Basics of time-series / dependent data analysis

Eli Gurarie

StatR 101 - Lecture 12b
December 11, 2012

December 11, 2012

Time Series Inferno:

"Lasciate ogni speranza (di indipendenza), voi ch'entrate!"[1]

[1] Translation: "Abandon all hope (of independence), ye who enter here!"

Linear models + 1/2 slide on generalized linear models

1. Simple:

$$\begin{aligned} Y_i &= \mathbf{X}\beta + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where $\mathbf{X}$ is the *linear predictor* (or "design matrix"), β are the parameters, and ϵ_i are i.i.d.

2. Generalized:

$$\begin{aligned} Y_i &\sim \text{Distribution}(\mu) \\ g(\mu) &= \mathbf{X}\beta \end{aligned}$$

where "Distribution" is exponential family distribution (Binomial, Poisson, Negative Binomial, etc.) and $g(\mu)$is the *link function*

Linear models + 1/2 slide on generalized linear models

1. Simple:

$$\begin{aligned} Y_i &= \mathbf{X}\beta + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where $\mathbf{X}$ is the *linear predictor* (or "design matrix"), β are the parameters, and ϵ_i are i.i.d.

2. Generalized:

$$\begin{aligned} Y_i &\sim \text{Distribution}(\mu) \\ g(\mu) &= \mathbf{X}\beta \end{aligned}$$

where "Distribution" is exponential family distribution (Binomial, Poisson, Negative Binomial, etc.) and $g(\mu)$is the *link function*

Linear models + 1/2 slide on generalized linear models

1. Simple:

$$\begin{aligned} Y_i &= \mathbf{X}\beta + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where $\mathbf{X}$ is the *linear predictor* (or "design matrix"), β are the parameters, and ϵ_i are i.i.d.

2. Generalized:

$$\begin{aligned} Y_i &\sim \text{Distribution}(\mu) \\ g(\mu) &= \mathbf{X}\beta \end{aligned}$$

where "Distribution" is exponential family distribution (Binomial, Poisson, Negative Binomial, etc.) and $g(\mu)$is the *link function*

Note: In both cases i is *unordered*!

- $Y_1, Y_2, Y_3 \ldots Y_n$ can be reshuffled: $Y_5, Y_{42}, Y_2, Y_n \ldots Y_3$
- Y_i and Y_j are independent: $\text{Cov}(Y_i, Y_j) = 0$

Basic Definitions

1. **Time-series:**
   - Any data (or realization of random process) that are indexed by *time*
   - $i \rightarrow t$: $Y_1, Y_2, Y_3 \ldots Y_n$ becomes $Y_{t_1}, Y_{t_2}, Y_{t_3} \ldots Y_{t_n}$

2. **Discrete-time time-series**:
   - Data collected at regular (Annual, Monthly, Daily, Hourly, etc.) intervals.
   - t usually denoted (and ordered) $1, 2, 3, \ldots n$.

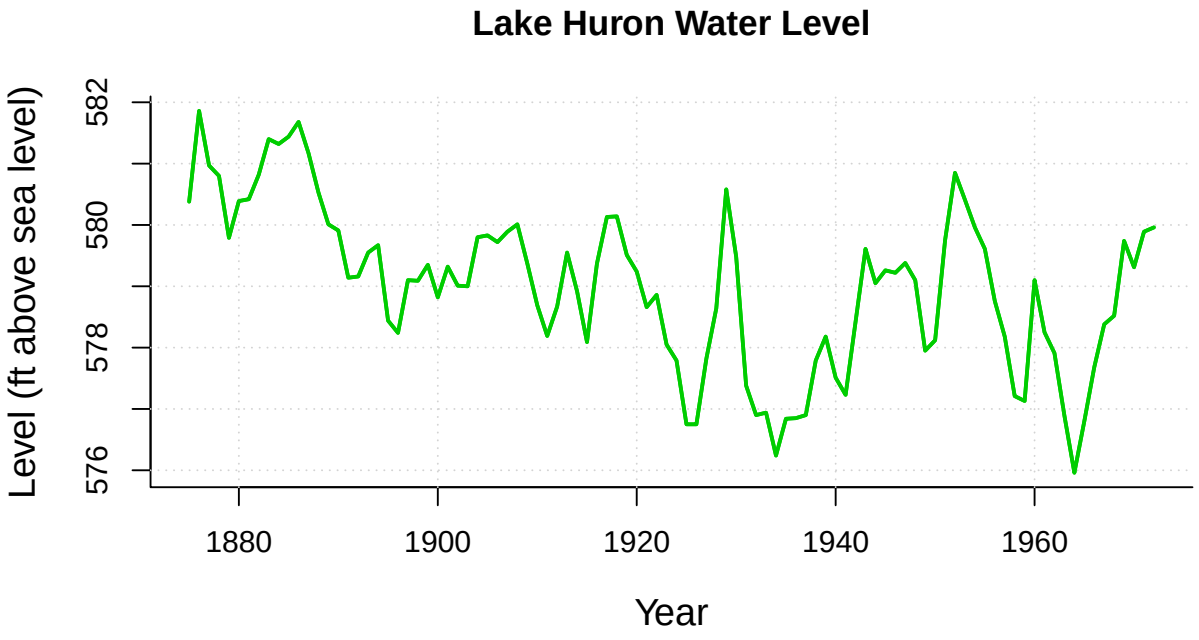
3. **Continuous-time time-series**:
   - Data collected at arbitrary intervals: $t_i \in \{T_{min}, T_{max}\}$.

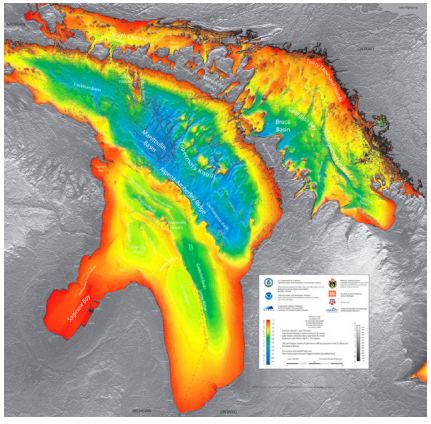
4. **Stochastic process:**
   - A *model* of a random process in time
   - $X_{t_n} = f(X_{t_1}, X_{t_2}, X_{t_3} \ldots X_{t_{n-1}})$, where $f(\cdot)$ is a random process

Some objectives of time-series analysis

- Characterizing a random process
  - Identifying trends, cycles, random structure
- *Identifying* and *estimating* the stochastic model for the time series
- Forecasting
- Inference
  - Accounting for lack of independence in linear modeling frameworks

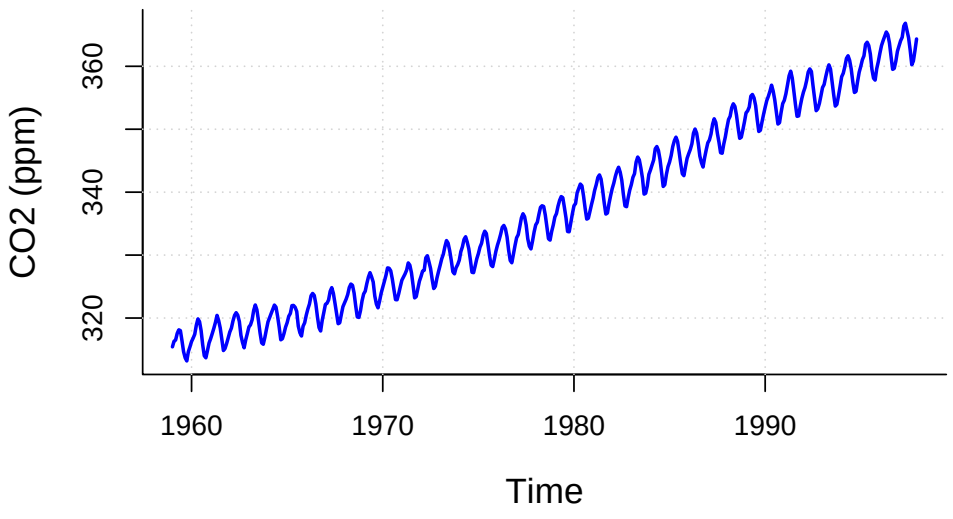
Example of time series (and questions)





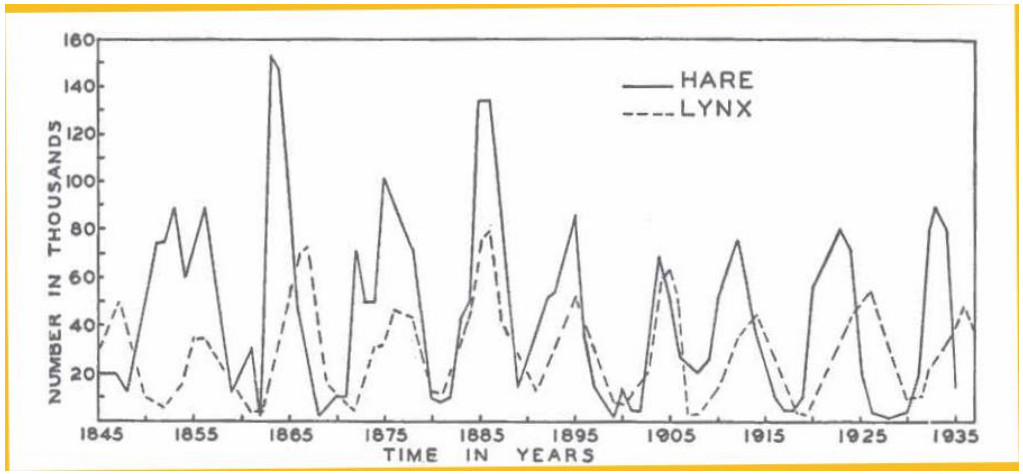
Question: Has the level of Lake Huron changed over 100 years?

Example of time series (and questions)



Question: Can we identify the trend / seasonal pattern / random fluctuations / make forecasts?

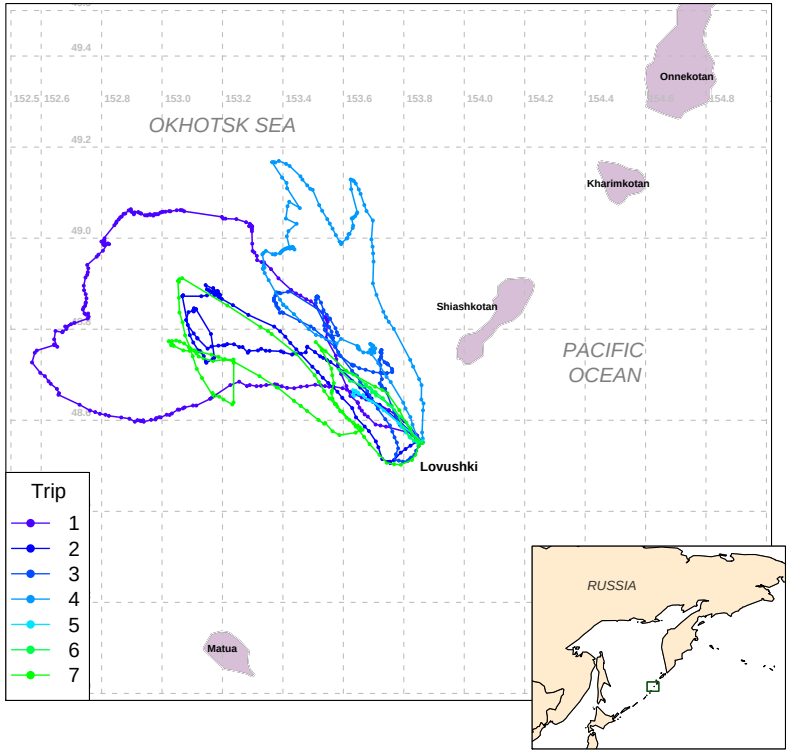
Example of time series (and questions)



Question: What are the dynamics and relationships within and between lynx and hare numbers?[2]

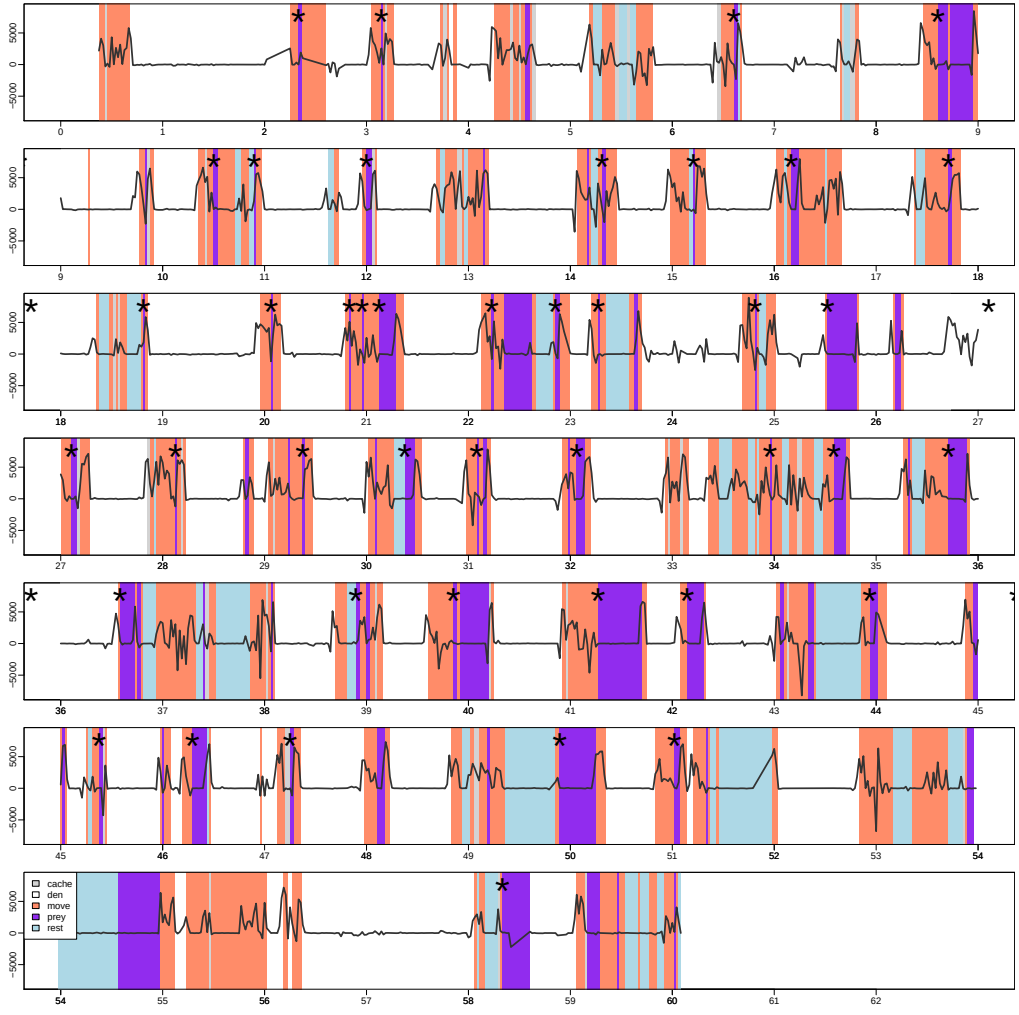
[2] MacLulich's *Fluctuations in the numbers of varying hare*, 1937

Example of time series (and questions)



Question: How can we analyze the movements of animals?
Note: Multidimensional state X, continuous time T

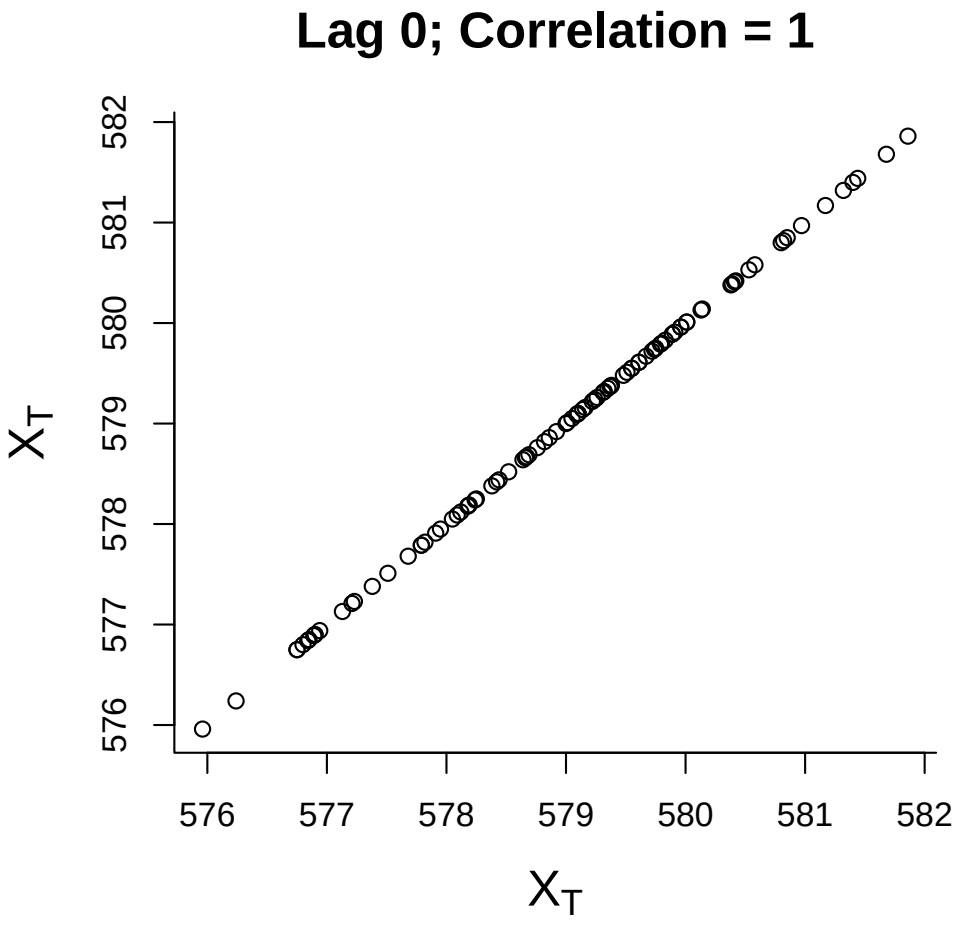
Example of time series (and questions)

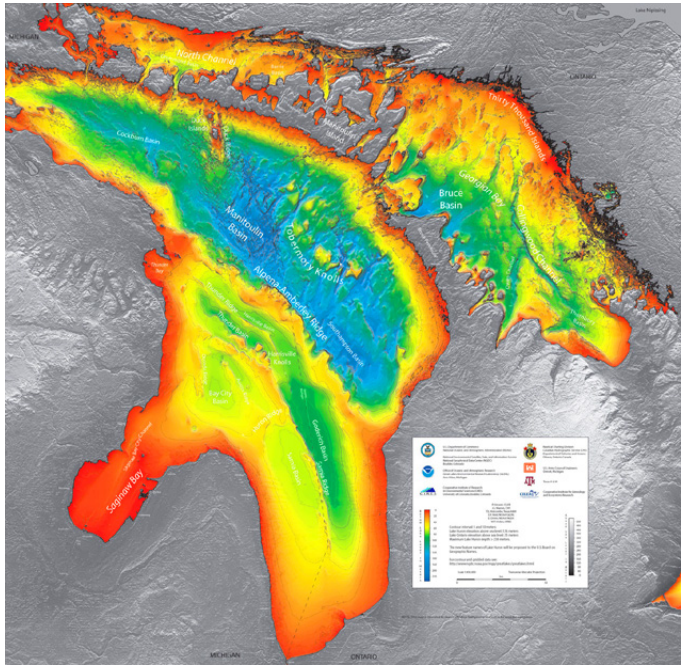


Question: How can we quantify the time budgets and behaviors of a wolf?

Note: Discrete states X, continuous time T

Concept 1: Autocorrelation

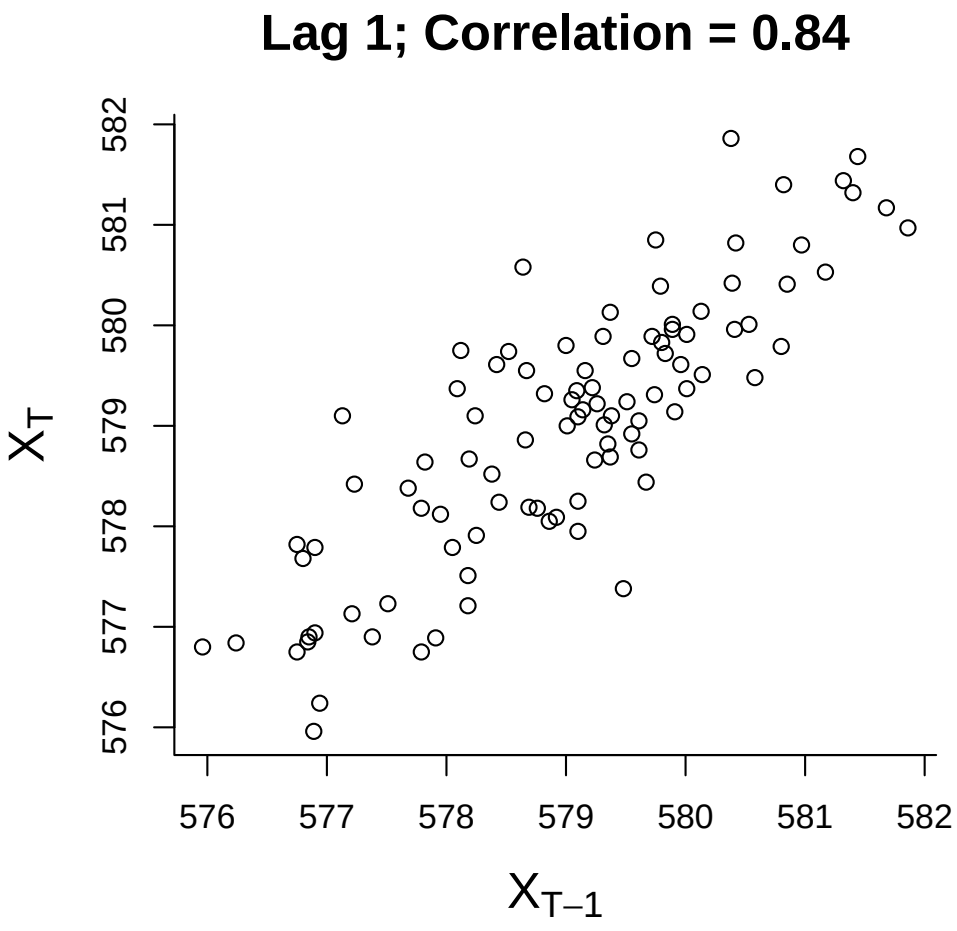


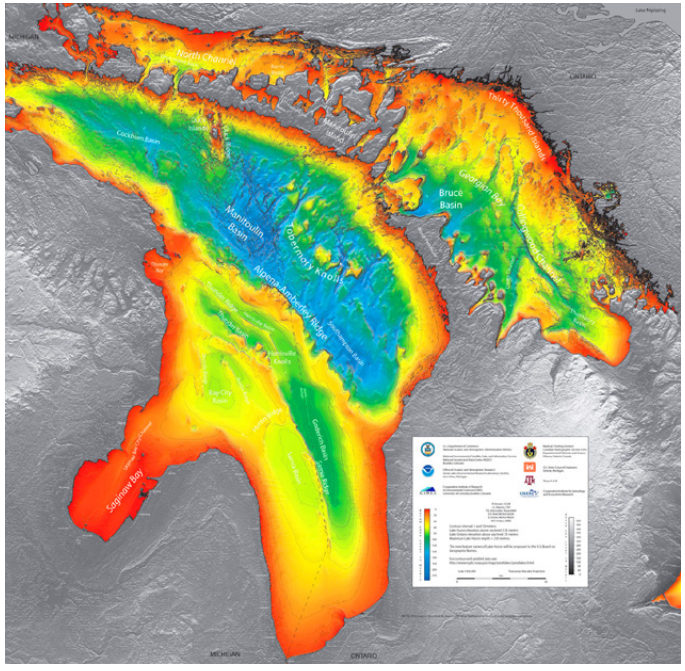


Autocovariance function: $\gamma(h) = \mathrm{Cov}\,(X_t, X_{t-h})$

Autocorrelation function: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{Cor}\,(X_t, X_{t+h})$

Concept 1: Autocorrelation

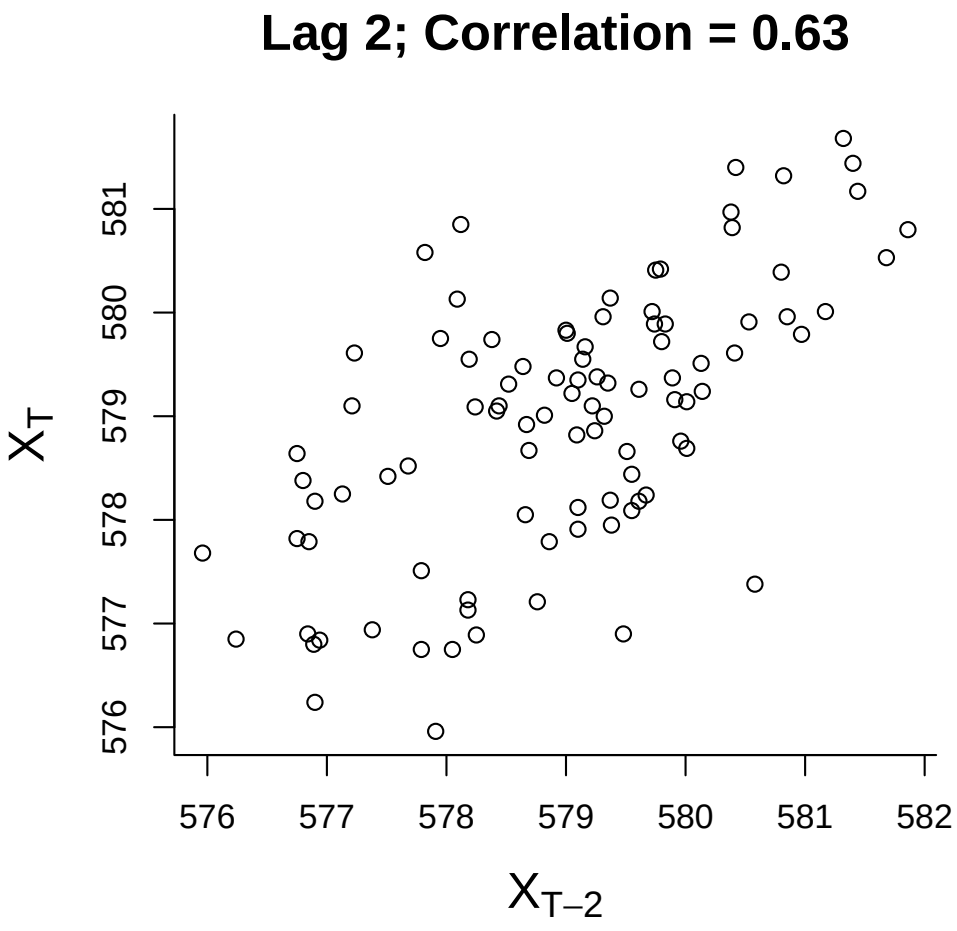


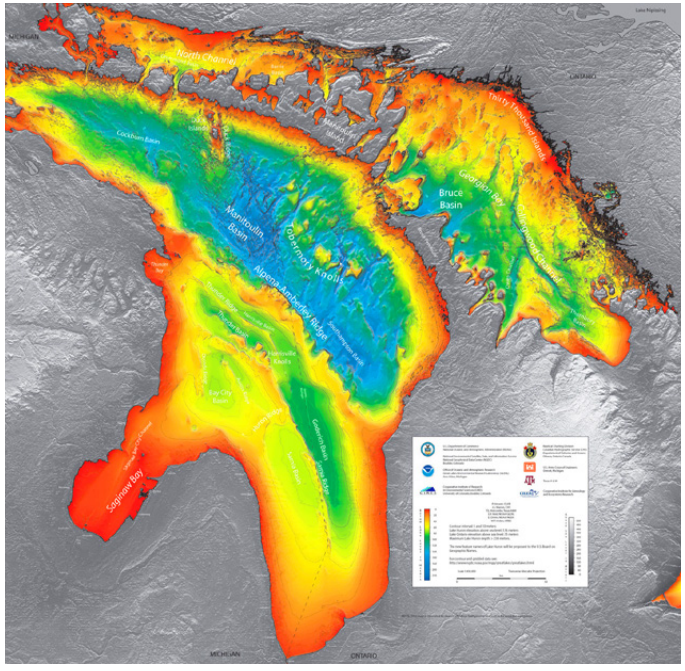


Autocovariance function: $\gamma(h) = \mathrm{Cov}\,(X_t, X_{t-h})$

Autocorrelation function: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{Cor}\,(X_t, X_{t+h})$

Concept 1: Autocorrelation

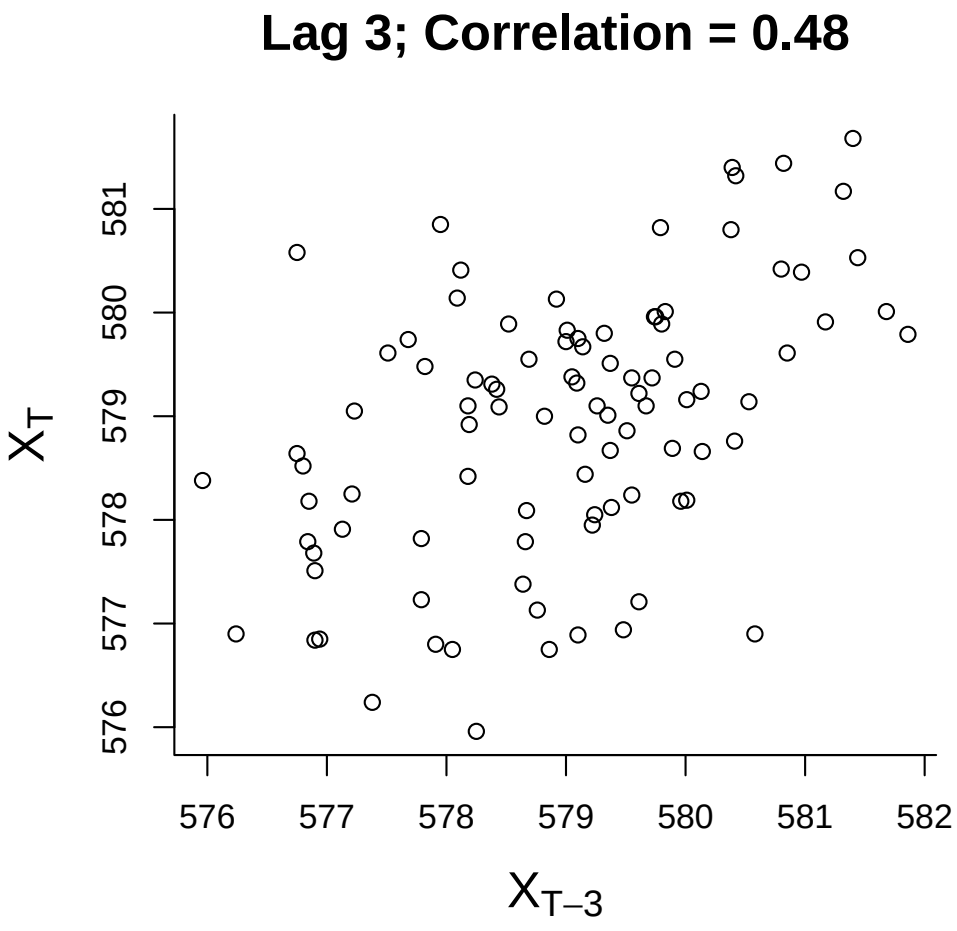


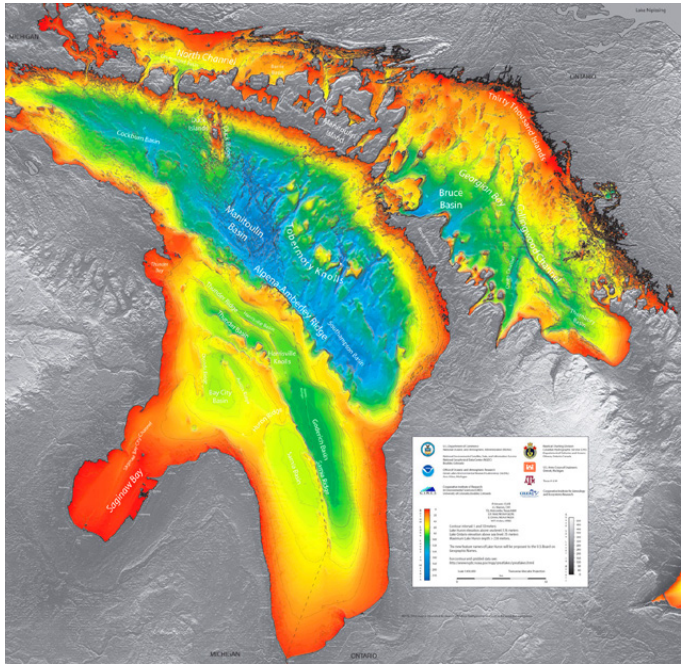


Autocovariance function: $\gamma(h) = \mathrm{Cov}\,(X_t, X_{t-h})$

Autocorrelation function: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{Cor}\,(X_t, X_{t+h})$

Concept 1: Autocorrelation

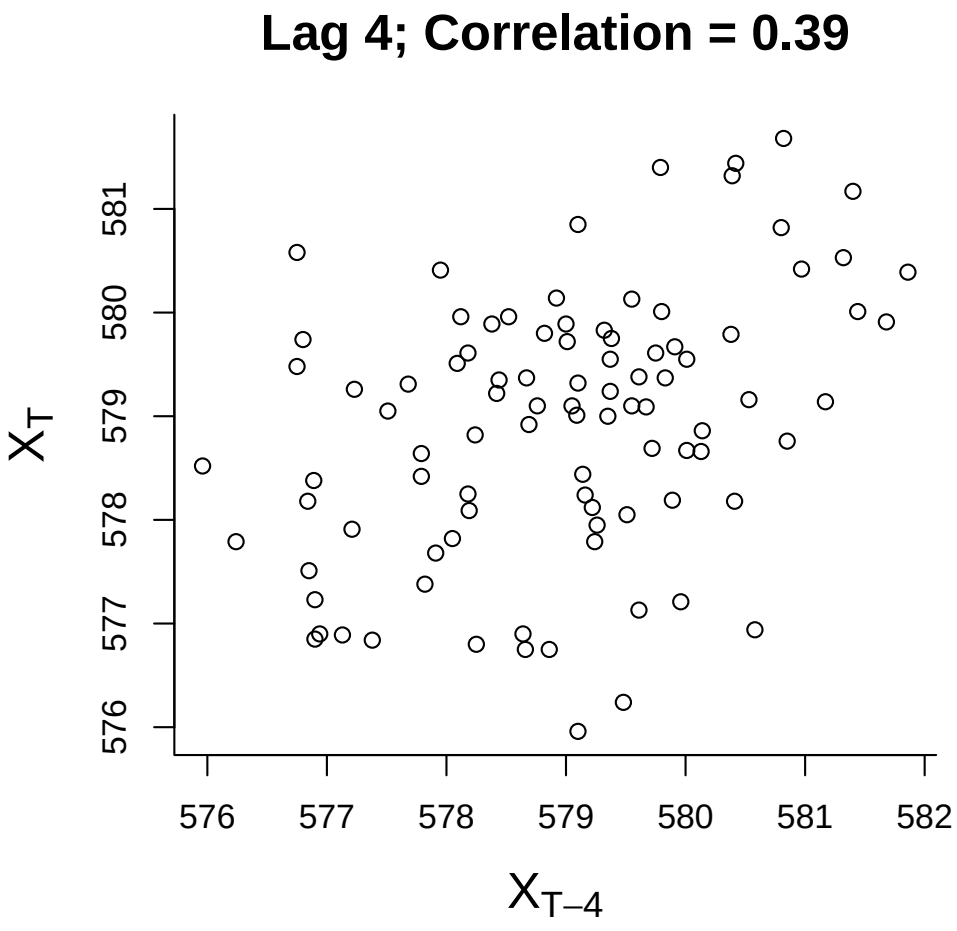


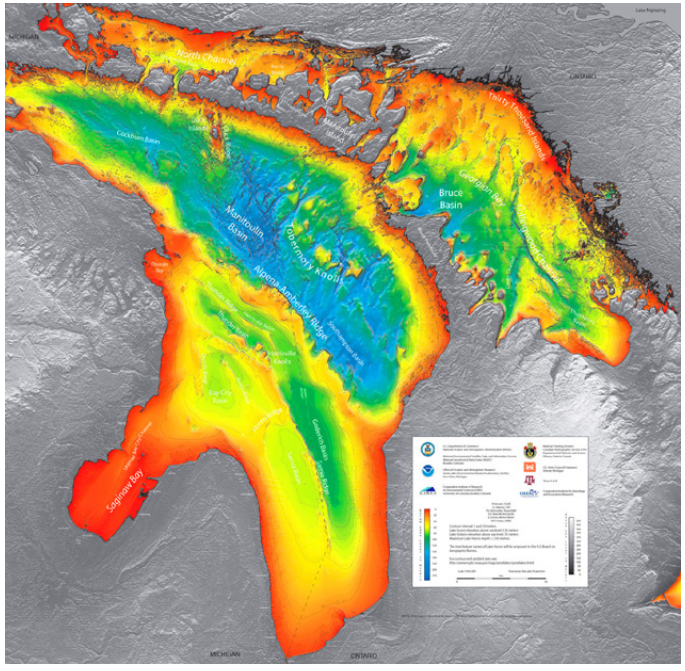


Autocovariance function: $\gamma(h) = \mathrm{Cov}\,(X_t, X_{t-h})$

Autocorrelation function: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{Cor}\,(X_t, X_{t+h})$

Concept 1: Autocorrelation





Autocovariance function: $\gamma(h) = \mathrm{Cov}\,(X_t, X_{t-h})$

Autocorrelation function: $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{Cor}\,(X_t, X_{t+h})$

Estimates

- Sample mean:

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$

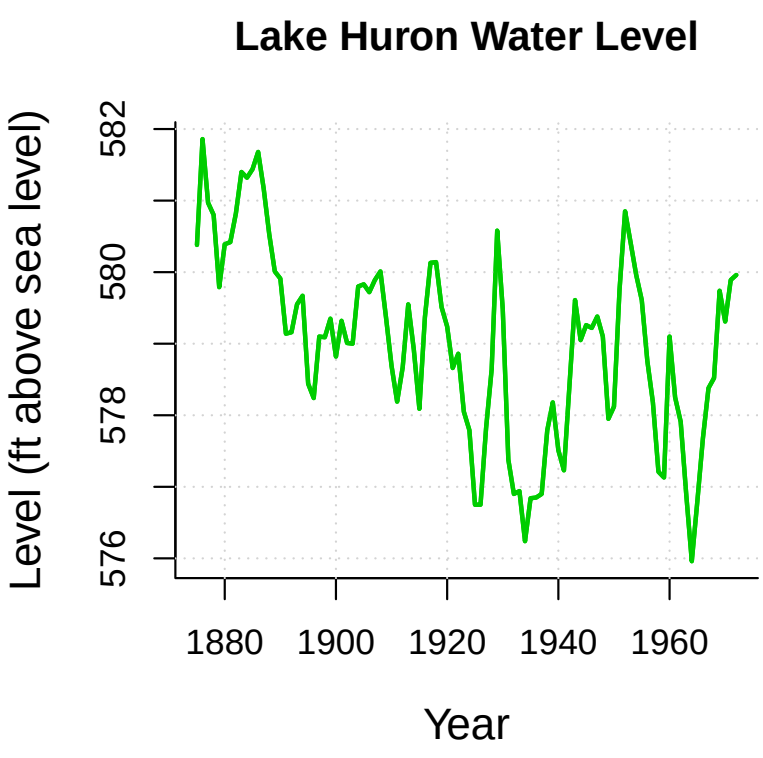
- Sample autocovariance:

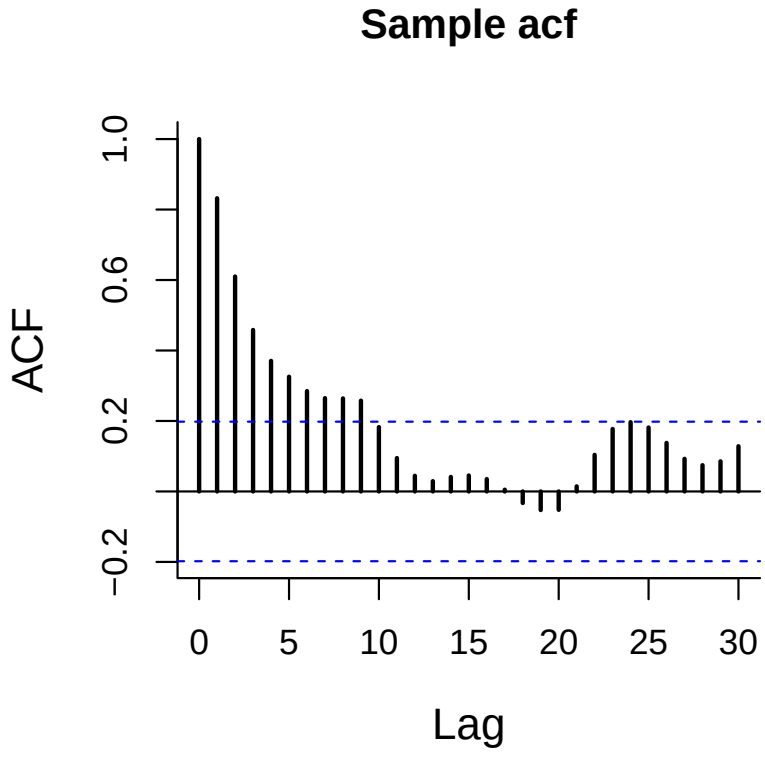
$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(X_{t+|h|} - \overline{X})(X_t - \overline{X})$$

- Sample autocorrelation:

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}$$

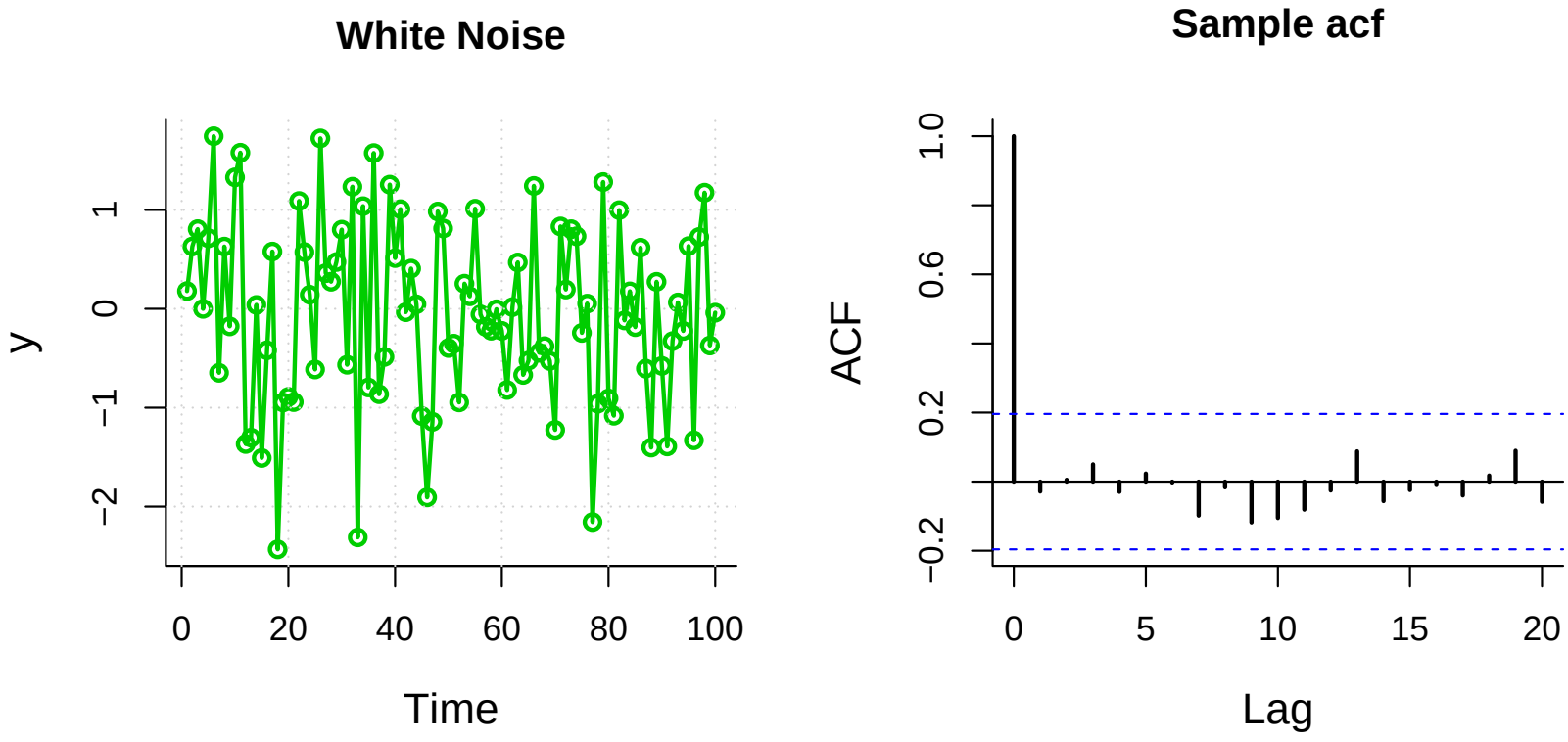
Sample autocorrelation





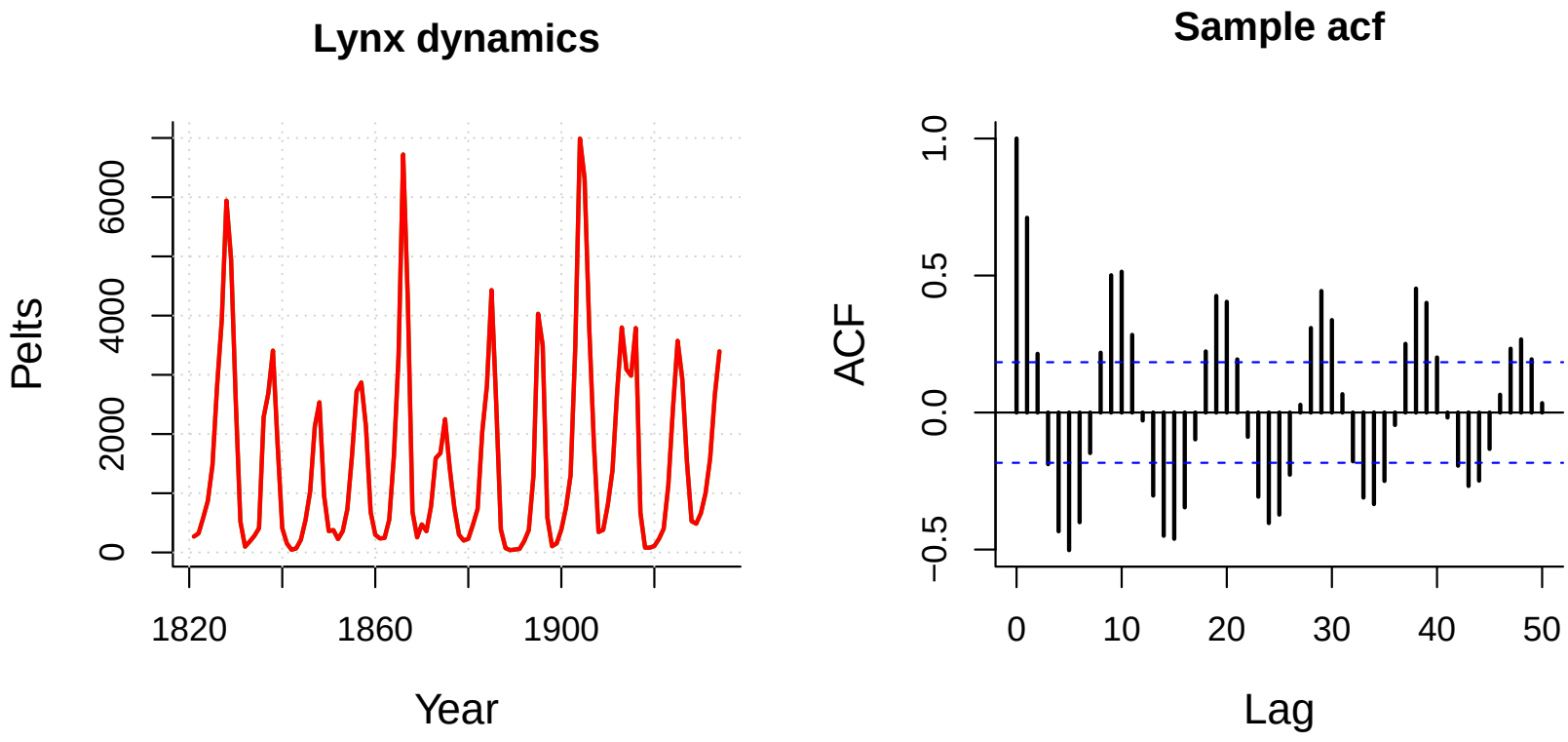
- Gives an immediate sense of the significance of correlation

Sample autocorrelation



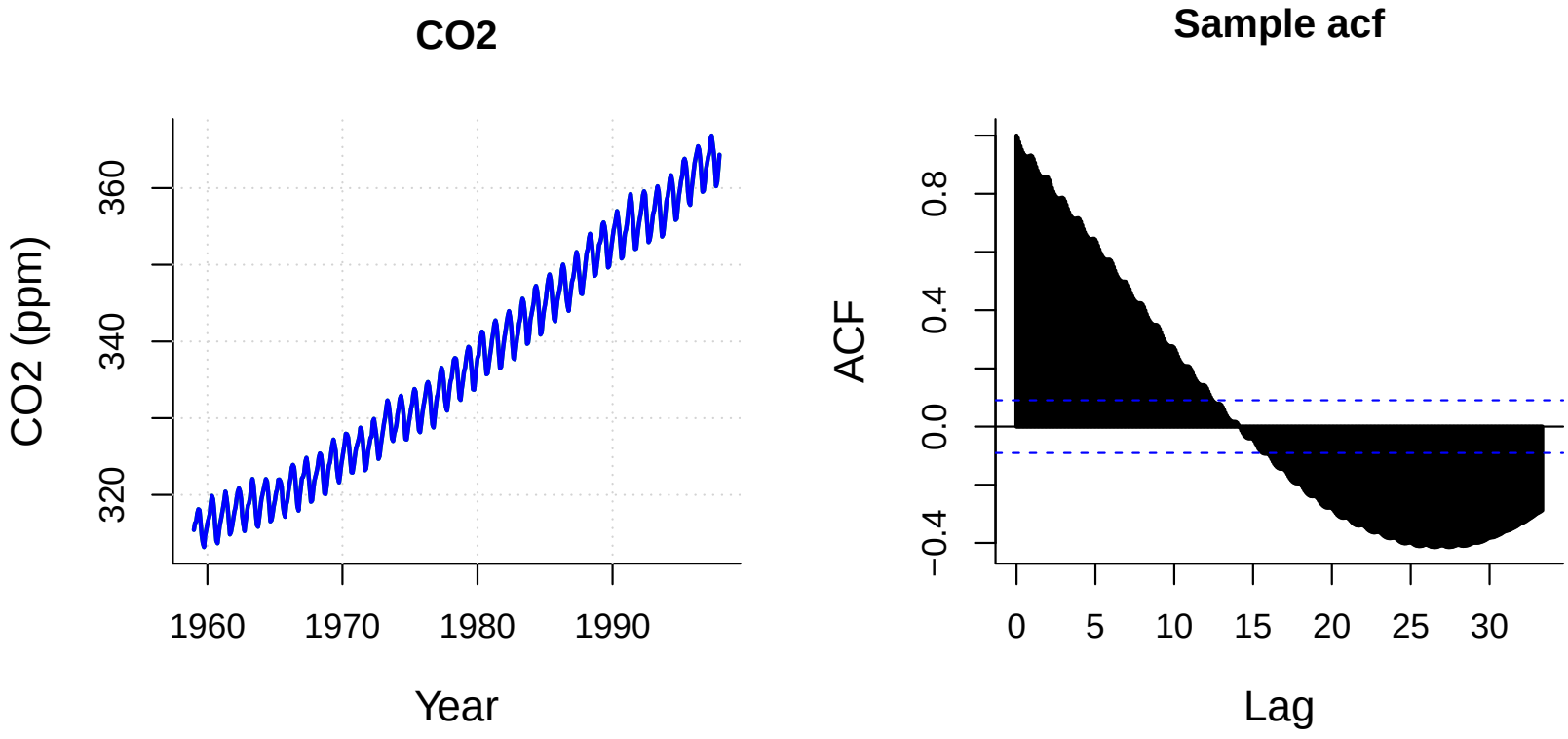
- Gives an immediate sense of the significance of correlation
  - Note, blue dashed line is $\frac{1.96}{\sqrt{n}}$, because expected sample autocorrelation for white noise is $\sim \mathcal{N}(0, 1/n)$

Sample autocorrelation



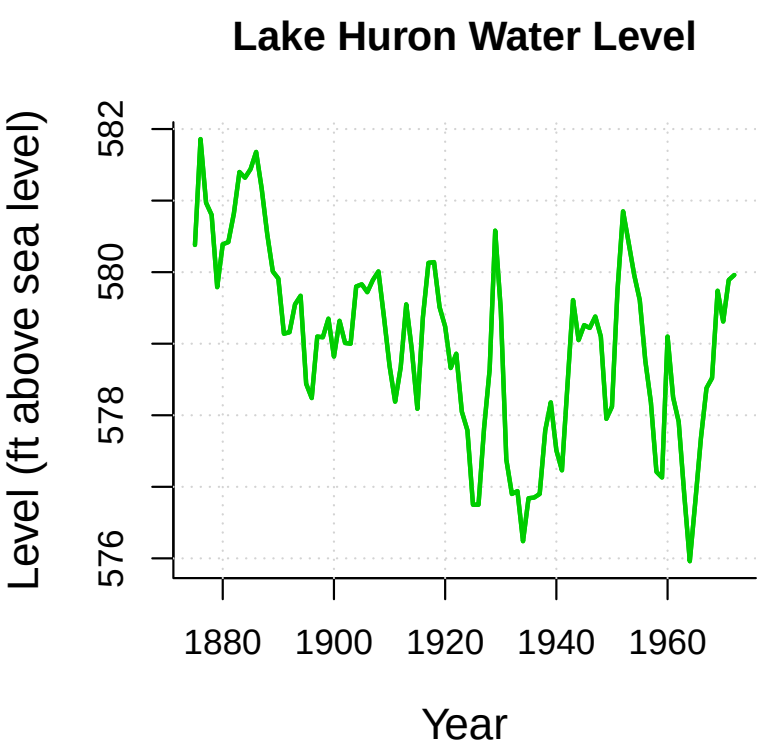
- Gives an immediate sense of the significance of correlation
- Gives a feel for "seasonal" patterns.

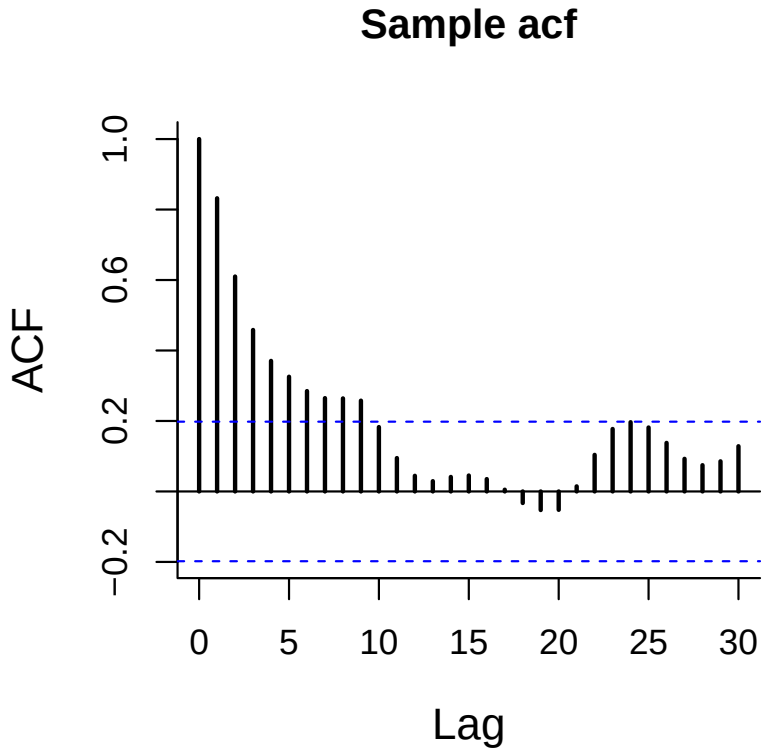
Sample autocorrelation



- Gives an immediate sense of the significance of correlation
- Gives a feel for "seasonal" patterns.
- Warning: Not very useful for time-series with very strong trends!

Sample autocorrelation





- Gives an immediate sense of the significance of correlation
- Gives a feel for "seasonal" patterns.
- Warning: Not very useful for time-series with very strong trends!
- In R: `acf(Y)`

The autoregressive model

- First order autoregressive model: AR(1)

$$X_t = \phi_1 X_{t-1} + Z_t$$

  where $Z_i \sim \mathcal{N}(0, \sigma^2)$ is **White Noise**.

- Second order autoregressive: AR(2)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

- p-th order autoregressive model: AR(p)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p Z_{t-p} + Z_t$$

- Note: these assume $\mathrm{E}(X) = 0$. Relaxing this gives (for AR(1)):

$$X_t = \phi_1 (X_{t-1} - \mu) + Z_t + \mu$$

The autoregressive model

- First order autoregressive model: AR(1)

$$X_t = \phi_1 X_{t-1} + Z_t$$

  where $Z_i \sim \mathcal{N}(0, \sigma^2)$ is **White Noise**.
- Second order autoregressive: AR(2)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

- p-th order autoregressive model: AR(p)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p Z_{t-p} + Z_t$$

- Note: these assume $\mathrm{E}(X) = 0$. Relaxing this gives (for AR(1)):

$$X_t = \phi_1 (X_{t-1} - \mu) + Z_t + \mu$$

The autoregressive model

- First order autoregressive model: AR(1)

$$X_t = \phi_1 X_{t-1} + Z_t$$

  where $Z_i \sim \mathcal{N}(0, \sigma^2)$ is **White Noise**.
- Second order autoregressive: AR(2)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

- p-th order autoregressive model: AR(p)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p Z_{t-p} + Z_t$$

- Note: these assume $\mathrm{E}(X) = 0$. Relaxing this gives (for AR(1)):

$$X_t = \phi_1 (X_{t-1} - \mu) + Z_t + \mu$$

The autoregressive model

- First order autoregressive model: AR(1)

$$X_t = \phi_1(X_{t-1} - \mu) + Z_t + \mu$$

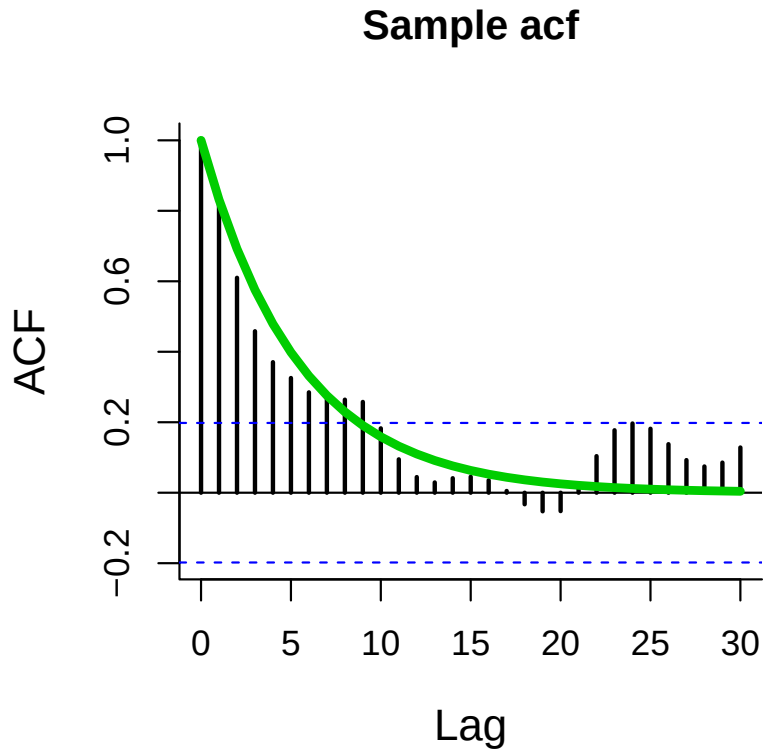
- Analogous to linear regression of X_t against X_{t-1}:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \epsilon_t$$

AR(1) Theoretical prediction

For the AR(1) model, the theoretical acf is:

$$\rho(h) = \phi_1^h$$

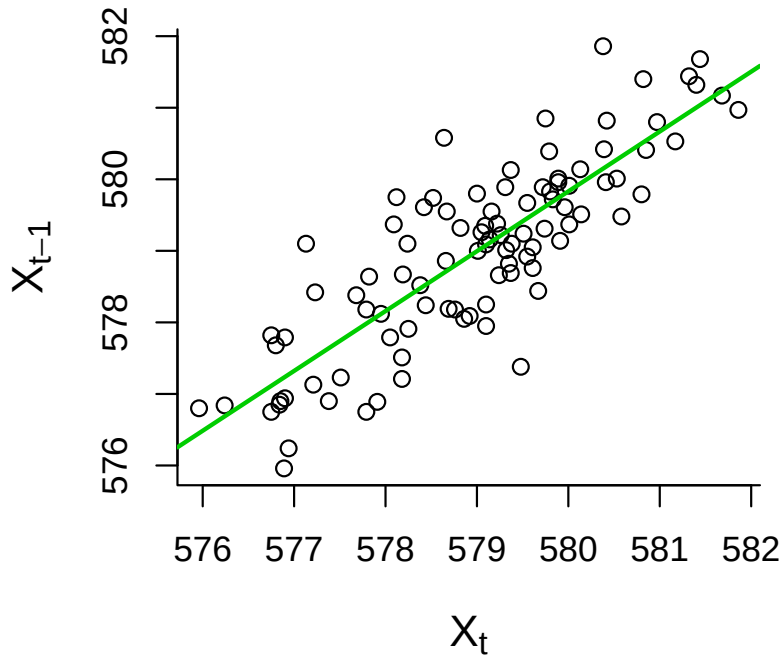


If the sample acf looks exponential - probably an AR(1) model.

Lake Huron AR(1)

Fit: $X_t = \phi_1(X_{t-1} - \mu) + Z_t + \mu$

In R: `LH.lm <- lm(LakeHuron[-1] ~ LakeHuron[-length(LakeHuron)])`



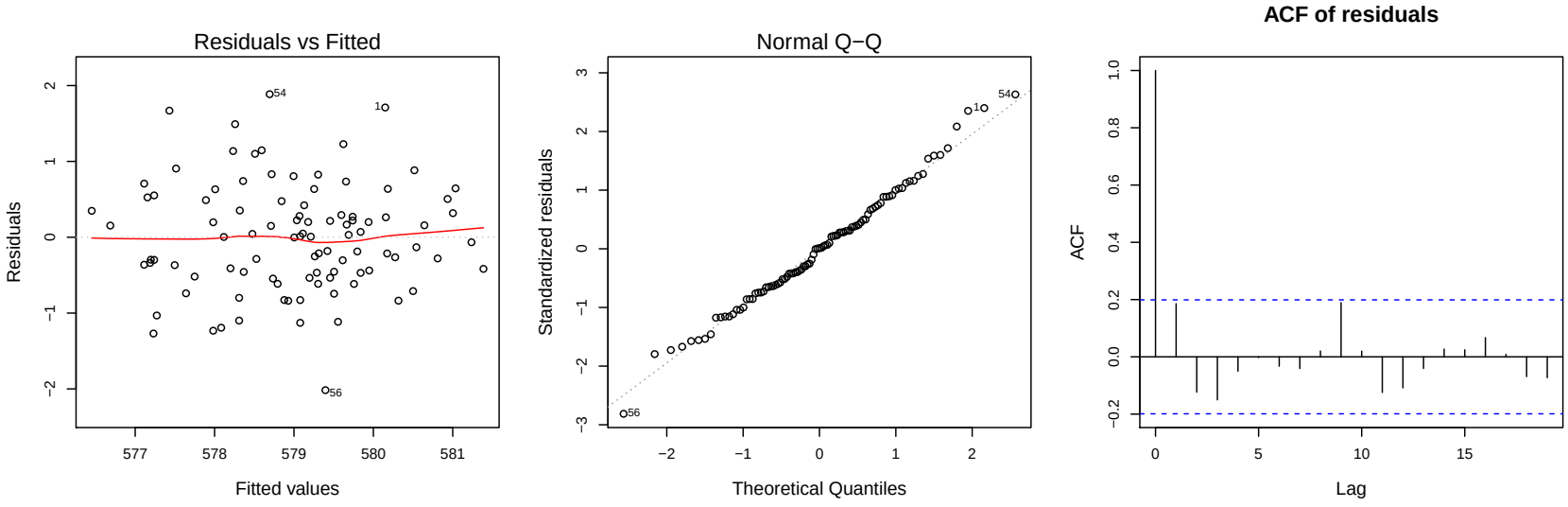
Results:

- $\phi_1 = 0.8364$; $\widehat{\mu} = 579$ ft; $\widehat{\sigma}^2 = 0.52$ ft^2

Note: $R^2 = 0.70$, i.e. about 70% percent of the variation observed in water levels is explained by previous years.

Diagnostic plots

```
plot(LakeHuron.lm); acf(residuals(LakeHuron.lm))
```



Check regression assumptions:

- Homoscedastic, Normal, Independent
- (note use of acf function to test assumption of independence!)

But what about the trend?

Decomposition with trend:

$$Y_t = m_t + X_t$$

where:

- m_t is the *slowly varying trend* component
- X_t is a random component
  - X_t can have serial autocorrelation
  - $\mathrm{E}(X_t) = 0$
  - X_t must be **stationary**.

Definition

- X_t is a **Stationary** process if $\mathrm{E}(X_t)$ is independent of t
- X_t is what is left over after the time-dependent part is removed

But what about the trend?

Decomposition with trend:

$$Y_t = m_t + X_t$$

where:

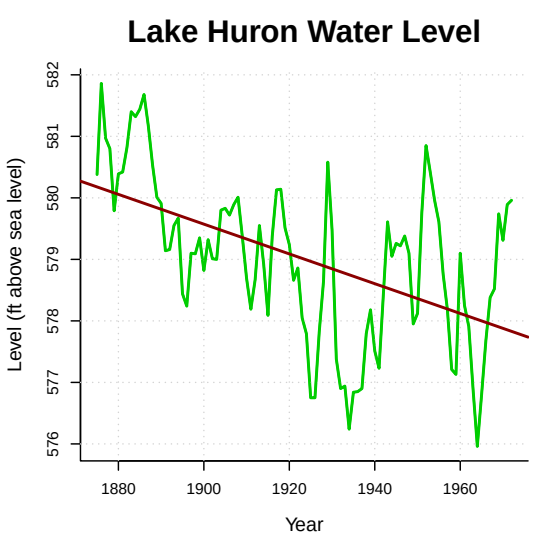
- m_t is the *slowly varying trend* component
- X_t is a random component
  - X_t can have serial autocorrelation
  - $\mathrm{E}(X_t) = 0$
  - X_t must be **stationary**.

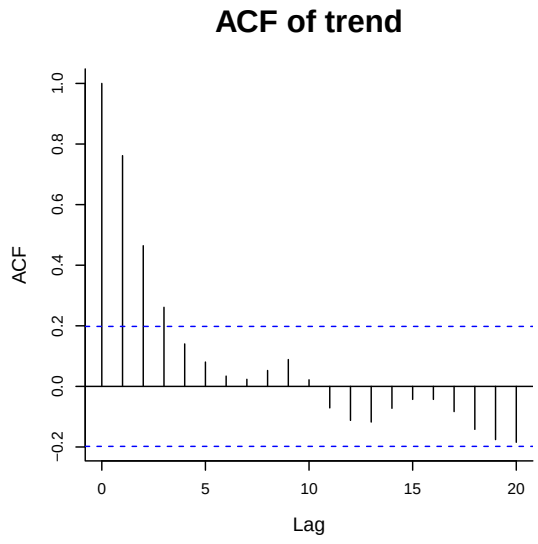
Definition

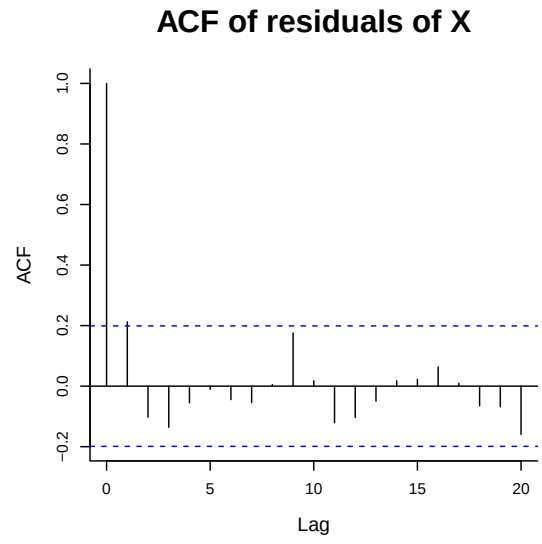
- X_t is a **Stationary** process if $\mathrm{E}(X_t)$ is independent of t
- X_t is what is left over after the time-dependent part is removed

Fitting a linear trend and correlation to Lake Huron

$$
\begin{aligned}
Y_T &= \beta_0 + \beta_1 T + X_T \\
X_T &= \phi_1 X_{T-1} + Z_T
\end{aligned}
$$







- $\widehat{\beta_0} = 625$ ft; $\widehat{\beta_1} = -0.024$ ft/year; $\widehat{\phi_1} = 0.79$; $\widehat{\sigma^2} = 0.513$ ft^2

Fitting a linear trend and correlation to Lake Huron in R

Version 1: Two steps

```
LH.trend <- lm(LakeHuron~time(LakeHuron))
X <- residuals(LH.trend)
X.lm <- lm(X[-1]~X[-length(X)])
```

Note that the simple linear regression gives a highly significant result:

	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	625.554918	7.764293	80.568	$<$ 2e-16 ***
time(LakeHuron)	-0.024201	0.004036	-5.996	3.55e-08 ***

Version 2: One step

```
require(nlme)
LH.gls <- gls(LakeHuron~time(LakeHuron), correlation =
corAR1(form=~1))
```

Fitting a linear trend and correlation to Lake Huron in R

Version 1: Two steps

```
LH.trend <- lm(LakeHuron~time(LakeHuron))
X <- residuals(LH.trend)
X.lm <- lm(X[-1]~X[-length(X)])
```

Note that the simple linear regression gives a highly significant result:

	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	625.554918	7.764293	80.568	$<$ 2e-16 ***
time(LakeHuron)	-0.024201	0.004036	-5.996	3.55e-08 ***

Version 2: One step

```
require(nlme)
LH.gls <- gls(LakeHuron~time(LakeHuron), correlation =
corAR1(form=~1))
```

Generalized least squares (gls) output

summary(LH.gls)

```
Model: LakeHuron ~ time(LakeHuron)
  AIC      BIC    logLik
  225.8304 236.0878 -108.9152

Correlation Structure: AR(1)
 Formula: ~ 1
 Parameter estimate(s):
      Phi
0.8247674

Coefficients:
                 Value Std.Error    t-value p-value
(Intercept)    616.4887 24.362632 25.304683  0.0000
time(LakeHuron) -0.0194  0.012664 -1.534616  0.1282
```

Note: The TIME effect is not significant in this model! Does this mean that there is no trend in Lake Huron levels?

Generalized least squares (gls) output

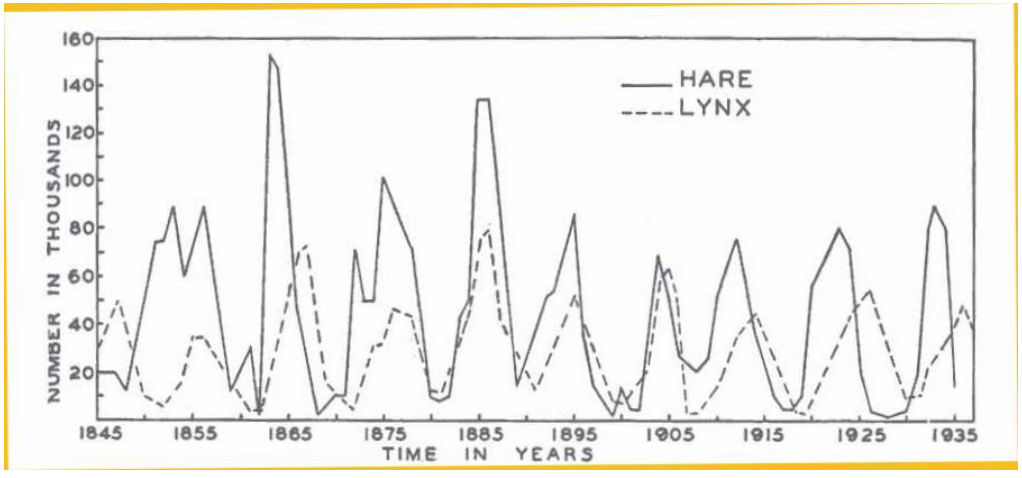
```
summary(LH.gls)
```

```
Model: LakeHuron ~ time(LakeHuron)
  AIC      BIC    logLik
225.8304 236.0878 -108.9152

Correlation Structure: AR(1)
Formula:  1
Parameter estimate(s):
      Phi
0.8247674

Coefficients:
                Value Std.Error    t-value    p-value
(Intercept)  616.4887 24.362632 25.304683 0.0000
time(LakeHuron) -0.0194  0.012664 -1.534616 0.1282
```

Note: The TIME effect is not significant in this model! Does this mean that there is no trend in Lake Huron levels?

Population cycles

Lynx and Hare



Interacting populations: Cross-correlation function

- Recall autocorrelation:
  - Autocorrelation function: $\rho(h) = \frac{\mathrm{E}[(X_t - \mu_x)(Y_{t+h} - \mu_y)]}{\sigma_x^2}$
  - Sample autocorrelation: $\widehat{\gamma}(h) = \frac{1}{ns_x^2} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \overline{X})(X_t - \overline{X})$
- Analogously
  - Cross-correlation function:

$$\rho_{xy}(h) = \frac{\mathrm{E}[(X_t - \mu_x)(Y_{t+h} - \mu_y)]}{\sigma_x \sigma_y}$$

  - Sample cross-correlation:

$$\widehat{\gamma}(h) = \frac{1}{ns_x s_y} \sum_{t=1}^{n-|h|} (Y_{t+|h|} - \overline{X})(X_t - \overline{X})$$

Cross-correlation allows identification of relationships BETWEEN time-series over different lags.

Interacting populations: Cross-correlation function

- Recall autocorrelation:
  - Autocorrelation function: $\rho(h) = \frac{\mathrm{E}[(X_t - \mu_x)(Y_{t+h} - \mu_y)]}{\sigma_x^2}$
  - Sample autocorrelation: $\widehat{\gamma}(h) = \frac{1}{ns_x^2} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \overline{X})(X_t - \overline{X})$

- Analogously
  - Cross-correlation function:

$$\rho_{xy}(h) = \frac{\mathrm{E}[(X_t - \mu_x)(Y_{t+h} - \mu_y)]}{\sigma_x \sigma_y}$$

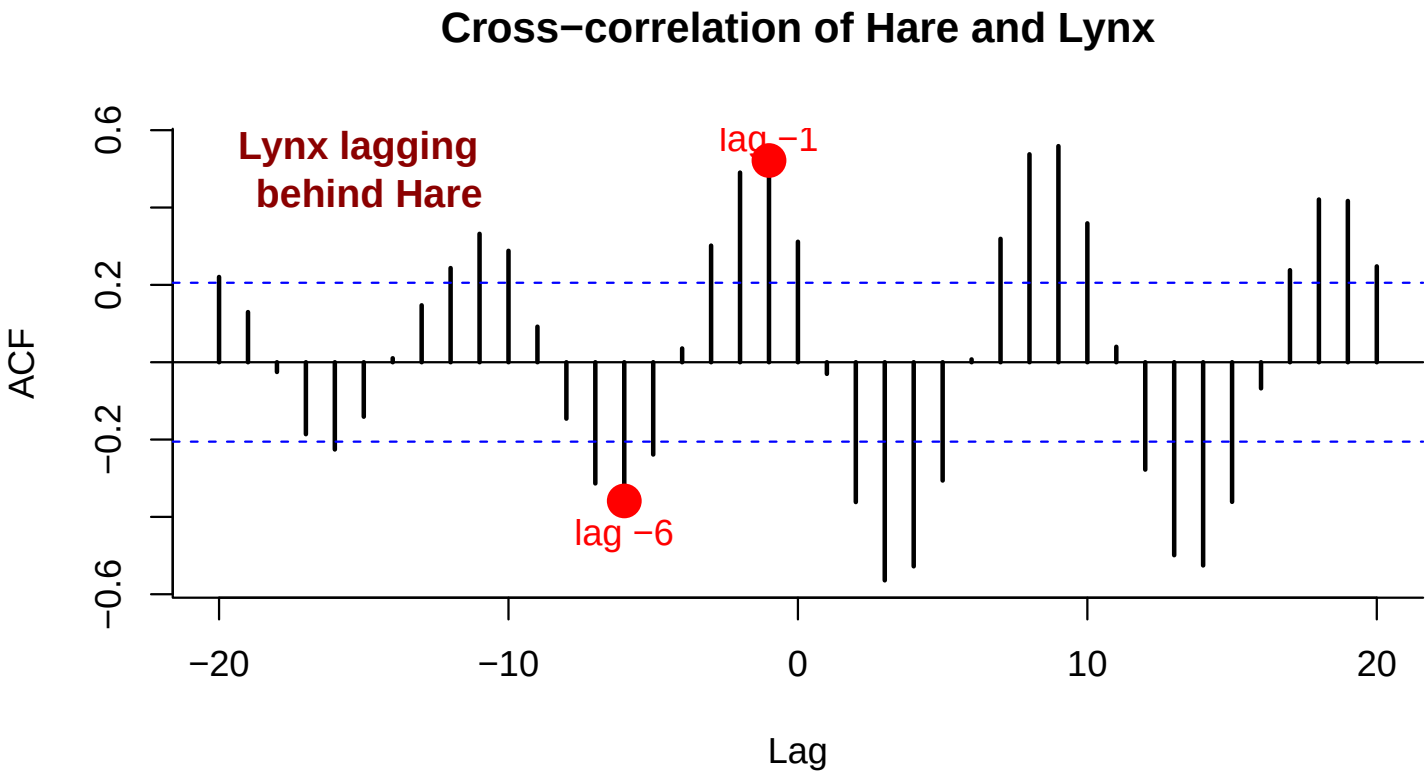
  - Sample cross-correlation:

$$\widehat{\gamma}(h) = \frac{1}{ns_x s_y} \sum_{t=1}^{n-|h|} (Y_{t+|h|} - \overline{X})(X_t - \overline{X})$$

Cross-correlation allows identification of relationships BETWEEN time-series over different lags.

Cross-correlation of lynx and hare

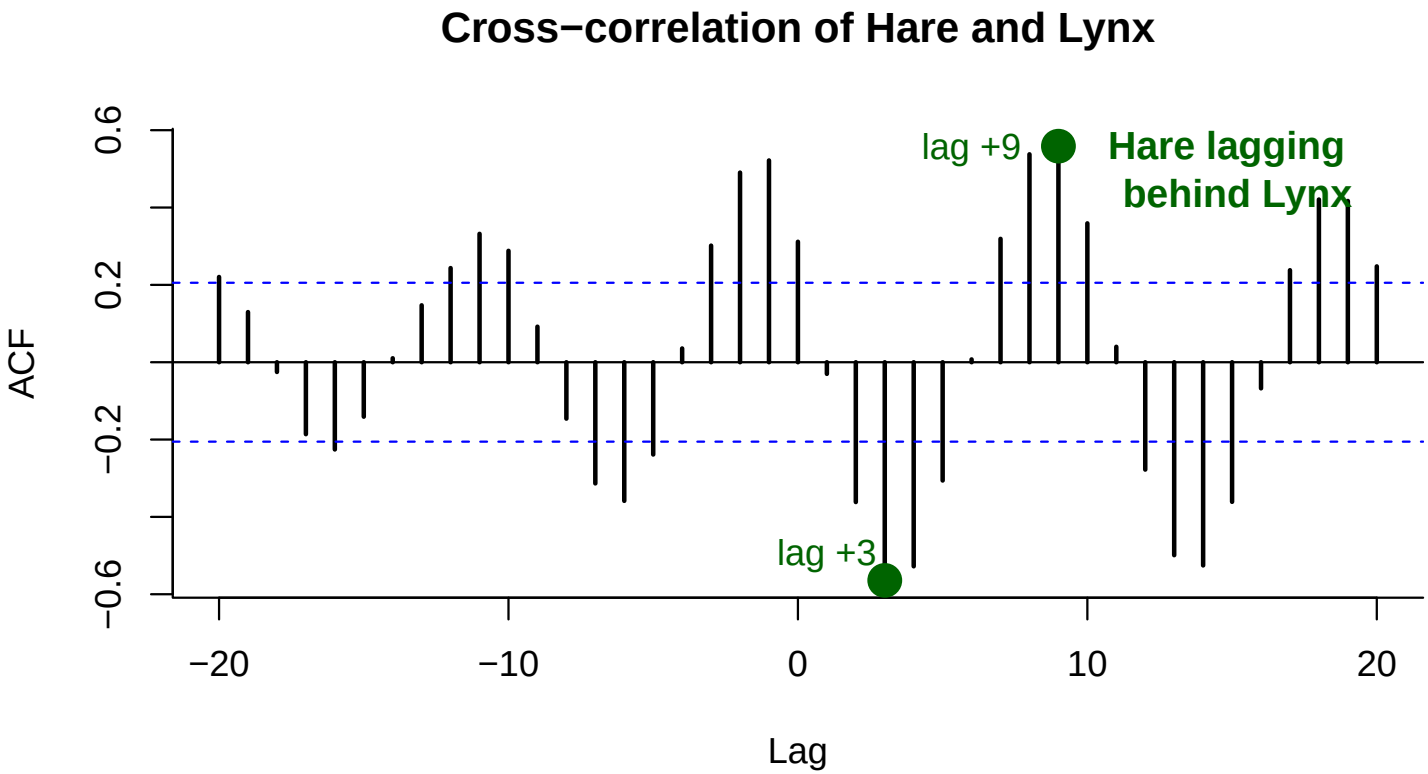
In R: `ccf(Hare, Lynx)`



Which of these effects is more important: Less Lynx leads to more Hare? More Hare leads to more Lynx? More Lynx leads to less Hare? Less Hare leads to less Lynx? **These are difficult, but not impossible, questions to parse!**

Cross-correlation of lynx and hare

In R: `ccf(Hare, Lynx)`



Which of these effects is more important: Less Lynx leads to more Hare? More Hare leads to more Lynx? More Lynx leads to less Hare? Less Hare leads to less Lynx? **These are difficult, but not impossible, questions to parse!**

Some intermediate conclusions

- If you do not account for correlation in your data, it is easy to make false inferences...
- BUT, there are interesting things to be learned from studying the (terrifying) lack of independence.
- A general strategy is to systematically extract all the patterns or otherwise reduce the data until you have extracted the stationary piece.
- Many basic concepts from regression are useful.
- There are many easy-to-use tools in R (and elsewhere) that, like Virgil, hold your hand as you make your journey through the inferno of dependent data.
- There are many ways to slice a time-series! As with any analysis - you have to be persistent and dig around and be creative.

Some intermediate conclusions

- If you do not account for correlation in your data, it is easy to make false inferences...
- BUT, there are interesting things to be learned from studying the (terrifying) lack of independence.
- A general strategy is to systematically extract all the patterns or otherwise reduce the data until you have extracted the stationary piece.
- Many basic concepts from regression are useful.
- There are many easy-to-use tools in R (and elsewhere) that, like Virgil, hold your hand as you make your journey through the inferno of dependent data.
- There are many ways to slice a time-series! As with any analysis - you have to be persistent and dig around and be creative.

Some intermediate conclusions

- If you do not account for correlation in your data, it is easy to make false inferences...
- BUT, there are interesting things to be learned from studying the (terrifying) lack of independence.
- A general strategy is to systematically extract all the patterns or otherwise reduce the data until you have extracted the stationary piece.
- Many basic concepts from regression are useful.
- There are many easy-to-use tools in R (and elsewhere) that, like Virgil, hold your hand as you make your journey through the inferno of dependent data.
- There are many ways to slice a time-series! As with any analysis - you have to be persistent and dig around and be creative.

Some intermediate conclusions

- If you do not account for correlation in your data, it is easy to make false inferences...
- BUT, there are interesting things to be learned from studying the (terrifying) lack of independence.
- A general strategy is to systematically extract all the patterns or otherwise reduce the data until you have extracted the stationary piece.
- Many basic concepts from regression are useful.
- There are many easy-to-use tools in R (and elsewhere) that, like Virgil, hold your hand as you make your journey through the inferno of dependent data.
- There are many ways to slice a time-series! As with any analysis - you have to be persistent and dig around and be creative.

Some intermediate conclusions

- If you do not account for correlation in your data, it is easy to make false inferences...
- BUT, there are interesting things to be learned from studying the (terrifying) lack of independence.
- A general strategy is to systematically extract all the patterns or otherwise reduce the data until you have extracted the stationary piece.
- Many basic concepts from regression are useful.
- There are many easy-to-use tools in R (and elsewhere) that, like Virgil, hold your hand as you make your journey through the inferno of dependent data.
- There are many ways to slice a time-series! As with any analysis - you have to be persistent and dig around and be creative.

Some intermediate conclusions

- If you do not account for correlation in your data, it is easy to make false inferences...
- BUT, there are interesting things to be learned from studying the (terrifying) lack of independence.
- A general strategy is to systematically extract all the patterns or otherwise reduce the data until you have extracted the stationary piece.
- Many basic concepts from regression are useful.
- There are many easy-to-use tools in R (and elsewhere) that, like Virgil, hold your hand as you make your journey through the inferno of dependent data.
- There are many ways to slice a time-series! As with any analysis - you have to be persistent and dig around and be creative.


Have Hope!