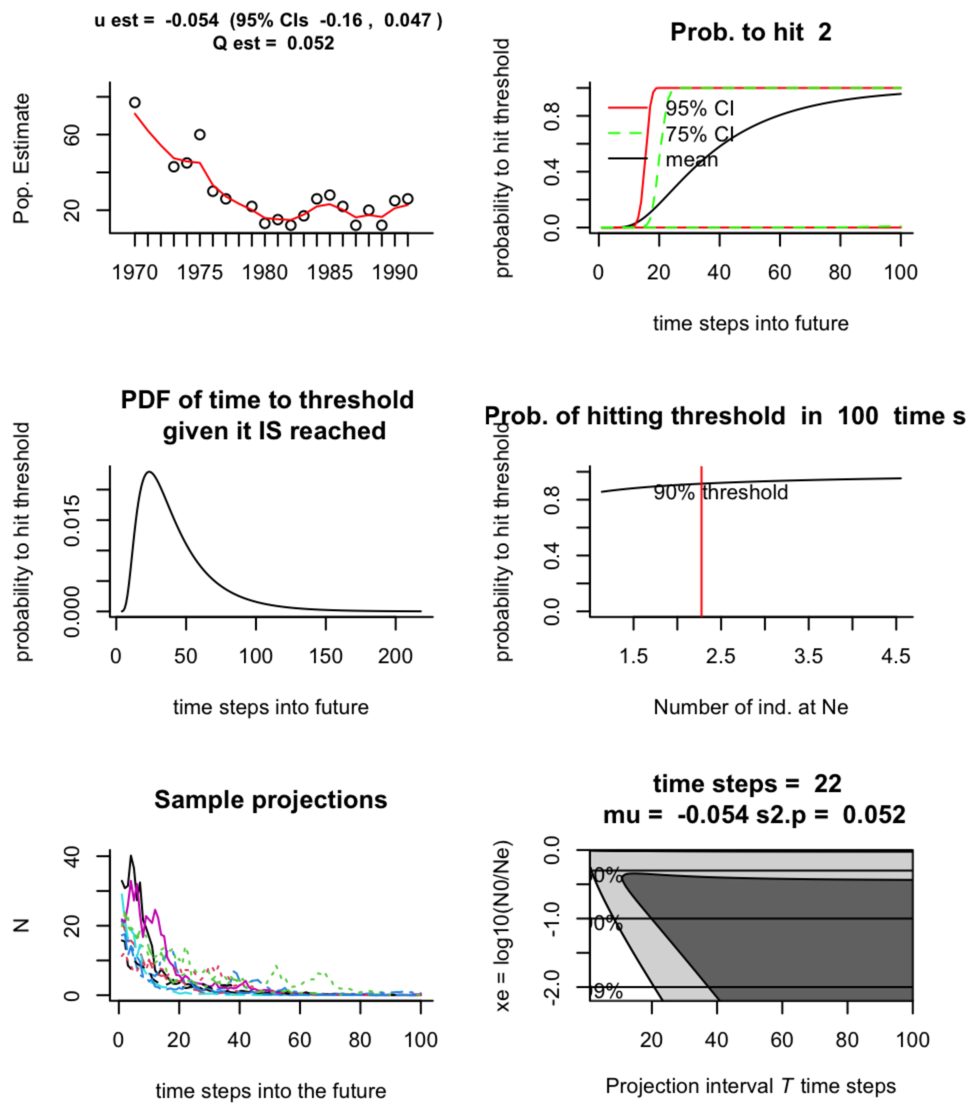
Process variability

Process error is the difference between the expected x_t given data up to time $t-1$ (x in the plot) and the actual x at time t.

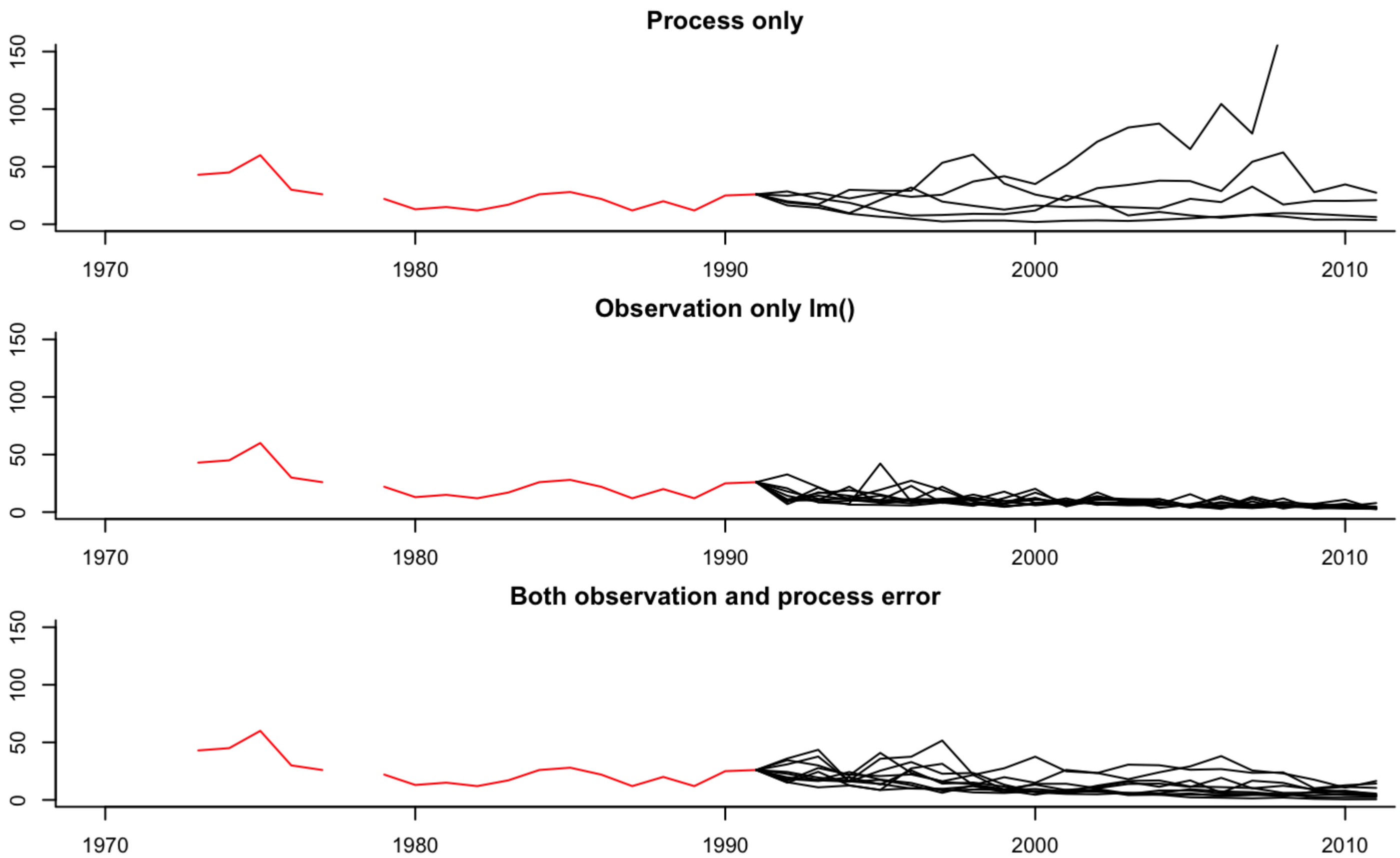


PVA example

One use of univariate state-space models is “count-based” population viability analysis (chap 7 HWS2014)

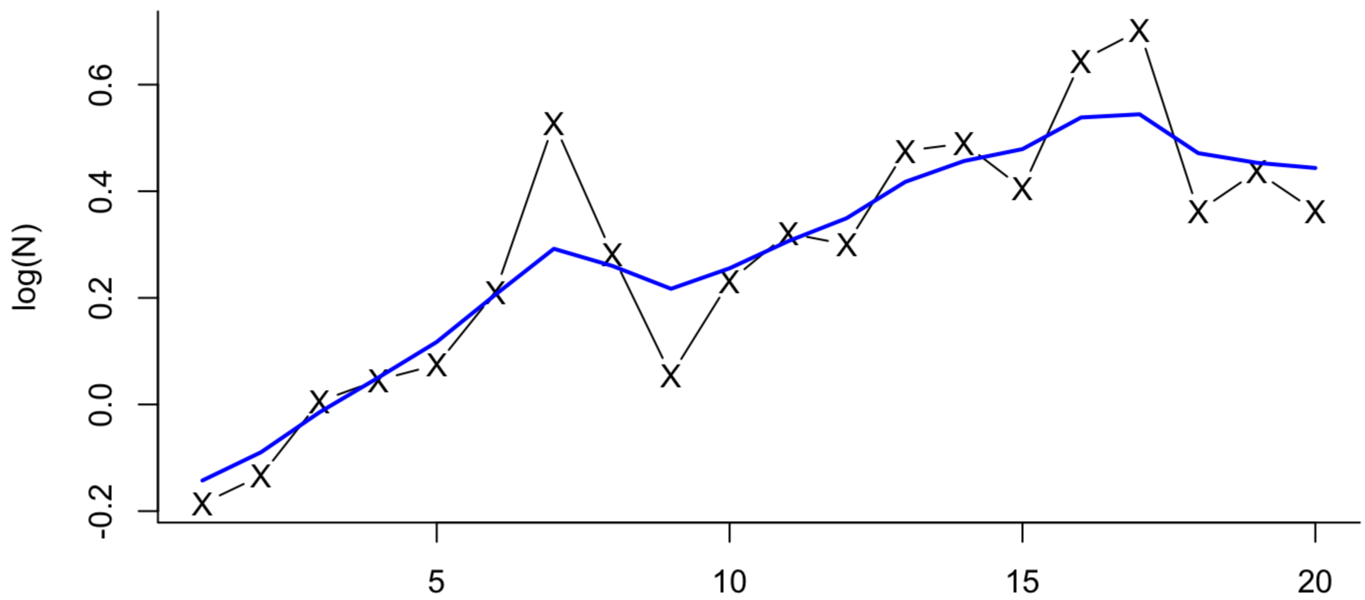

u est = -0.054 (95% CIs -0.16 , 0.047)
Q est = 0.052
Pop. Estimate
1970 1975 1980 1985 1990
Prob. to hit 2
probability to hit threshold
95% CI
75% CI
mean
time steps into future
PDF of time to threshold
given it IS reached
Prob. of hitting threshold in 100 time s
90% threshold
Number of ind. at Ne
Sample projections
N
time steps into the future
time steps = 22
mu = -0.054 s2.p = 0.052
xe = log10(N0/Ne)
Projection interval T time steps

How you model your data has a large impact on your forecasts

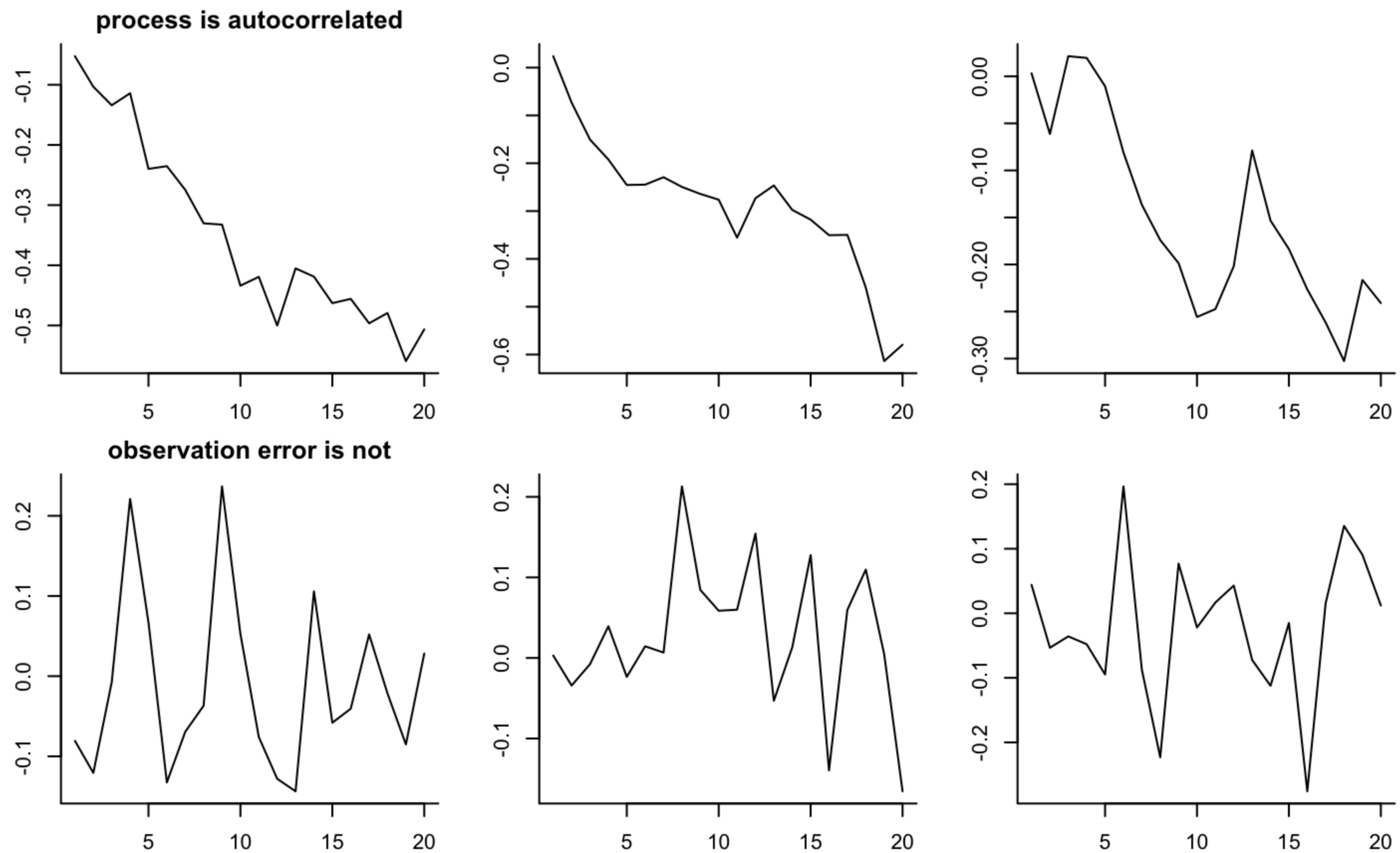


How can we separate process and non-process variance?

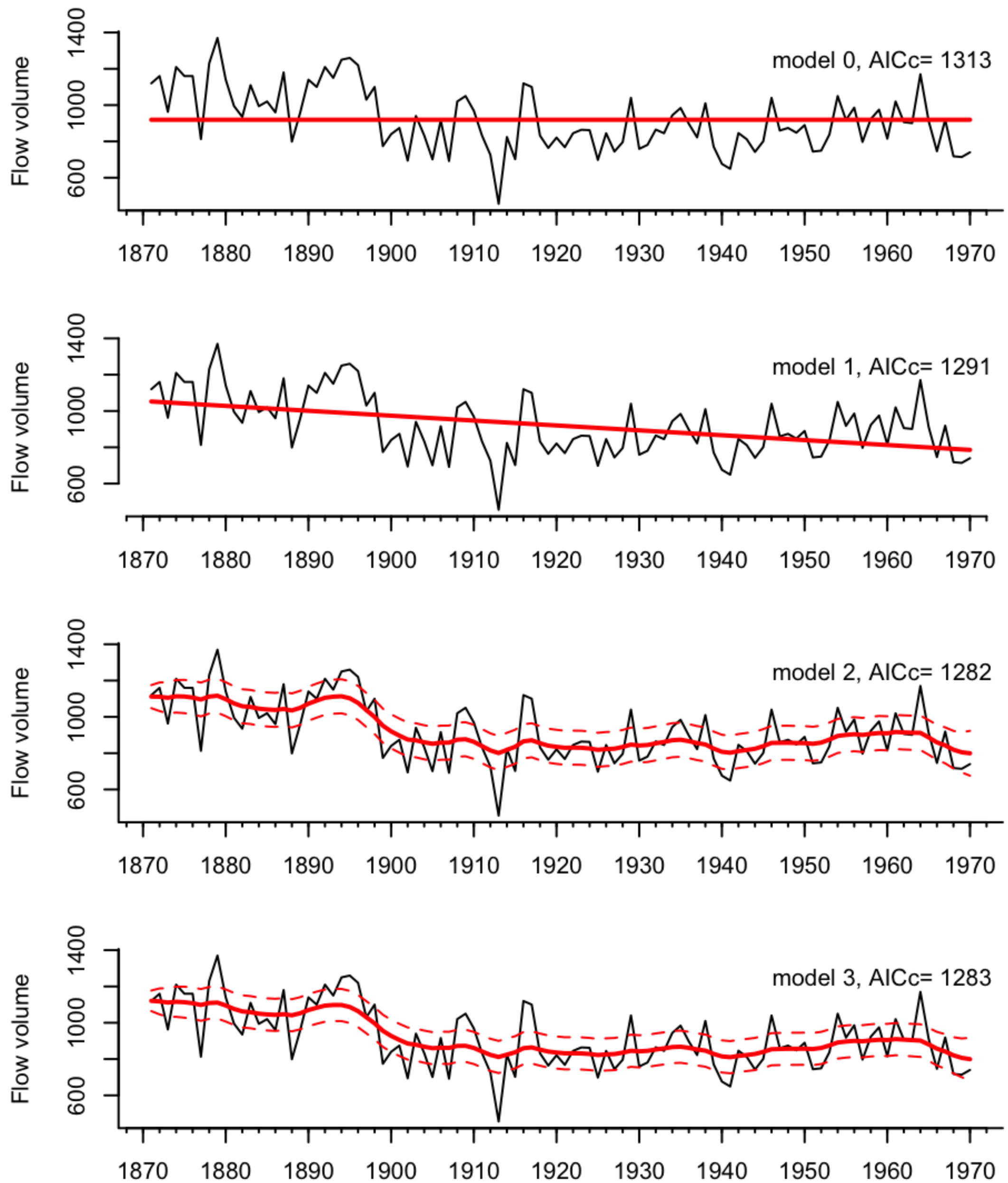
Wouldn't these two variances be impossible to separate?



They have different temporal patterns.



Nile River example



Kalman filter and smoother

The Kalman filter and smoother is an algorithm for computing the expected value of the x_t from the data and the model parameters.

$$x_t = x_{t-1} + u + w_t, \quad w_t \sim N(0, q)$$

$$y_t = x_t + v_t, \quad v_t \sim N(0, r)$$

Diagnostics

Innovations residuals aka, one-step ahead residuals, same ones we used for ARMA models

data at time t minus model predictions given data up to $t-1$

$$\hat{y}_t = E[Y_t|y_{t-1}]$$

In the MARSS package, the one-step ahead residuals are returned by

```
residuals(fit)
```

This is fairly standard for models that fit state-space models.

Standard diagnostics

- ACF
- Normality

`MARSS` package

We will be using the MARSS package to fit univariate and multivariate state-space models.

The main function is `MARSS()`:

```
fit <- MARSS(data, model=list())
```

data are a vector or a matrix with time going along the columns.

model list is a list with the structure of all the parameters.

MARSS model notation

$$x_t = \mathbf{B}x_{t-1} + \mathbf{U} + w_t, \quad w_t \sim N(0, \mathbf{Q})$$

$$y_t = \mathbf{Z}x_t + \mathbf{A} + v_t, \quad v_t \sim N(0, \mathbf{R})$$

The MARSS model list follows this notation one-to-one.

$$x_t = x_{t-1} + u + w_t, \quad w_t \sim N(0, q)$$

$$y_t = x_t + v_t, \quad v_t \sim N(0, r)$$

Write as where everything bold is a matrix.

$$x_t = \mathbf{B} x_{t-1} + \mathbf{U} + w_t, \quad w_t \sim N(0, \mathbf{Q})$$

$$y_t = \mathbf{Z} x_t + \mathbf{A} + v_t, \quad v_t \sim N(0, \mathbf{R})$$

```
mod.list <- list(
  U = matrix("u"),
  x0 = matrix("x0"),
  B = matrix(1),
  Q = matrix("q"),
  Z = matrix(1),
  A = matrix(0),
  R = matrix("r"),
  tinitx = 0
)
```

Diagnostics and plotting

Use

```
autoplot(fit)
```

where `fit` is returned by `MARSS()` to see the standard diagnostics.

Output

```
fit <- MARSS()
```

- `coef(fit)` to get the estimated parameters
- `tidy(fit)` to get estimated parameters with CIs
- `tsSmooth()` to get the estimates states or use `fit$states`
- `fitted()` to get the model estimates of mean y
- `fr <- forecast(fit, h=5, interval="prediction")` predictions
- `autoplot(fr)` plot the forecast

Let’s see some examples

We will go through these in class

- example 1
- example 2
- example 3
- example 4