Statistical Techniques for Data Science & Robotics

Week 7

Quiz

Jackknife and Bootstrap

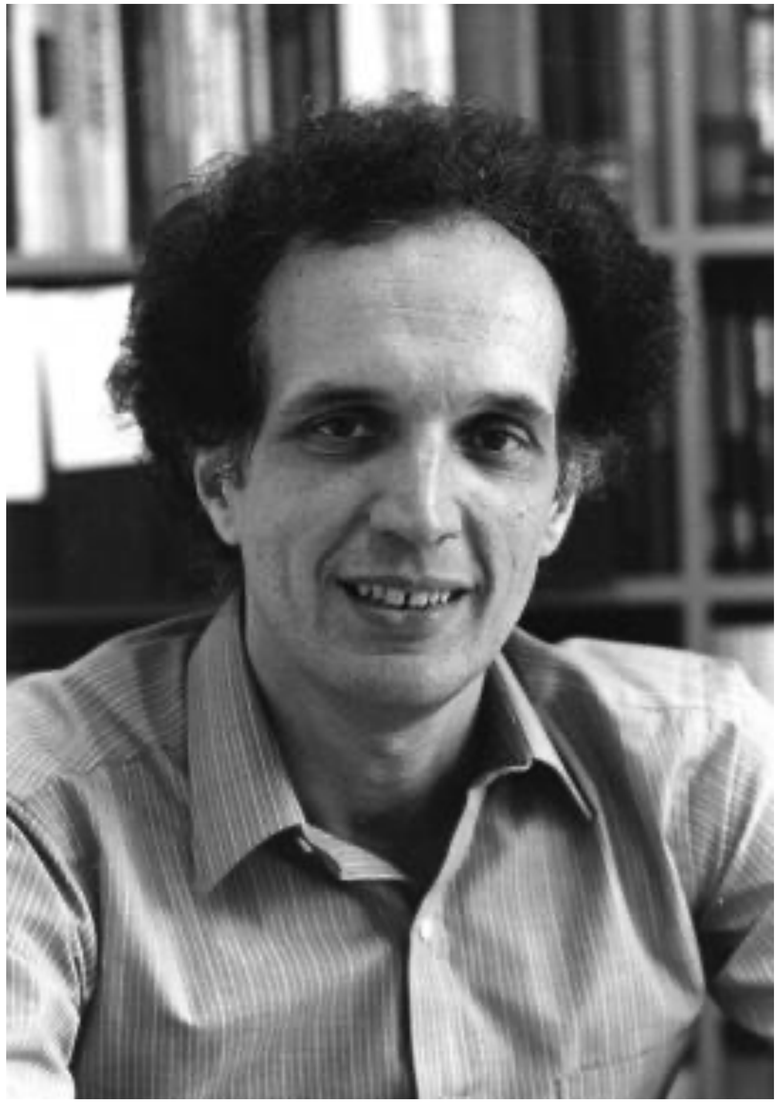
Objectives (for today)

- to understand
  - Jackknife
  - Bootstrapping
- to apply
  - Bootstrapping

Resampling

versus classical

- The **classical statistical methods** are based on the statistical properties of the estimators
- They have **a simple closed form,** which can be analyzed mathematically
- Usually, these methods have **unrealistic assumptions**
- **Resampling methods** provide inference on a wide range of statistics under very general conditions



Resampling

- Resampling methods involve constructing hypothetical ‘populations’ derived from the observations.
- Resampling the original data **preserves whatever distributions are truly present**.
-

Jackknife

The jackknife is a simple resampling method for approximating the bias and variance of an estimator

Jackknife. Step by Step (1/4)

We start with informal description of the idea of jackknife.
The jackknife estimation of a parameter θ is an iterative process

Step 1. The parameter θ is estimated from the whole sample X

For example the original sample $X_4 = (3, 2, 8, 4)$.

We can use the sample to calculate value of T():

$$T(X_n) \equiv T(X_4) \equiv \phi_n(X_n) \equiv \phi_4(X_4)$$

Jackknife. Step by Step (2/4)

Step 2. Then each element of X_n is, in turn, dropped from the sample and the parameter θ is estimated from this smaller sample.
This estimation is called *a partial estimate* (or also a jackknife replication)

This is where the resampling happens.

From the original sample $(3, 2, 8, 4)$ we resample the following 4 subsamples (by removal of one of values):
$X_{[1]} = (2, 8, 4)$, $X_{[2]} = (3, 8, 4)$,
$X_{[3]} = (3, 2, 4)$, $X_{[4]} = (3, 2, 8)$

We calculate partial estimates: $\phi_3(X_{[1]})$, $\phi_3(X_{[2]})$, $\phi_3(X_{[3]})$, $\phi_3(X_{[4]})$

Jackknife. Step by Step (3/4)

Step 3. *A pseudo-values* (ps_i) are then computed as the difference between the whole sample estimate ($\phi_n(X)$) and each partial estimate ($\phi_{n-1}(X_{[i]})$).

$$
\begin{aligned}
ps_1 &= 4 * \phi_4(X_4) - 3 * \phi_3(X_{[1]}),\\
ps_2 &= 4 * \phi_4(X_4) - 3 * \phi_3(X_{[2]}),\\
ps_3 &= 4 * \phi_4(X_4) - 3 * \phi_3(X_{[3]}),\\
ps_4 &= 4 * \phi_4(X_4) - 3 * \phi_3(X_{[4]}),
\end{aligned}
$$

This is important: Such calculation of pseudo-values reduces the (linear) bias of the partial estimate (because the bias is eliminated by the subtraction between the two estimates). It means that each pseudo-value is a bias-corrected estimator

Jackknife. Step by Step (4/4)

Step 4. Finally, the pseudo-values are used instead of the original values to estimate the parameter of interest.

Standard deviation of the pseudo-values is used to estimate the parameter standard error which can then be used for null hypothesis testing and for computing confidence intervals.

Resampling: Bootstrap

Motivation of the bootstrap

- The bootstrap is a nonparametric method for computing **standard errors and confidence intervals**.

- The traditional approach to statistical inference relies on asymptotic theory and **are not available for small samples**

- The bootstrapping method, introduced by Efron in 1979, is a computational intensive resampling method, which is widely applicable and allows the treatment of more realistic models.

A quick view of Bootstrap

- **It has minimum assumptions.** It is merely based on the assumption that the sample is a good representation of the unknown population

- **It is not a black box method.** It works for the majority of problems but it may be problematic for some others

- In practice it is **computationally demanding**, but the progress on computer speed makes it easily available in everyday practice

How?

- Given a sample $X^n = (X_1, \ldots, X_n)$
- Build an ECDF $\hat{F}_n$
- Sample from the ECDF with replacement
  - i.e. with returns
- Calculate statistic of interest for B resamples (e.g. $B \geq 1000$)
- Finally, you got a bootstrap sampling distribution of a statistic

The Bootstrap schema

Real World

Unknown probability distribution

Observed random sample

$$P \longrightarrow X = (X_1, \ldots, X_n)$$

$$\downarrow$$

$$\hat{\theta} = s(X)$$

Statistic of interest

The Bootstrap schema

Real World

Unknown probability distribution

Observed random sample

$$P \longrightarrow X = (X_1, \ldots, X_n)$$

$$\downarrow$$

$$\hat{\theta} = s(X)$$

Statistic of interest

$$\Longrightarrow$$

Bootstrap World

Empirical distribution

Bootstrap sample

$$\hat{P} \longrightarrow X^* = (X_1^*, \ldots, X_n^*)$$

$$\downarrow$$

$$\hat{\theta}^* = s(X^*)$$

Bootstrap replication

Bootstrap: Formal Procedure

The general bootstrap algorithm

1. Generate a sample $\mathbf{x}^*$ of size n from $\hat{F}_n$.

Bootstrap: Formal Procedure

The general bootstrap algorithm

1. Generate a sample $\mathbf{x}^*$ of size n from $\hat{F}_n$.
2. Compute $\hat{\theta}^*$ for this bootstrap sample

Bootstrap: Formal Procedure

The general bootstrap algorithm

1. Generate a sample $\mathbf{x}^*$ of size n from $\hat{F}_n$.
2. Compute $\hat{\theta}^*$ for this bootstrap sample
3. Repeat steps 1 and 2, B time.

Bootstrap: Formal Procedure

The general bootstrap algorithm

1. Generate a sample $\mathbf{x}^*$ of size n from $\hat{F}_n$.
2. Compute $\hat{\theta}^*$ for this bootstrap sample
3. Repeat steps 1 and 2, B time.

By this procedure we end up with bootstrap values $\hat{\theta^*} = (\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*)$. We will use these bootstrap values for calculating all the quantities of interest.

A Numeric Example

Data: Mouse data

- Survival times of 16 mice after a test surgery
- 7 mice in treatment group (new medical treatment)
- 9 mice in control group (no treatment)

Group	*Survival time (in days)*									Mean
Treatment	94	197	16	38	99	141	23			86.86
Control	52	104	146	10	51	30	40	27	46	56.22

Question: Did treatment prolong survival?

$$\hat{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

- in this case $\hat{\theta}_i^*$ is mean, thus we estimate the distribution of means

i

$\hat{\theta}_i^*$

A Numeric Example

Data: Mouse data

- Survival times of 16 mice after a test surgery
- 7 mice in treatment group (new medical treatment)
- 9 mice in control group (no treatment)

Group	*Survival time (in days)*									Mean
Treatment	94	197	16	38	99	141	23			86.86
Control	52	104	146	10	51	30	40	27	46	56.22

Question: Did treatment prolong survival?

$$\hat{\theta}^* = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i^*$$

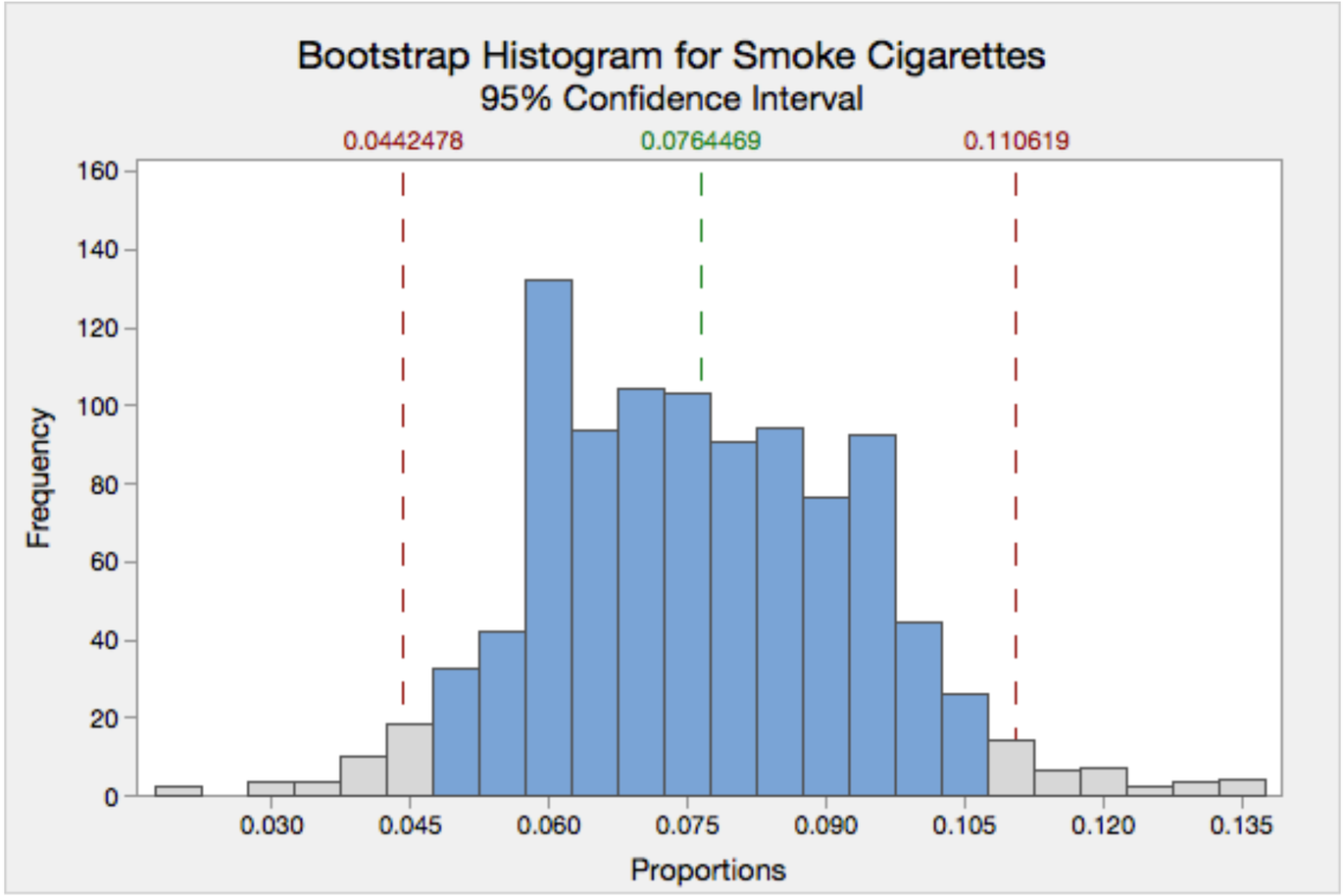
- in this case $\hat{\theta}_i^*$ is mean, thus we estimate the distribution of means

i								$\hat{\theta}_i^*$
1	38	141	94	16	99	197	23	86.9
2	94	23	197	16	141	38	94	86.1
3	16	141	94	23	94	38	99	72.1
4	16	94	94	23	99	197	16	77.0
5	38	141	16	99	16	141	141	84.6
6	197	16	197	94	16	16	16	78.9
7	99	23	94	23	38	197	99	81.9
8	38	38	38	23	16	99	38	41.4
9	23	38	141	94	23	94	23	62.3
10	38	23	141	94	38	141	197	96.0
11	38	38	38	99	197	141	141	98.9
12	38	23	38	99	23	38	99	51.1
13	23	94	197	99	99	16	99	89.6
14	38	16	16	38	141	38	141	61.1
15	94	38	16	94	23	38	141	63.4
16	23	197	94	16	38	99	99	80.9
17	38	99	16	38	16	197	38	63.1
18	197	16	141	16	16	94	197	96.7
19	141	38	94	197	38	23	16	78.1
20	23	99	23	16	197	99	23	68.6

Types of Bootstrap

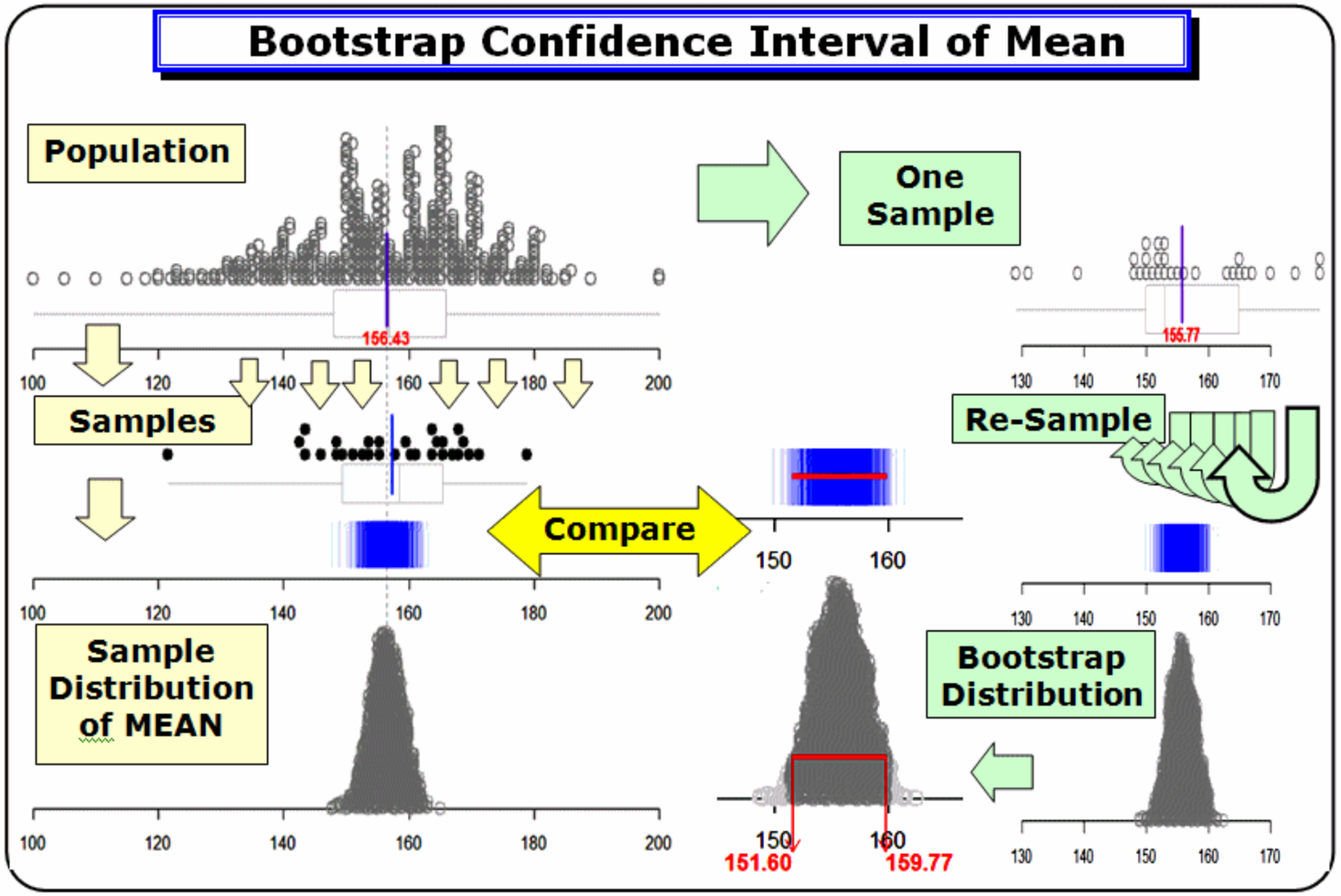
- **Balanced Bootstrap**:
  - In the original sample, we can found extreme values very rarely. Thus, bootstrapped samples are biased and we need to correct this
- **Parametric Bootstrap**:
  - We know that F belongs to a parametric family of distributions and we just estimate its parameters from the sample. We generate samples from F using the estimated parameters.
- **Non-parametric Bootstrap**:
  - We do not know the form of F and we estimate it by $\hat{F}_n$ the empirical distribution obtained from the data
  - e.g. we take 2.5 and 97.5 percentiles from the distribution of bootstrap samples

Visualization



- https://onlinecourses.science.psu.edu/stat200/book/export/html/97

Visualization



- https://maths.nayland.school.nz/Year_13_Maths/3.10_Inference/Images/ScreenShot145.gif

Exercise

- Derive the probability that a given observation is part of a bootstrap sample.
  - Suppose that we obtain a bootstrap sample from a set of n observations.
  - What is the probability that the first bootstrap observation in the resample is **not** the jth observation from the original sample?
    (Hint: $(1 - 1/n)^n$)

Exercise: Homework

- Investigate numerically the probability that a bootstrap sample of size n = 100 contains the j-th observation.

Calculating Confidence intervals

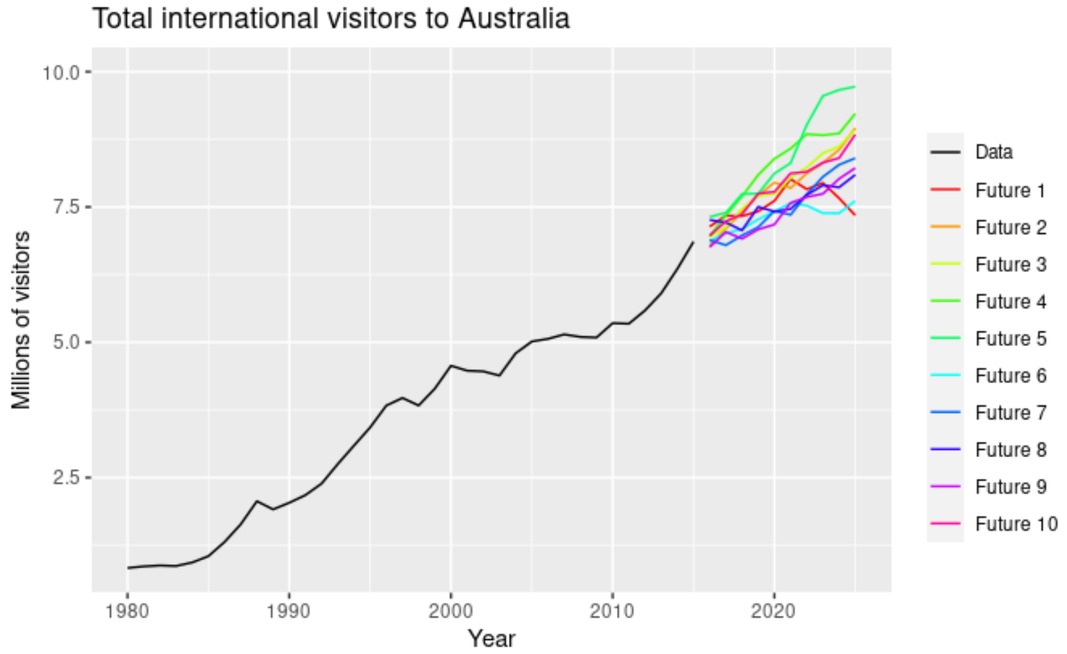
- Check the code:
  - https://github.com/rasbt/data-science-tutorial/blob/master/code/bootstrapping.ipynb

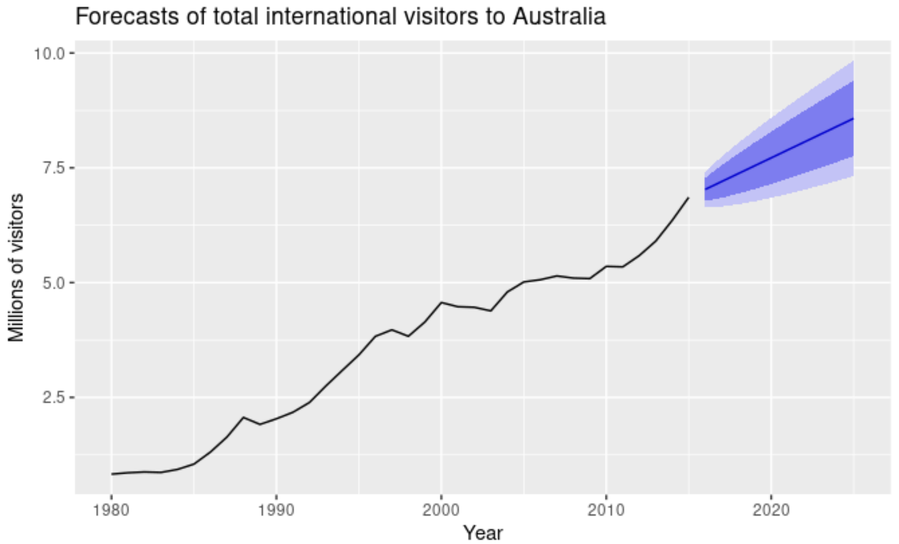
Break

Bootstrap in Time series

Time series. Forecasting and residuals

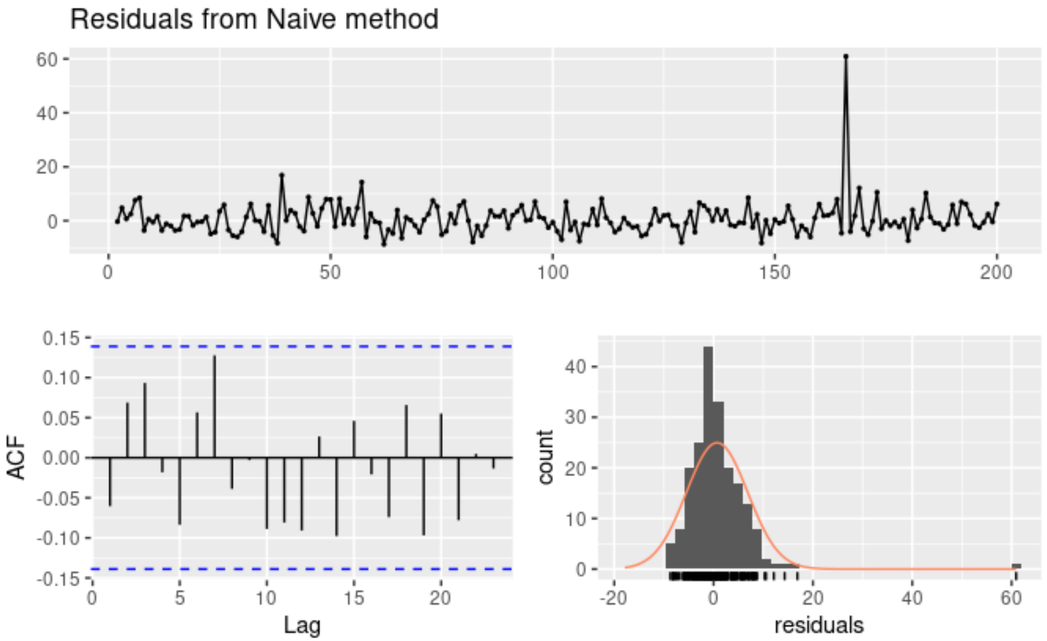
- Series $y_1, y_2, \ldots, y_t, \ldots, y_n$
- When we talk about the “forecast” we usually mean the average value of the forecast distribution
- We put a “hat”
  - $\hat{y}$ is a forecasted value
- If we take into account all previous observations, we write $\hat{y}_{t|t-1}$

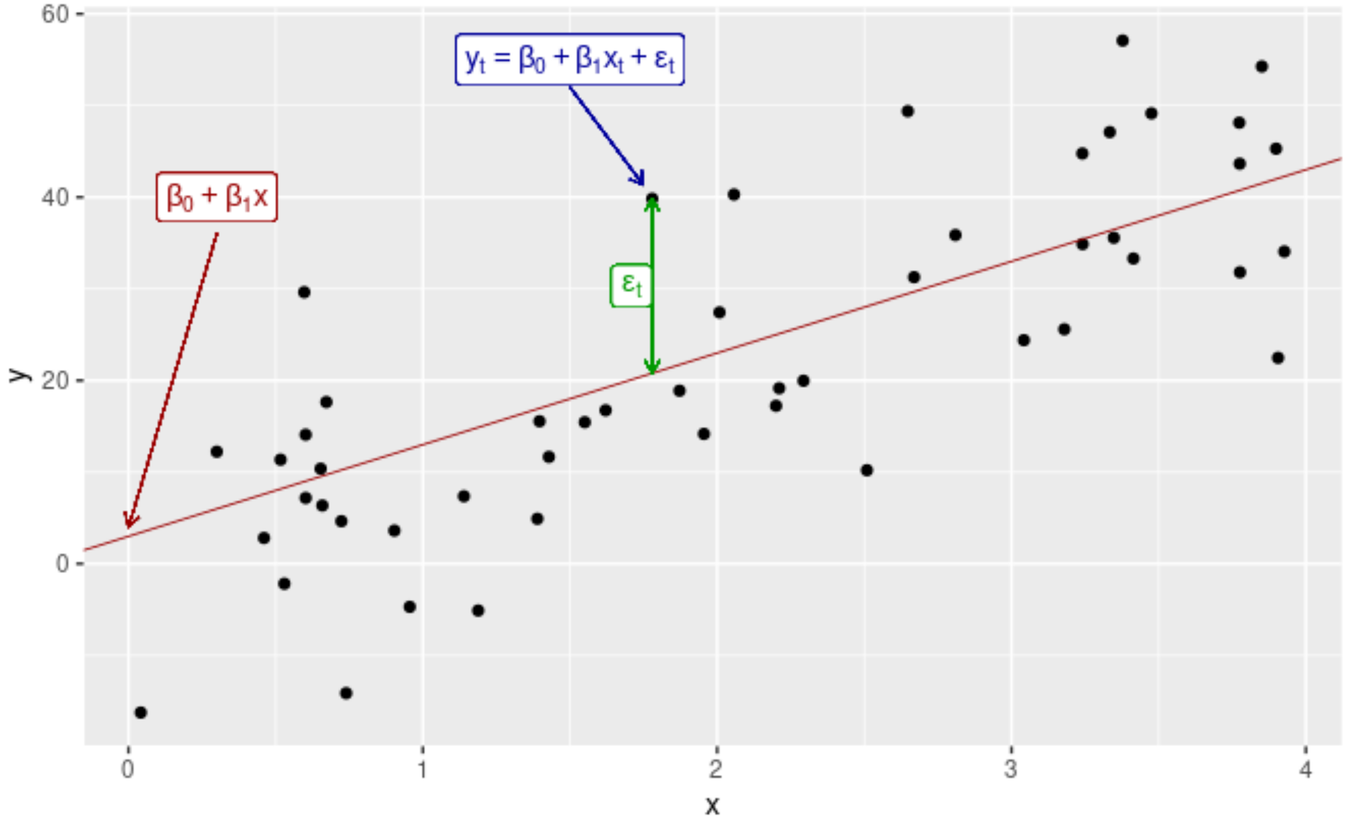




Time series. Residuals

- The “residuals” in a time series model are what is left over after fitting a model.



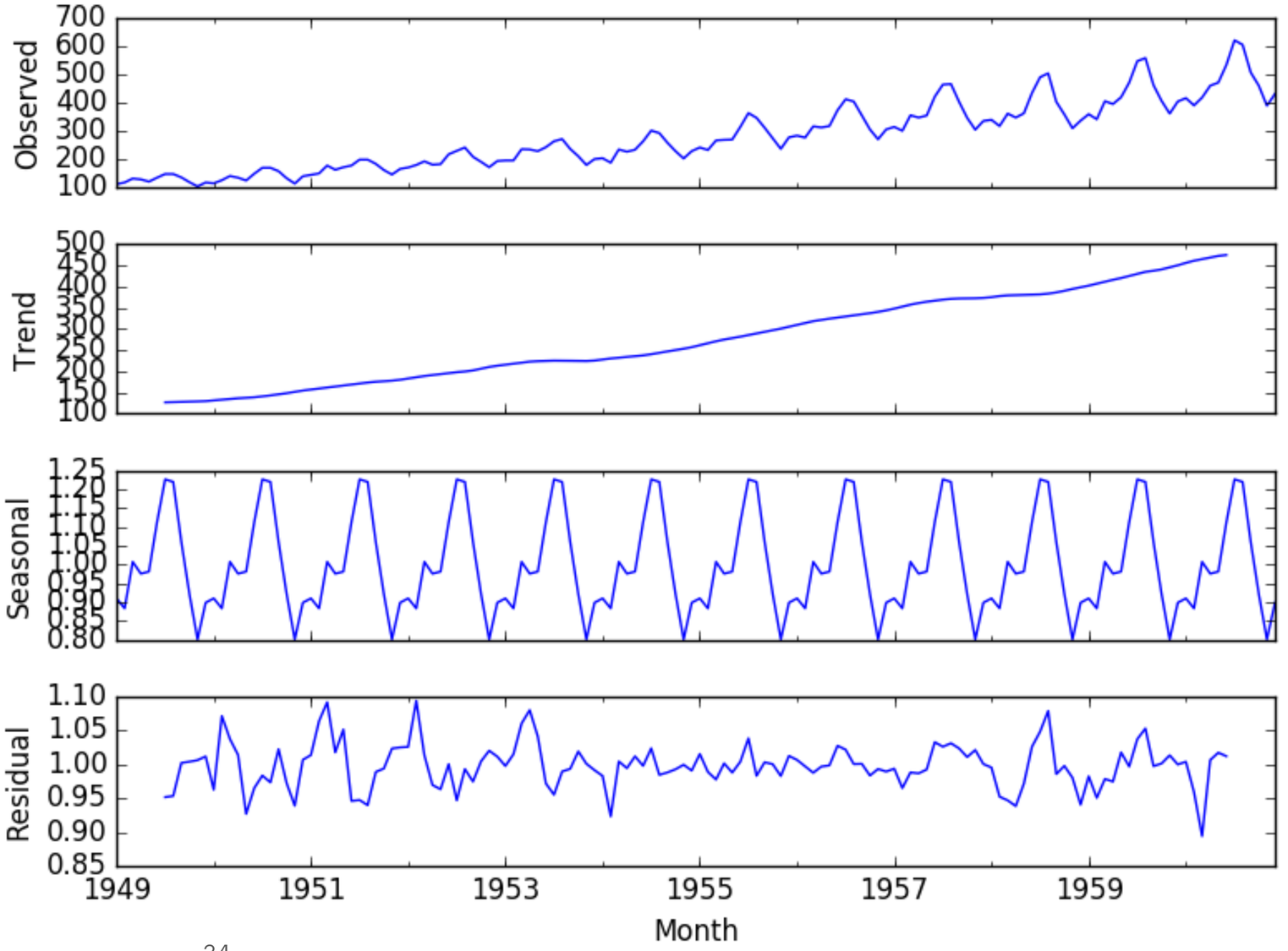


a simple linear model

Time Series. Decomposition

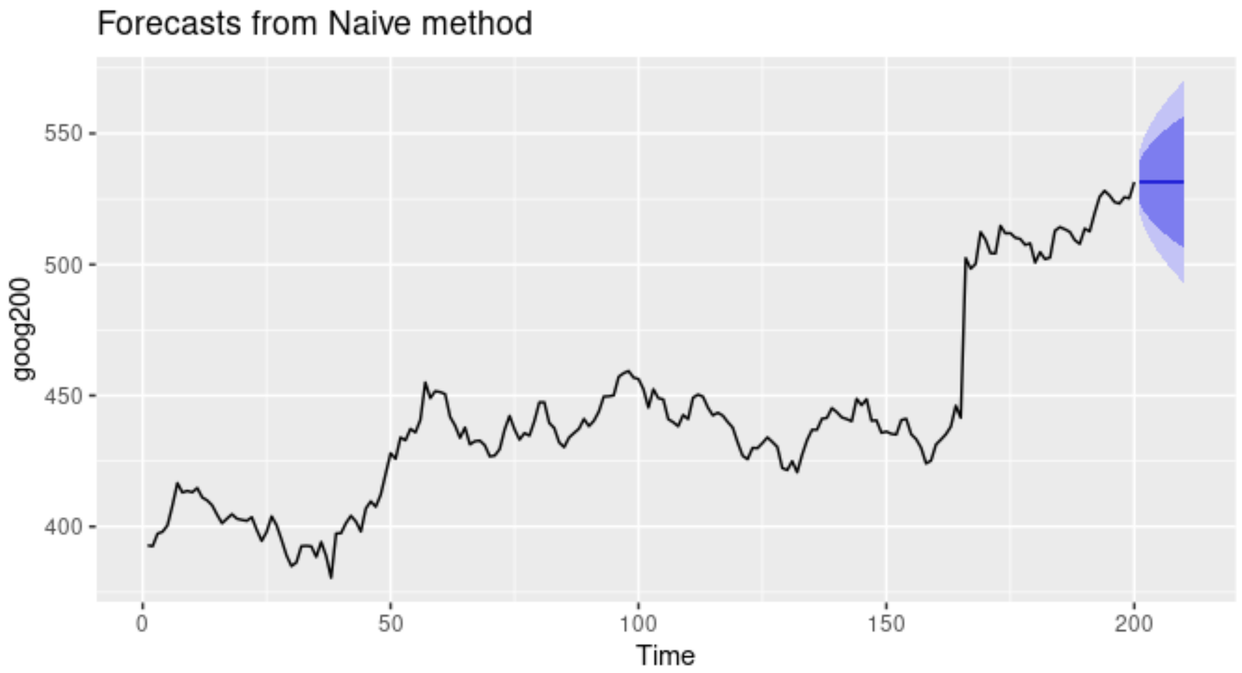
- Common approach is to decompose a time series into 3 components:
  - **Trend,**
  - **Seasonal,**
  - **Residual**

- So, the time series is a summation of them



Forecasting and Prediction intervals

- A prediction interval gives an interval within which we expect $y_{T+h|T}$ to lie with a specified probability.
- For example, assuming that the forecast errors are normally distributed, a 95% prediction interval
  - $y_{T+h|T} \pm 1.96\sigma_h$
  - where σ_h is the standard deviation of the h-step forecast distribution.

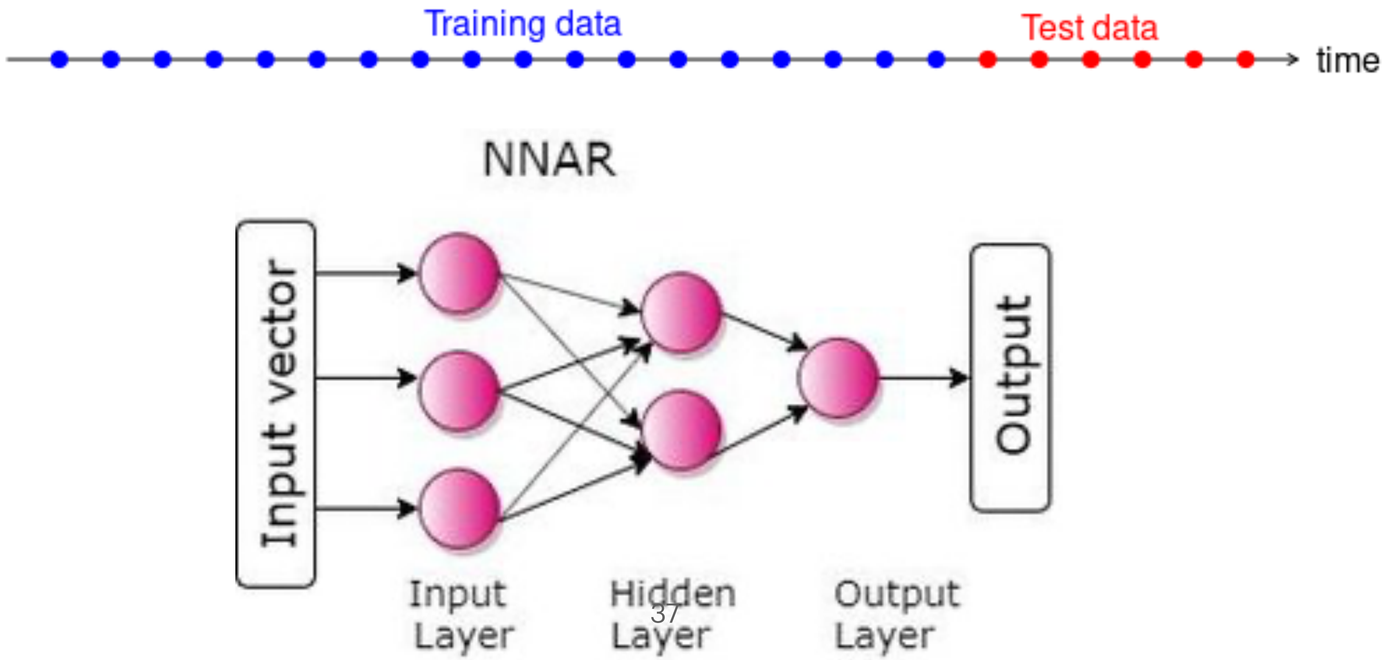


Models for forecasting

- There are many models with nice statistical properties:
  - AR
  - ARMA
  - ARIMA
  - SARIMA
  - SARIMAX
  - VARIMA
  - …
- But of course, we are NOT going to use them,
  - because we like **Neural Networks**!!!

Neural network autoregression, NNAR

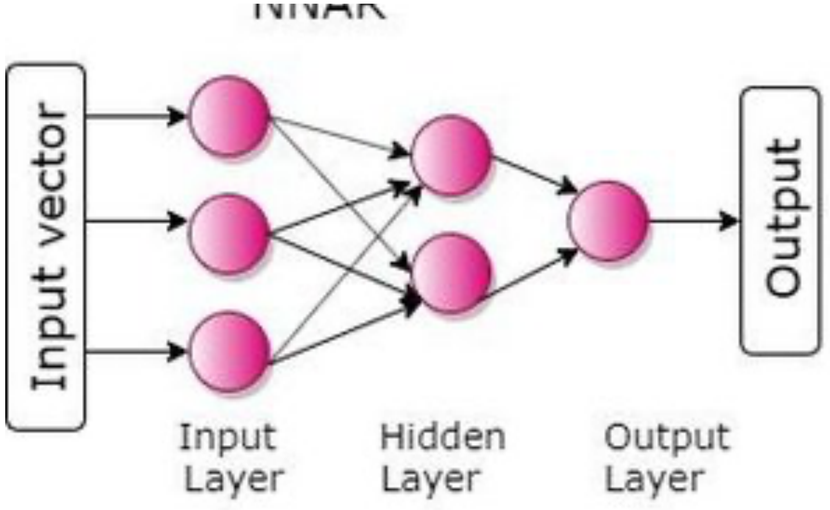
- With time series data, lagged values of the time series can be used as inputs to a neural network
- When it comes to the forecasting, the network is applied iteratively.
- For forecasting one step ahead, we simply use the available historical inputs



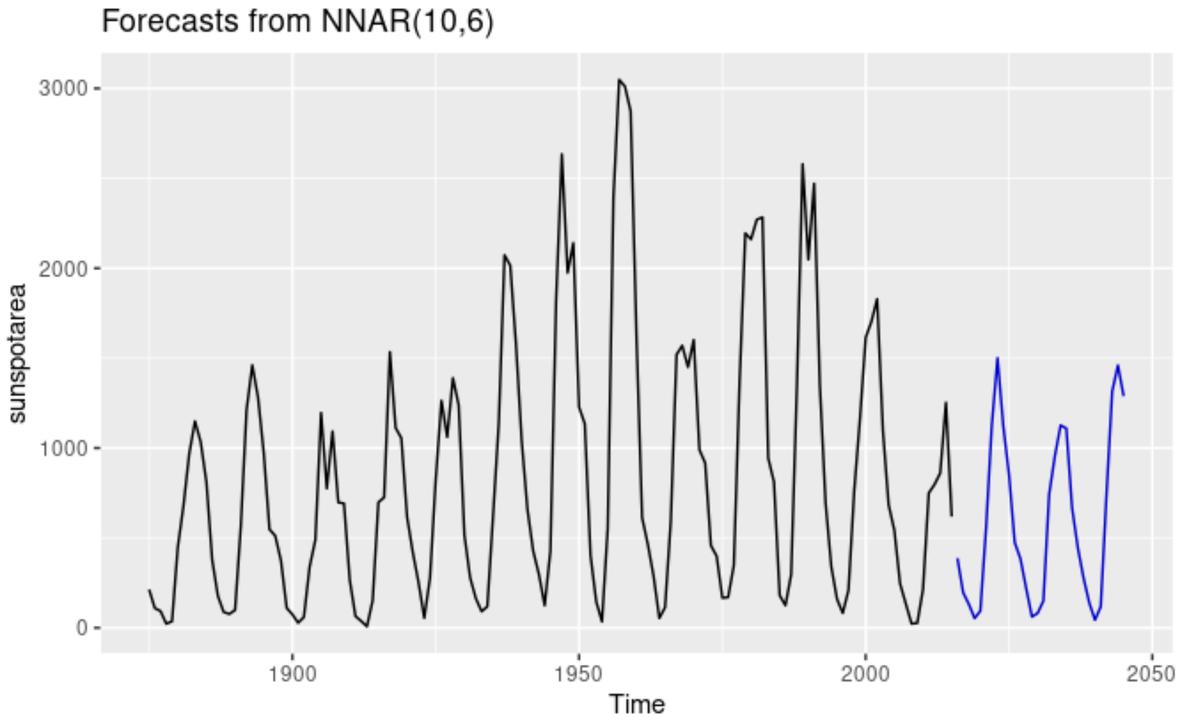
Neural network autoregression, NNAR

- The neural network fitted to the data can be written as

- $y_t = f(\boldsymbol{y}_{t-1}) + \varepsilon_t$

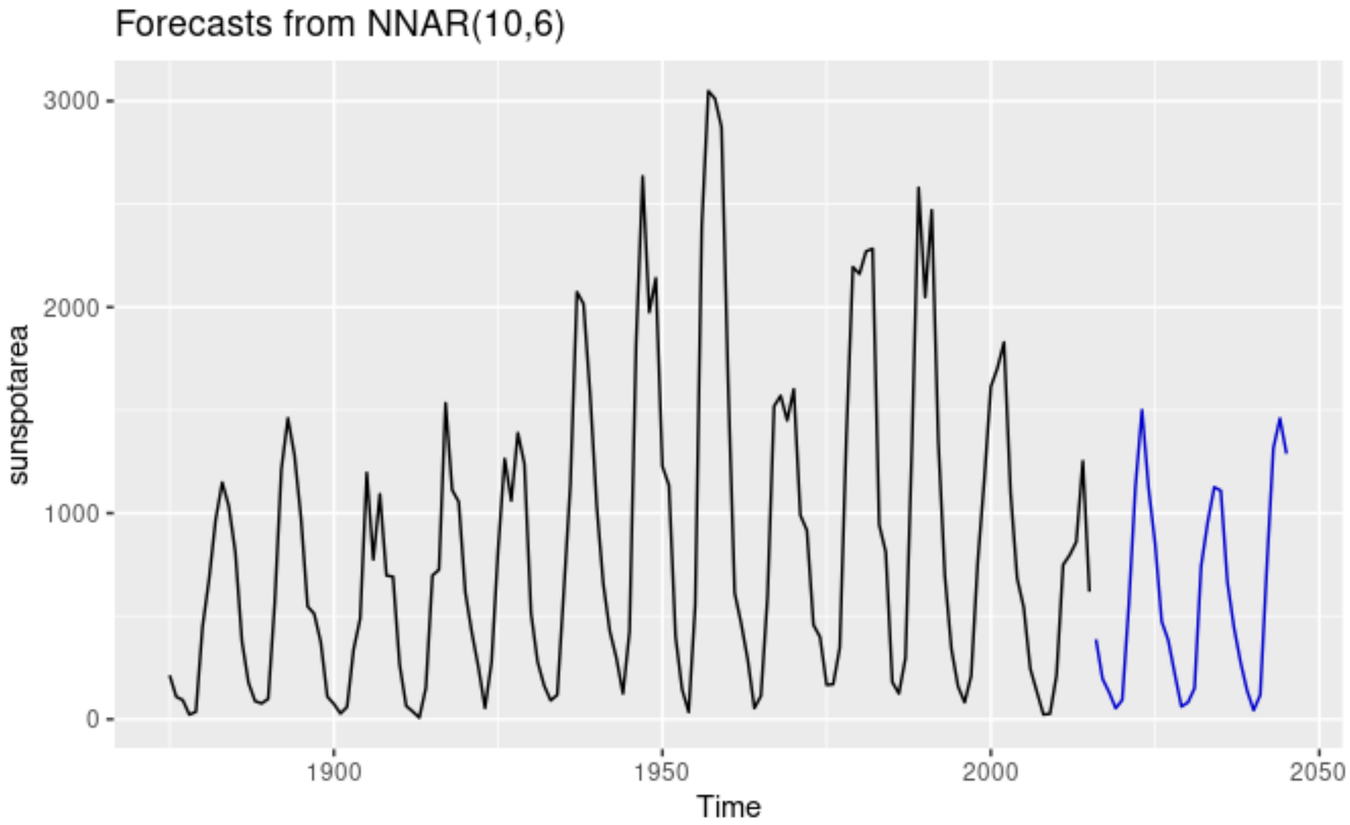


Forecasting



Discussion: How to assess the error of the prediction?

- We would like to have something like this
  - $y_{T+h|T} \pm 1.96\hat{\sigma}_h$
  - but we cannot assume normal distribution of the residuals,



Prediction Intervals

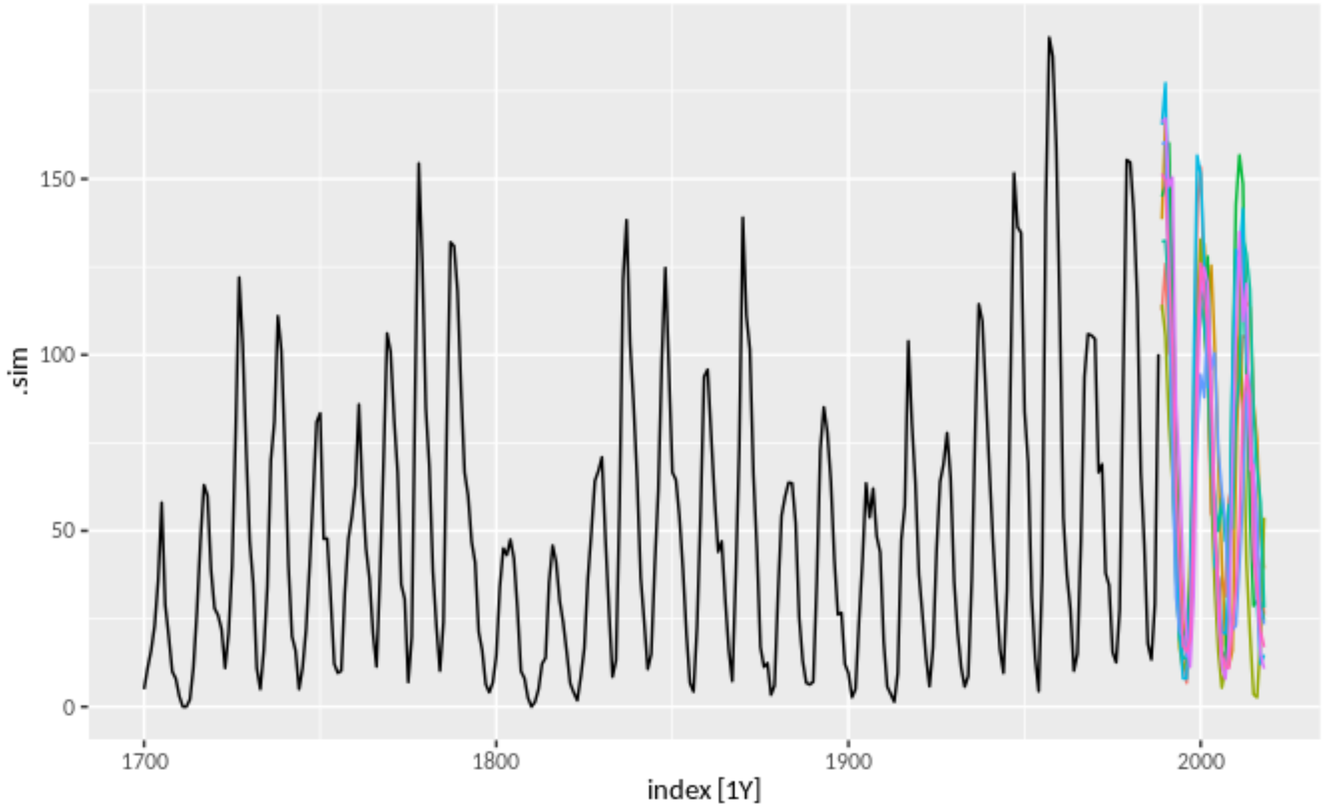
- Neural networks are not based on a well-defined stochastic model, and so it is not straightforward to derive prediction intervals for the resultant forecasts.
- However, we can still compute prediction intervals using simulation where future sample paths are generated using bootstrapped residuals

  - $y_t = f(\mathbf{y_{t-1}}) + \varepsilon_t$
  - $\mathbf{y_{t-1}}$ is a vector containing lagged values of the series

- How can we generate ε_t from historical data using bootstrap?

Example

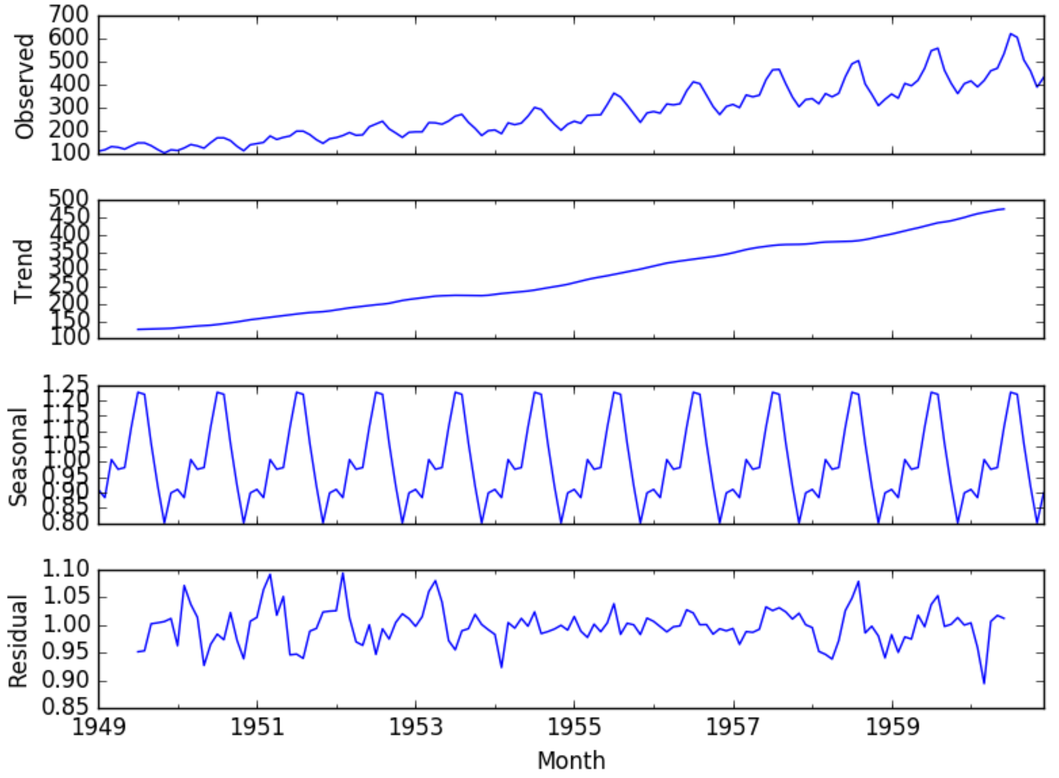
- Here is a simulation of 9 possible future sample paths
- Each sample path covers the next 30 years after the observed data.



Bootstrapped residuals

- First, the time series is transformed, and then decomposed into 3 components
  - trend,
  - seasonal and
  - remainder

- Then we obtain shuffled versions of the remainder component to get bootstrapped remainder series.

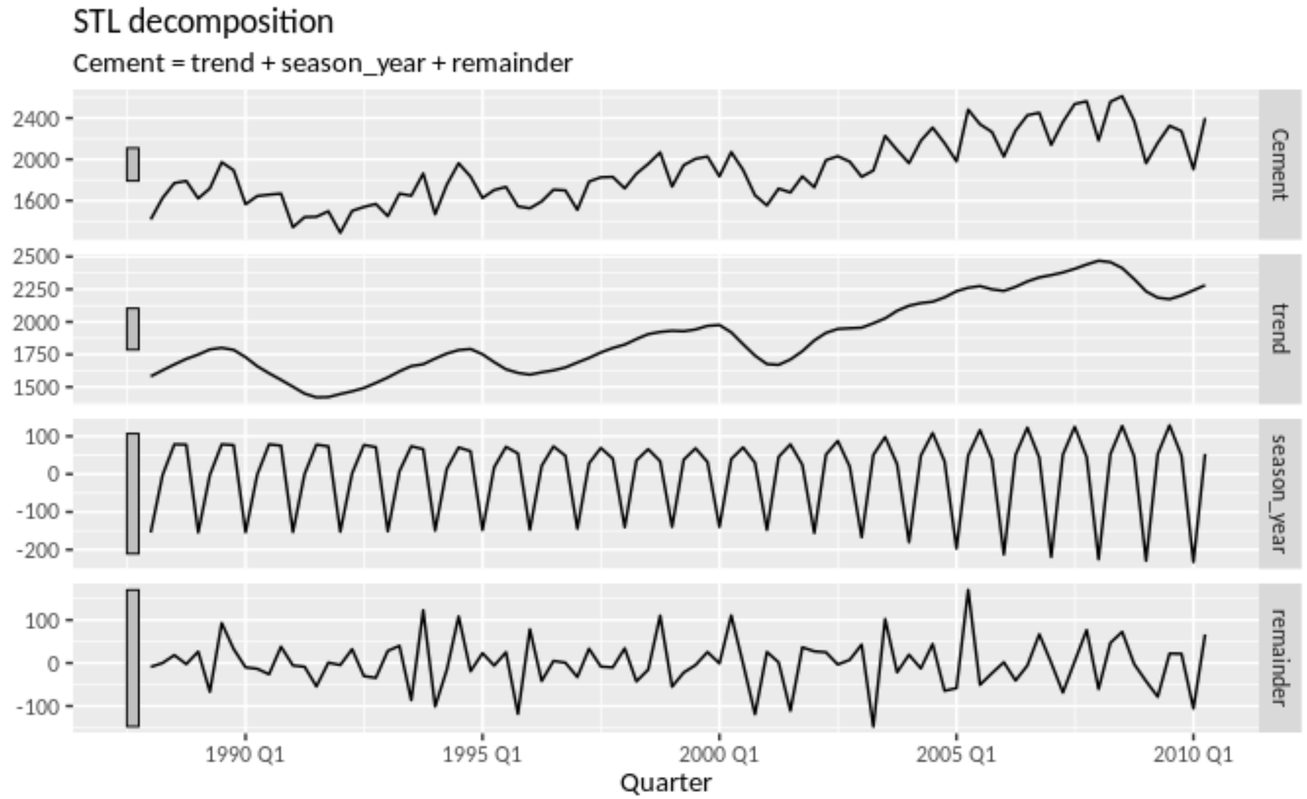


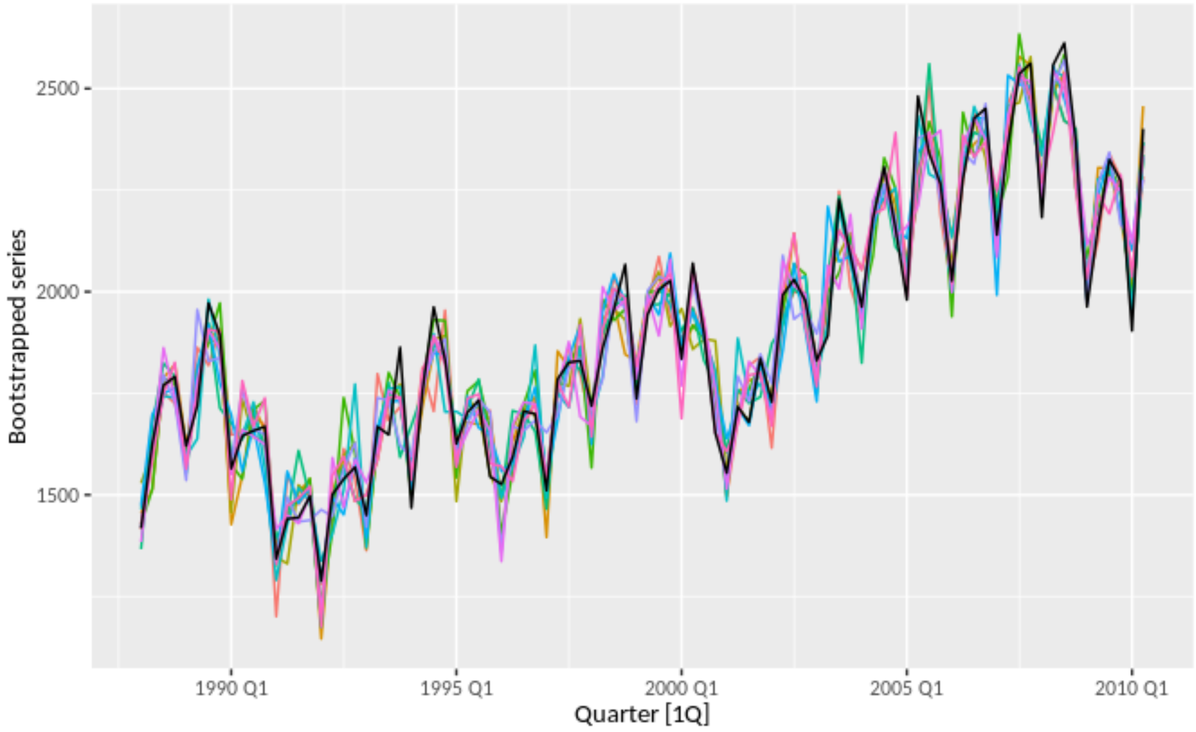
Bootstrapped residuals

- We use a “blocked bootstrap,” where contiguous sections of the remainder time series are selected at random and joined together.

- These bootstrapped remainder series are added to the trend and seasonal components, and the transformation is reversed to give variations on the original time series.

Example

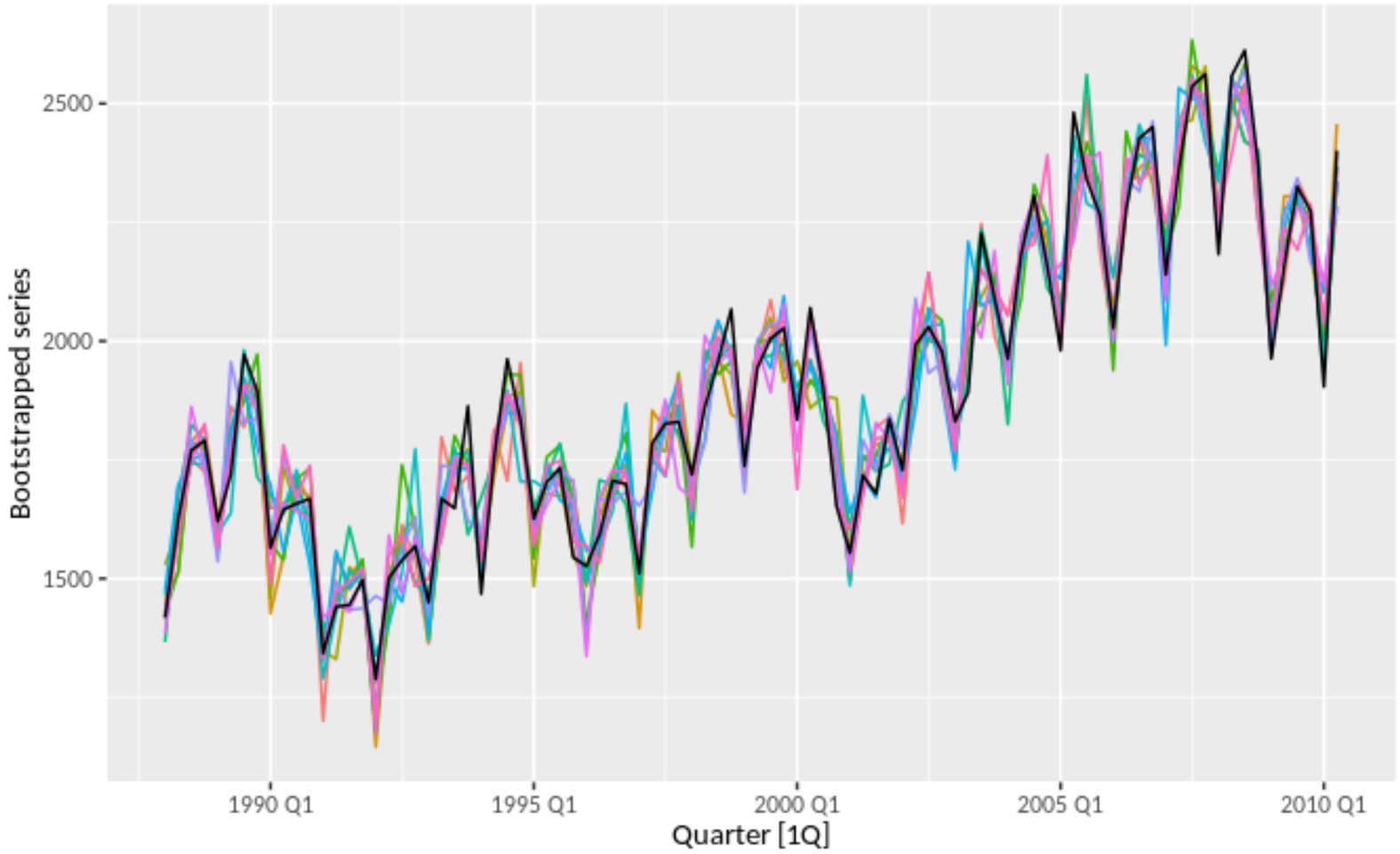




https://otexts.com/fpp3/bootstrap.html

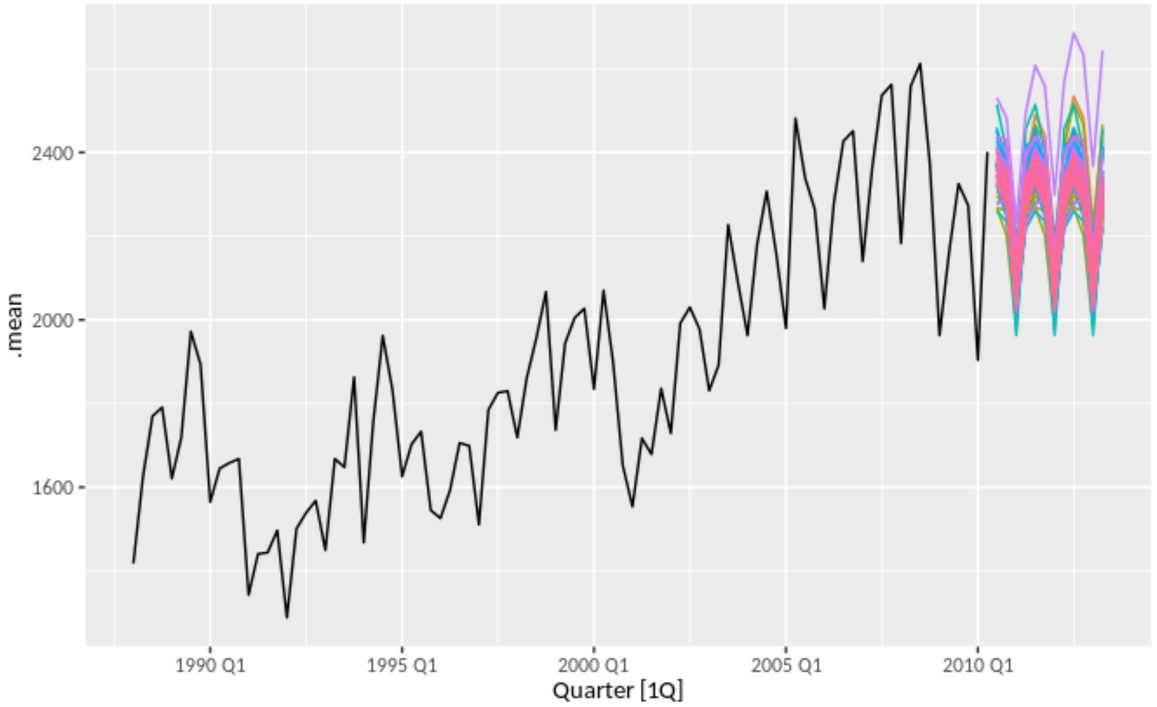
How to use the series / bootstrapped trajectories?

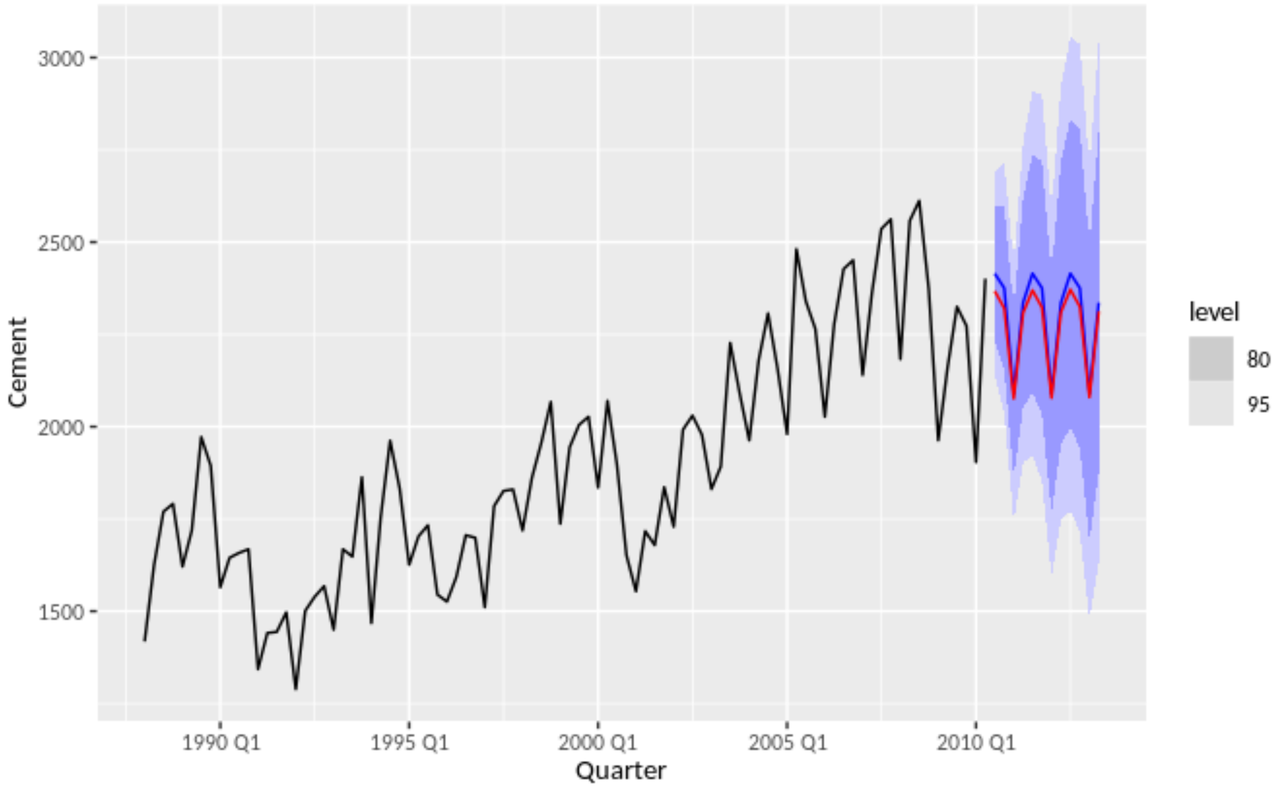
- So, we have several trajectories derived through bootstrap!



Bagging (“bootstrap aggregating”) !

- Predict for each trajectory and aggregate!
- It is shown that bagging gives better forecasts than classic methods such as ETS





Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation. International Journal of Forecasting, 32(2), 303–312. https://doi.org/10.1016/j.ijforecast.2015.07.002

Summary

References

- https://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474007#.U2o7MVdMzTo
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1-26.
- Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap, Chapman & Hall, New York.
- Michael Eichler's Statistics 24600 Course. Handouts. 2004. http: //galton.uchicago.edu/~eichler/stat24600/Handouts/bootstrap.pdf
- Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statist. Sci. 1 (1986), no. 1, 54–75. doi:10.1214/ss/1177013815. http://projecteuclid.org/euclid.ss/1177013815.
- http://notstatschat.tumblr.com/post/156650638586/when-the-bootstrap-doesnt-work
- https://github.com/rasbt/data-science-tutorial/blob/master/code/bootstrapping.ipynb