Paleoclimate
Time Series Analysis

Manfred Mudelsee
Climate Risk Analysis, Germany, www.climate-risk-analysis.com
Alfred Wegener Institute, Germany

TIES, Hyderabad, 5 January 2012

 Courtesy Dominik Fleitmann, Oeschger Centre, Switzerland
Part 1: Paleoclimate
Part 2: Time Series Analysis
Part 3: Examples

Mathematical formulas: no sweat!
Physics, Geology: no sweat!
Part 1: Paleoclimate

Climate

Part 1: Paleoclimate

Part 1: Paleoclimate

Paleoclimate archive

- 5278 ± 77 a
- 7525 ± 55 a
- 10062 ± 83 a
- 3010 ± 70 a
- -1.24 ± 0.08 ‰
- -2.23 ± 0.08 ‰
Figure 1.7. Speleothem data: oxygen isotope record from stalagmite Q5 from southern Oman over the past 10,300 years. Along the growth axis of the nearly 1 m long sta-
Part 1: Paleoclimate

Spacing and uncertain timescale

Figure 1.15. Spacing of selected climatic timeseries. 

- **a**: ODP846 δ¹⁸O;  
- **b**: Vostok CO₂;  
- **c**: Vostok δD;  
- **d**: NGRIP SO⁴;  
- **e**: Q5 δ¹⁸O. Data are given in Figs. 1.2, 1.3, 1.4 and 1.7.

In **d**, \(d(i)\) is shown for \(D(i) = 0.5\) cm data; the time series with \(t(i) = 1\) a (Fig. 1.4) is obtained from the 0.5-cm data using “downsampling.” The ice core data (**b–d**) reflect to some degree the effects of ice compaction, that means, \(d(i)\) increases with \(t(i)\). The Q5 speleothem spacing time series (**e**) suggests visually a strong negative correlation with the speleothem δ¹⁸O series (Fig. 1.7). This is explained as follows. Low (high) δ¹⁸O means strong (weak) Indian monsoonal rainfall, this in turn faster (slower) movement of the rainwater through the soil, weaker (stronger) uptake of soil-CO₂, lower (higher) pH of the water, reduced (enhanced) solution of soil-carbonate, less (more) material for calcite precipitation, small (large) annual stalagmite layers and, finally, a higher (lower) temporal spacing because the depth spacing is nearly constant (Fig. 1.7). Note that at places with other soil properties, the relation δ¹⁸O–spacing may be different (Burns et al. 2002). The values of the average spacing, \(\bar{d}\), and the coefficient of variation of spacing, CV\(\bar{d}\), which is defined as the standard deviation of the spacing divided by \(\bar{d}\), are as follows.

- **a**: \(\bar{d} = 2.40\) a, CV\(\bar{d} = 0.41\);  
- **b**: \(\bar{d} = 1.46\) a, CV\(\bar{d} = 0.82\);  
- **c**: \(\bar{d} = 0.13\) a, CV\(\bar{d} = 0.85\);  
- **d**: \(\bar{d} = 0.32\) a, CV\(\bar{d} = 0.47\);  
- **e**: \(\bar{d} = 5.62\) a, CV\(\bar{d} = 0.49\).
Part 1: Paleoclimate

Noise and statistical distribution

Histogram: inference of PDF

Residuals, “detrended version” of the $x(i)$
Part 1:  Paleoclimate

Persistence

Lag-1 scatterplot $x(i-1)$ vs $x(i)$: inference about persistence

Residuals, “detrended version” of the $x(i)$

Book, Fig. 1.12 (j, stalagmite Q5 $\delta^{18}O$)
Part 1: Paleoclimate

Climate equation

Climate equation, discrete time

\[ X(i) = X_{\text{trend}}(i) + X_{\text{out}}(i) + S(i) \cdot X_{\text{noise}}(i), \quad (1.2) \]

Climate equation, discrete time

\( T(i) \), time points

\( X(i) \equiv X(T(i)), \) etc.
Climate is a paradigm of a complex system. Analysing climate data is an exciting challenge, which is increased by non-normal distributional shape, serial dependence, uneven spacing and timescale uncertainties. This book presents bootstrap resampling as a computing-intensive method able to meet the challenge. It shows the bootstrap to perform reliably in the most important statistical estimation techniques: regression, spectral analysis, extreme values and correlation.

This book is written for climatologists and applied statisticians. It explains step by step the bootstrap algorithms (including novel adaptions) and methods for confidence interval construction. It tests the accuracy of the algorithms by means of Monte Carlo experiments. It analyses a large array of climate time series, giving a detailed account on the data and the associated climatological questions.
Part 2: Time Series Analysis

Estimation

\[ X(i) = X_{\text{trend}}(i) + X_{\text{out}}(i) + S(i) \cdot X_{\text{noise}}(i), \quad (1.2) \]

Climate equation, stochastic process \( X(i) \)

Time series \( \{t(i), x(i)\}_{i=1}^{n} \)

Time series analysis uses sample to learn about process: trend parameters, probability of extremes, cycles, etc.

Note relation between data and question(s).
Part 2: Time Series Analysis

First-order autoregressive model, uneven spacing

\[ X_{\text{noise}}(1) = \mathcal{E}_{N(0, 1)}(1), \]
\[ X_{\text{noise}}(i) = \exp\left\{ - \frac{T(i) - T(i - 1)}{\tau} \right\} \cdot X_{\text{noise}}(i - 1) + \mathcal{E}_{N(0, 1 - \exp\{-2\frac{T(i) - T(i - 1)}{\tau}\})}(i), \quad i = 2, \ldots, n. \] (2.9)

AR(1) process, uneven spacing

\[ S(\tilde{\tau}) = \sum_{i=2}^{n} \left[ x_{\text{noise}}(i) - \exp\left\{ - \frac{t(i) - t(i - 1)}{\tilde{\tau}} \right\} \cdot x_{\text{noise}}(i - 1) \right]^2, \]

Least-squares estimation

Residuals \( x_{\text{noise}}(i) \), “detrended version” of the \( x(i) \)
Part 2: Time Series Analysis

Bootstrap principle

\[
\{t(i), x(i)\}_{i=1}^{n} \quad \hat{\theta} \\
\{t^*(i), x^*(i)\}_{i=1}^{n} \quad \hat{\theta}^*_{1} \\
\{t^*(2)(i), x^*(2)(i)\}_{i=1}^{n} \quad \hat{\theta}^*_{2} \\
\{t^*(B)(i), x^*(B)(i)\}_{i=1}^{n} \quad \hat{\theta}^*_{B} \\
\text{Sample, estimate} \\
\text{Confidence interval} \\
\text{Replications} \\
\text{Resamples} \\
\text{Sample, estimate}
\]

Book, Fig. 3.3
Part 2: Time Series Analysis

Bootstrap principle

Bootstrap

Resampling

Moving block bootstrap (MBB)
Autoregressive bootstrap (ARB)

Other

Confidence interval construction

Normal
Student’s $t$
Percentile
BCa
Bootstrap resampling

**Algorithm 3.1. Moving block bootstrap algorithm (MBB).**

Step 1: Data

\[ \{ t(i), x(i) \}_{i=1}^n \]

Step 2: Resampled times unchanged

\[ \{ t^*(i) \}_{i=1}^n = \{ t(i) \}_{i=1}^n \]

Step 3: Blocks

\[ j \) (see above) \]

\[ \{ x(i) \}_{j+l-1} \]

\[ i = j, j = 1, \ldots, n-l+1 \]

Step 4: Set counter \( c = 1 \)

Start resampling

Step 5: Draw random block \( j^* \) \( j^* \in \{ 1, \ldots, n-l+1 \} \)

Step 6: Insert block data

\[ \{ x^*(i) \}_{c+l-1} \]

\[ i = c = \{ x(i) \}_{j^*+l-1} \]

\[ i = j^* \]

If \( x^*(n) \) has been inserted Stop inserting and exit

Step 7: Increase counter \( c \to c+l \)

Step 8: Go to Step 5

End resampling

---

Moving block bootstrap (MBB)
Bootstrap resampling

Block length selector (MBB) after Carlstein (1986)

\[ l_{\text{opt}} = \text{NINT} \left\{ \left[ \frac{6^{1/2} \cdot \hat{a} / \left(1 - \hat{a}^2\right)}{1 - \hat{a}^2} \right]^{2/3} \cdot n^{1/3} \right\}, \quad (3.28) \]

Other block length selectors via
- persistence time
- autocorrelation function
### Part 2: Time Series Analysis

#### Bootstrap resampling

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data</td>
<td>( {t(i), x(i)}_{i=1}^{n} )</td>
</tr>
<tr>
<td>2</td>
<td>Resampled times</td>
<td>( {t^*(i)}<em>{i=1}^{n} = {t(i)}</em>{i=1}^{n} ) unchanged</td>
</tr>
<tr>
<td>3</td>
<td>Residuals (Eq. 3.29)</td>
<td>( r(i) = \left[ x(i) - \hat{x}<em>{trend}(i) - \hat{x}</em>{out}(i) \right] / \hat{S}(i) )</td>
</tr>
<tr>
<td>4</td>
<td>Apply MBB</td>
<td>(Algorithm 3.1)</td>
</tr>
<tr>
<td></td>
<td>to residuals</td>
<td>( {r(i)}_{i=1}^{n} )</td>
</tr>
<tr>
<td>5</td>
<td>Resampled residuals</td>
<td>( {r^*(i)}_{i=1}^{n} )</td>
</tr>
<tr>
<td>6</td>
<td>Use resampled residuals</td>
<td>( x^<em>(i) = \hat{x}<em>{trend}(i) + \hat{x}</em>{out}(i) + \hat{S}(i) \cdot r^</em>(i) )</td>
</tr>
</tbody>
</table>

**Algorithm 3.3.** MBB for realistic climate processes, which comprise trend, outlier and variability components.
Part 2: Time Series Analysis

Bootstrap confidence intervals

The surrogate data approach (Algorithm 3.6), related to ARB, is a simulation rather than a resampling method. No residuals are drawn as in the ARB. Instead, climate equation residuals $\{r^*_i\}_{i=1}^n$ are obtained by numerical simulation (Step 8) from the persistence model with estimated (and bias-corrected) parameters. Because also the distributional shape is specified, the surrogate data approach is bounded stronger by parametric restrictions than the ARB. Therein lies its danger: it is more prone than the ARB to systematic errors from violated assumptions.

3.4 Bootstrap confidence intervals

Estimation of $\theta$ is repeated for the resamples, $\{\tilde{t}^*_b(i), \tilde{x}^*_b(i)\}_{i=1}^B$. This yields the bootstrap replications, $\{\hat{\theta}^*_b\}_{b=1}^B$. The replications are used to construct equi-tailed $(1 - 2\alpha)$ confidence intervals, $\text{CI}_{\hat{\theta}^*, 1 - 2\alpha}$, see Fig. 3.3.

Two approaches, standard error based and percentile based, dominate theory and practice of bootstrap CI construction. The estimated bootstrap standard error is the sample standard error of the replications, $\hat{\text{se}}_{\hat{\theta}^*} = \left\{ \sum_{b=1}^{B} \left[ \hat{\theta}^*_b - \langle \hat{\theta}^*_b \rangle \right]^2 \right\}^{1/2} / (B - 1)$, (3.30)

where $\langle \hat{\theta}^*_b \rangle = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*_b / B$.

Bootstrap standard error
Part 2: Time Series Analysis

Bootstrap confidence intervals

$$\text{CI}_{\hat{\theta}, 1-2\alpha} = [\hat{\theta}^*(\alpha); \hat{\theta}^*(1 - \alpha)]$$, \hspace{1cm} (3.33)

Percentile confidence interval

$\alpha$ small $\Rightarrow B$ large

typically

$\alpha = 0.025$ (95% CI) or $\alpha = 0.05$ (90% CI), $B \approx 2000$
Part 3: Examples

Spectrum estimation, stalagmite Q5, Holocene

Oxygen isotope record (Q5)
Monsoon rainfall proxy
Trend: ramp + sinusoid
Interval [2.7 ka; 8.0 ka]
$\bar{d} = 5.4 \text{ a}, n = 973$

Welch I taper, 6 segments, 50% overlap

Estimated spectrum
Lomb–Scargle
Oversampling factor 64
Highest $f = 1.0 f_{Ny}$
Part 3: Examples

Spectrum estimation, stalagmite Q5, Holocene

**Spectrum**

6-dB bandwidth

\[ B_s \approx 0.001 \text{ a}^{-1} \]

**Peaks**

\[ T_{\text{period}} = 10.9 \text{ a} \quad (I) \]
\[ T_{\text{period}} = 107 \text{ a} \quad (II) \]
\[ T_{\text{period}} = 137 \text{ a} \quad (III) \]
\[ T_{\text{period}} = 221 \text{ a} \quad (IV) \]
\[ T_{\text{period}} = 963 \text{ a} \quad (V) \]

**AR(1) alternatives**

90%, 95%, 99%, 99.9%

---

Book, Fig. 5.8; 6-dB bandwidth: decay on flanks of a peak to \( 10^{-6/\nu} \approx 0.251 \times \text{peak maximum} \)
Part 3: Examples

Spectrum estimation, stalagmite Q5, Holocene

Aliasing
Likely not a problem here

Sampling length 3.9 years
Monsoon rainfall mostly during JAS, not full year
Statistical test: AR(1) + annual cycle \( \Rightarrow \) no peaks \( (T_{\text{period}}: I-V) \)
Timescale error effects?
Test using $t(i)^*$ simulation
Part 3: Examples

Spectrum estimation, stalagmite Q5, Holocene

Step 1 Time series \{t(i), x(i)\}_{i=1}^n
Step 2 Bias-corrected Lomb–Scargle spectrum (Algorithm 5.3) \(\hat{h}'(f_j)\)
Step 3 Estimated, bias-corrected persistence time \(\hat{\tau}'\)
Step 4 Determine area under spectrum within \([0; (2\bar{d})^{-1}]\)
Step 5 Generate AR(1) data (Eq. 2.9) \{t(i), x^*(i)\}_{i=1}^n
Step 6 Use timescale model to resample times \{t^*(i)\}_{i=1}^n
Step 7 Bias-corrected Lomb–Scargle spectrum estimate \(\hat{h}'^{*b}(f_j)\)
for \{t^*(i), x^*(i)\}_{i=1}^n
(b, counter),
scaled to area \(A_{\hat{h}}\)
Step 8 Go to Step 5 until \(b = B\) replications exist
Step 9 Test at each \(f_j\) whether \(\hat{h}'(f_j)\) exceeds a pre-defined upper percentile of \(\{\hat{h}'^{*b}(f_j)\}_{b=1}^B\)
Part 3: Examples

Spectrum estimation, stalagmite Q5, Holocene

Timescale error effects?
Test using $t(i)^*$ simulation
Peak I likely affected, possibly also peaks II to IV
Peak V ($T_{\text{period}} = 963$ a) robust
Part 3: Examples

Spectrum estimation, stalagmite Q5, Holocene

Oxygen isotope record (Q5)
Monsoon rainfall proxy

Use spectrum to filter out frequency ranges of interest.
Study physics of Sun–monsoon system.
Take into account all relevant records (multiple test).
Part 3: Examples

Correlation estimation, Agulhas current, Pleistocene

Recent:
- Temperature
- Agulhas flow

Paleo:
- Temperature
- (via sediment core)

Inference target:
- Paleo Agulhas flow

Beal et al. (2011) *Nature* 472:429
Recent:
Temperature ($X$)
Agulhas flow ($Y$)

Paleo:
Temperature
(via sediment core)

Inference target:
Paleo Agulhas flow

Method:
Look for $\max(r_{XY})$
with bootstrap CI
Correlation estimation, Agulhas current, Pleistocene

\[ X(i) = X_{\text{trend}}(i) + X_{\text{out}}(i) + S(i) \cdot X_{\text{noise}}(i) \]
\[ = \mu_X + S_X \cdot X_{\text{noise}}(i). \]

Climate equation, discrete time

\[ Y(i) = \mu_Y + S_Y \cdot Y_{\text{noise}}(i). \]

Climate equation, discrete time

Bivariate setting, equal time points

Absent trends, absent outliers, constant variability

Time series

\[ \{t(i), x(i)\}_{i=1}^{n} \]

Time series

\[ \{t(i), y(i)\}_{i=1}^{n} \]
Part 3: Examples

Correlation estimation, Agulhas current, Pleistocene

The correlation coefficient is then defined as

\[ r_{XY} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X(i) - \bar{X}}{S_{n,X}} \right) \cdot \left( \frac{Y(i) - \bar{Y}}{S_{n,Y}} \right), \]  

(7.5)

Pearson’s correlation coefficient \((r_{XY}\) estimates \(\rho_{XY}\))

\[-1 \leq r_{XY} \leq 1\]

\[ \bar{X} = \sum_{i=1}^{n} \frac{X(i)}{n} \quad \bar{Y} = \sum_{i=1}^{n} \frac{Y(i)}{n} \]

Sample means

\[ S_{n,X} = \left\{ \sum_{i=1}^{n} \frac{[X(i) - \bar{X}]^2}{n} \right\}^{1/2} \]

\[ S_{n,Y} = \left\{ \sum_{i=1}^{n} \frac{[Y(i) - \bar{Y}]^2}{n} \right\}^{1/2} \]

Sample standard deviations (denominator \(n\))
Correlation estimation, Agulhas current, Pleistocene

Pairwise-moving block bootstrap (pairwise-MBB)

\[
l_{opt} = \text{NINT} \left\{ \left[ 6^{1/2} \cdot \hat{a}'_{XY} / \left( 1 - \hat{a}'_{XY}^2 \right) \right]^{2/3} \cdot n^{1/3} \right\}.
\]

Block length selector

\[
\hat{a}'_{XY} = \left[ \hat{a}'_X \cdot \hat{a}'_Y \right]^{1/2}
\]
Table 7.4. Monte Carlo experiment, Pearson’s correlation coefficient with Fisher’s $z$-transformation for bivariate lognormal AR(1) processes. The number of Monte Carlo simulations and the properties of \{T(i), X(i), Y(i)\}$_{i=1}^n$ are identical to those in the first experiment (Table 7.1), with the exception that $\rho_X$ is here given via Eq. (7.24). The construction of CIs followed Algorithms 7.1 and 7.2.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\gamma_{r_{XY}}^a$</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>True correlation, $\rho_{XY}$</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>CI type</td>
<td>Bootstrap</td>
<td>Classical</td>
</tr>
<tr>
<td></td>
<td>Student’s $t$</td>
<td>BCa</td>
</tr>
<tr>
<td>10</td>
<td>0.820</td>
<td>0.701</td>
</tr>
<tr>
<td>20</td>
<td>0.876</td>
<td>0.808</td>
</tr>
<tr>
<td>50</td>
<td>0.932</td>
<td>0.875</td>
</tr>
<tr>
<td>100</td>
<td>0.939</td>
<td>0.866</td>
</tr>
<tr>
<td>200</td>
<td>0.941</td>
<td>0.879</td>
</tr>
<tr>
<td>500</td>
<td>0.907</td>
<td>0.876</td>
</tr>
<tr>
<td>1000</td>
<td>0.911</td>
<td>0.885</td>
</tr>
</tbody>
</table>

Alert! MC runs suffer from programming bug (which affects small $r_{XY}$).
### Table 7.6. Monte Carlo experiment, Pearson’s and Spearman’s correlation coefficients with Fisher’s $z$-transformation for bivariate lognormal AR(1) processes: calibrated CI coverage performance. The number of Monte Carlo simulations and the properties of $\{T(i), X(i), Y(i)\}_{i=1}^{n}$ are identical to those in the first experiment (Table 7.1), with $\rho_{\varepsilon}$ given by Eq. (7.24) and Table 7.8, respectively. Calibrated Student’s $t$ CIs were constructed after Eq. (3.47) using two loops of pairwise-MBB resampling with block length selection after Eqs. (7.31) and (7.32). The first loop (bootstrap of samples) used $B = 2000$ resamplings, the second loop (bootstrap of resamples) used 1000 resamplings. In the second loop, the block length was not re-estimated but overtaken from the first loop. The spacing of the $\lambda$ values for the calibration (Eq. 3.45) is 0.001.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\gamma_{r_{XY}}^a$</th>
<th>$\gamma_{r_{S}}^a$</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True correlation, $\rho_{XY}$</strong></td>
<td><strong>True rank correlation, $\rho_{S}$</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.8</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>0.917</td>
<td>0.836</td>
<td>0.950</td>
</tr>
<tr>
<td>20</td>
<td>0.959</td>
<td>0.937</td>
<td>0.950</td>
</tr>
<tr>
<td>50</td>
<td>0.964</td>
<td>0.947</td>
<td>0.950</td>
</tr>
<tr>
<td>100</td>
<td>0.969</td>
<td>0.947</td>
<td>0.950</td>
</tr>
<tr>
<td>200</td>
<td>0.966</td>
<td>0.946</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Alert! MC runs suffer from programming bug (which affects small $r_{XY}$).
Monte Carlo experiments, in terms of CI coverage accuracy

1. Classical CIs fail completely for non-Gaussian shapes.

2. Usage of Spearman’s $r_s$ instead of Pearson’s $r_{XY}$ advised:
   - If distributional shapes Gaussian, then both perform similar
   - else if shapes are non-Gaussian (e.g., skewed), then Pearson’s $r_{XY}$ performs badly.

3. Calibration (expensive) increases accuracy (small $n$) strongly.
Paleoclimate time series analysis is exciting!
Conclusion

Paleoclimate time series analysis is exciting!

Thanks!

Postdoc job: www.climate-risk-analysis.com
Part 1: Paleoclimate

Paleoclimate time series

Appendix

Book, Figs. 4.18 (modified); Fleitmann et al. (2004: Fig. 1a) Quaternary Science Reviews 23:935
Why not interpolate?

Monte Carlo experiment

\( n = 50, \) uneven spacing \((\sigma_d)\),
\( \tau = -1/\ln(0.7) \approx 2.804, \)
\( \bar{a} = \exp(-\bar{d}/\tau) = 0.7 \)

Equivalent autocorrelation coefficient

\( \hat{a} = \exp(-\bar{d}/\hat{\tau}) \)

Estimated equivalent autocorrelation coefficient averaged over \( n_{\text{sim}} = 10,000 \) simulations
Part 2: Time Series Analysis

Climate theory

Climate, $X(T)$

\[
\frac{dX(T)}{dT} = F(X(T), Y(T)), \quad \text{timescale } \tau_X,
\]

Weather, $Y(T)$

\[
\frac{dY(T)}{dT} = G(X(T), Y(T)), \quad \text{timescale } \tau_Y,
\]

Weather components

“leave a trace” in climate.

\[
\frac{dX(T)}{dT} \approx F(X(0), Y(T)), \quad \text{timescale } \tau_X,
\]

\[
= W(T),
\]

\[
X(T + 1) = X(T) + \mathcal{E}_{N(0, \sigma^2)}(T).
\]

Climate does not run away: negative feedback ($\sim X$)

\[
X(T + 1) = a \cdot X(T) + \mathcal{E}_{N(0, \sigma^2)}(T),
\]

$W(T)$, Wiener process (“continuous-time noise increment”)

Hasselmann (1976) *Tellus* 28:473; Book, Eqs. (2.25), (2.26), (2.28), (2.29), (2.30), Section 2.5.1
Support of simple AR(1) climate noise model

1. Weather components (Hasselmann 1976)

2. Climate equation

\[ X(i) = X_{\text{trend}}(i) + X_{\text{out}}(i) + S(i) \cdot X_{\text{noise}}(i), \]

3. We have also trend, outliers/extremes and variability to describe climate.
Part 2: Time Series Analysis

Appendix

Bootstrap resampling

Autoregressive bootstrap (ARB)
- Take noise residuals $r(i)$
- Fit AR(1) model
- Take white-noise residuals $\epsilon(i) = [r(i) - \hat{a} \cdot r(i - 1)]$
- Draw randomly from $\epsilon(i)$ with replacement (ordinary bootstrap)
- Plug in resampled $\epsilon(i)$
- Plug in resampled $r(i)$
Bootstrap confidence intervals

Estimation of $\theta$ is repeated for the resamples, $\{t^*_b(i), x^*_b(i)\}_{i=1}^n$, $b = 1, \ldots, B$. This yields the bootstrap replications, $\{\hat{\theta}^*_b\}_{b=1}^B$. The replications are used to construct equi-tailed $(1 - \frac{2}{B})$ confidence intervals, $\text{CI} \hat{\theta}, 1 - \frac{2}{B}$, see Fig. 3.3.

Two approaches, standard error based and percentile based, dominate theory and practice of bootstrap CI construction. The estimated bootstrap standard error is the sample standard error of the replications, $\widehat{\text{se}}\hat{\theta}^* = \left\{ \sum_{b=1}^B \left[ \hat{\theta}^*_b - \langle \hat{\theta}^*_b \rangle \right]^2 / (B - 1) \right\}^{1/2}$, (3.30)

Bootstrap standard error

\[ \widehat{\text{se}}\hat{\theta}^* = \left\{ \sum_{b=1}^B \left[ \hat{\theta}^*_b - \langle \hat{\theta}^*_b \rangle \right]^2 / (B - 1) \right\}^{1/2}, \]
Part 2: Time Series Analysis

Bootstrap confidence intervals

3.4 Bootstrap confidence intervals

Step 1 Data \{t(i), x(i)\}$_{i=1}^n$

Step 2 Resampled times \{t^*(i)\}$_{i=1}^n = \{t(i)\}$_{i=1}^n$ unchanged

Step 3 Estimated trend, \{\hat{x}_{\text{trend}}(i)\}$_{i=1}^n$, outliers, \{\hat{x}_{\text{out}}(i)\}$_{i=1}^n$, variability \hat{S}(i)\$_{i=1}^n$

Step 4 Climate equation \{r(i)\}$_{i=1}^n$ residuals (Eq. 3.29)

Step 5 Assume \{r(i)\}$_{i=1}^n$ to come from specific model (shape, persistence)

Step 6 Estimate model parameters

Step 7 Bias correction

Step 8 Simulate climate equation residuals \{r^*(i)\}$_{i=1}^n$ from estimated model

Step 9 Simulated data \(x^*(i) = \hat{x}_{\text{trend}}(i) + \hat{x}_{\text{out}}(i) + \hat{S}(i) \cdot r^*(i)\), i = 1, ..., n

Algorithm 3.6. Surrogate data approach.

3.4.1 Normal confidence interval

The bootstrap normal confidence interval, already given in Fig. 3.3, is

\[
\text{CI}_{\hat{\theta}, 1-2\alpha} = \left[ \hat{\theta} + z(\alpha) \cdot \hat{\sigma}_{\hat{\theta}^*}; \hat{\theta} - z(\alpha) \cdot \hat{\sigma}_{\hat{\theta}^*} \right], \quad (3.31)
\]

Normal confidence interval

\(z(\alpha)\) Percentage point of normal distribution

Example

\(z(1 - 0.025) \approx 1.959964\)

“±2-sigma interval is 95% CI.”
Bootstrap confidence intervals

Student’s $t$ confidence interval

$$\text{CI}_{\hat{\theta}, 1-2\alpha} = \left[\hat{\theta} + t_{\nu}(\alpha) \cdot \hat{\text{se}}_{\hat{\theta}}; \hat{\theta} - t_{\nu}(\alpha) \cdot \hat{\text{se}}_{\hat{\theta}}\right], \quad (3.32)$$

$t_{\nu}(\alpha)$ Percentage point of Student’s $t$-distribution

- with $\nu$ degrees of freedom ($\nu = n – \text{number of parameters}$)
- Takes into account that not only standard error, but also fit value estimated.
- Negligible difference to normal CI if $\nu$ above $\sim 30$
Part 2: Time Series Analysis

Bootstrap confidence intervals

\[
\{ t^*(i), x^*(i) \}_{i=1}^{n}
\]

\[
\{ t^*(i), x^*(i) \}_{i=1}^{n}
\]

\[
\{ t^*(i), x^*(i) \}_{i=1}^{n}
\]

\[
\{ t(i), x(i) \}_{i=1}^{n}
\]

\[ \hat{\theta} \]

\[ \hat{\theta}_{1} \]

\[ \hat{\theta}_{2} \]

\[ \hat{\theta}^B \]

\[ \text{Confidence interval} \]

\[ \text{Replications} \]

\[ \text{Resamples} \]

\[ \text{Sample, estimate} \]

Book, Fig. 3.3
Bootstrap confidence intervals

\[ \text{CI}_{\hat{\theta}, 1-2\alpha} = \left[ \hat{\theta}^*(\alpha); \hat{\theta}^*(1 - \alpha) \right], \quad (3.33) \]

Percentile confidence interval

\( \alpha \) small \( \Rightarrow \) \( B \) large

- typically
  - \( \alpha = 0.025 \) (95% CI) or \( \alpha = 0.05 \) (90% CI), \( B \approx 2000 \)
Appendix

Part 2: Time Series Analysis

Bootstrap confidence intervals

Percentile confidence interval

Assume that our estimator underestimates, $E[\hat{\theta}] < \theta$. Then likely also the bootstrap replications are “shifted”.

There is an option for bias correction.

$$CI_{\hat{\theta}, 1-2\alpha} = [\hat{\theta}^*(\alpha); \hat{\theta}^*(1 - \alpha)]$$

(3.33)
Part 2: Time Series Analysis

Bootstrap confidence intervals

Construction of equi-tailed percentile CIs

1. \( \hat{\theta}, \{\hat{\theta}^*_b\}_{b=1}^B \)

2. Percentile CI

5%  90%  5%
Appendix

Part 2: Time Series Analysis

Bootstrap confidence intervals

Construction of equi-tailed percentile CIs

1. $\hat{\theta}, \{\hat{\theta}_b\}_{b=1}^B$

2. Percentile CI

3. $\hat{\theta}, \text{med}\{\hat{\theta}_b\}_{b=1}$

4. Bias correction

5% 90% 5%
Part 2: Time Series Analysis

Bootstrap confidence intervals

\[
\text{CI}_{\hat{\theta},1-2\alpha} = \left[ \hat{\theta}^*(\alpha 1); \hat{\theta}^*(\alpha 2) \right],
\]

(3.34)

Bias-corrected and accelerated (BCa) confidence interval

\[
\alpha 1 = F \left( \hat{z}_0 + \frac{\hat{z}_0 + z(\alpha)}{1 - \hat{a} [\hat{z}_0 + z(\alpha)]} \right)
\]

\[
\alpha 2 = F \left( \hat{z}_0 + \frac{\hat{z}_0 + z(1 - \alpha)}{1 - \hat{a} [\hat{z}_0 + z(1 - \alpha)]} \right)
\]

\[
\hat{z}_0 = F^{-1} \left( \frac{\# \left\{ \hat{\theta}^* b < \hat{\theta} \right\}}{B} \right)
\]

\[
\hat{a} = \frac{\sum_{j=1}^{n} \left[ \langle \hat{\theta}(j) \rangle - \hat{\theta}(j) \right]^3}{6 \left\{ \sum_{j=1}^{n} \left[ \langle \hat{\theta}(j) \rangle - \hat{\theta}(j) \right]^2 \right\}^{3/2}}.
\]

Bias correction

\(\alpha\), acceleration; \(\diamond\), average; \(\hat{\theta}_j\), jackknife value (leave one point, \(j\), out)
Part 2: Time Series Analysis

Bootstrap confidence intervals

The standard normal (Gaussian) distribution has the following PDF:

\[ f(x) = \left(2\pi\right)^{-1/2} \exp\left(-x^2/2\right) \]

Standard normal probability density function (PDF)

The distribution function, \( F(x) \), is given by:

\[ F(x) = \int_{-\infty}^{x} f(x') \, dx' \]

Standard normal probability distribution function

---

Book, Eqs. (3.48), (3.49)
Correlation estimation, Agulhas current, Pleistocene

Step 1  Bivariate time series  \( \{t(i), x(i), y(i)\}_{i=1}^{n} \)

Step 2  Pearson’s \( r_{XY} \) (Eq. 7.5)

Step 3  Fisher’s \( z \)-transformation  \( z = \tanh^{-1} (r_{XY}) \)

Step 4  Estimated, bias-corrected persistence time, process \( X(i) \), \( \hat{\tau}'_X \)
using mean-detrended time series, \( \{x(i) - \bar{x}\}_{i=1}^{n} \)

Step 5  Analogously, process \( Y(i) \) \( \hat{\tau}'_Y \)

Step 6  Estimated, bias-corrected equivalent autocorrelation coefficient, process \( X(i) \) \( \hat{a}'_X = \exp \left( -\bar{d}/\hat{\tau}'_X \right) \)

Step 7  Analogously, process \( Y(i) \) \( \hat{a}'_Y = \exp \left( -\bar{d}/\hat{\tau}'_Y \right) \)
Part 3: Examples

Correlation estimation, Agulhas current, Pleistocene

Step 8 Effective data size,

\[ n'_\rho = n \left\{ 1 + \frac{2}{n} \frac{1}{1 - a_X a_Y} \left[ a_X a_Y \left( n - \frac{1}{1 - a_X a_Y} \right) \right] \right\}^{-1} \]

obtained by plugging in

\( \hat{a}'_X \) for \( a_X \) and \( \hat{a}'_Y \) for \( a_Y \)

in Eq. (2.38)

Step 9 Approximate, classical

normal CI for \( r_{XY} \),

\[ \text{CI}_{r_{XY}, 1-2\alpha} = \left[ \tanh \left[ z + z(\alpha) \cdot (n'_\rho - 3)^{-1/2} \right] \right] \]

obtained from

\[ \tanh \left[ z - z(\alpha) \cdot (n'_\rho - 3)^{-1/2} \right] \]

re-transforming \( z \)
Step 1. Bivariate time series  
\( \{t(i), x(i), y(i)\}_{i=1}^{n} \)

Step 2. Pearson’s \( r_{XY} \) (Eq. 7.5)

Step 3. Fisher’s z-transformation  
\[ z = \tanh^{-1}(r_{XY}) \]

Step 4. Estimated, bias-corrected persistence time, process \( X(i) \), \( \widehat{\tau}'_X \)

using mean-detrended time series, \( \{x(i) - \bar{x}\}_{i=1}^{n} \)

Step 5. Analogously, process \( Y(i) \) \( \widehat{\tau}'_Y \)

Step 6. Select block length \( l \)

Step 7. Apply MBB with \( l \)  
\( \{x^*b(i)\}_{i=1}^{n} = \{x(f(i))\}_{i=1}^{n} \)  
(Algorithm 3.1) to \( x \) values \( (b, \text{counter}) \)

Step 8. Overtake bootstrap index \( f(i) \)

for resampled \( y \) values  
\( \{y^*b(i)\}_{i=1}^{n} = \{y(f(i))\}_{i=1}^{n} \)

Step 9. Resample  
\( \{x^*b(i), y^*b(i)\}_{i=1}^{n} \)
Correlation estimation, Agulhas current, Pleistocene

Algorithm 7.2 (continued)

Step 1 Bivariate time series 
\{t(i), x(i), y(i)\}_{i=1}^{n}

Step 2 Pearson's $r_{XY}$ (Eq. 7.5)

Step 3 Fisher's $z$-transformation $z = \tanh^{-1}(r_{XY})$

Step 4 Estimated, bias-corrected persistence time, process $X(i), \hat{\tau}_X$

using mean-detrended time series, $\{x(i) - \bar{x}\}_{i=1}^{n}$

Step 5 Analogously, process $Y(i), \hat{\tau}_Y$

Step 6 Select block length $l$

Step 7 Apply MBB with $l \times x^*(i), y^*(i)_{i=1}^{n} = l \times f(i), y^*(i)_{i=1}^{n}$ (Algorithm 3.1) to $x$ values ($b$, counter)

Step 8 Overtake bootstrap index $f(i)$ for resampled $y$ values $\{y^*(i)\}_{i=1}^{n} = l \times f(i), y^*(i)_{i=1}^{n}$

Step 9 Resample $\{x^*(i), y^*(i)\}_{i=1}^{n}$

Step 10 Bootstrap replications, Pearson's $r_{XY}$ and Fisher's $z$

$\hat{r}_{XY}, \hat{z}^{(b)} = \tanh^{-1}(r^{(b)}_{XY})$

Step 11 Go to Step 7 until $b = B$

(usually $B = 2000$)

replications exist $\{z^{(b)}\}_{b=1}^{B}$

Step 12 Calculate CI (Section 3.4) for Fisher's $z$

$\text{CI}_{z,1-2\alpha} = [z_l; z_u]$

Step 13 Re-transform lower and upper endpoints to obtain pairwise-MBB CI for $r_{XY}$

$\text{CI}_{r_{XY,1-2\alpha}} = [\tanh(z_l); \tanh(z_u)]$
Calibration of equi-tailed bootstrap confidence intervals

1. \( \hat{\theta}, \{\hat{\theta}^*_b\}_{b=1}^B \)

2. \([\hat{\theta}^*_1(\lambda); \hat{\theta}^*_1(\lambda)]_{b=1} \)

3. \([\hat{\theta}^*_2(\lambda); \hat{\theta}^*_2(\lambda)]_{b=2} \)

4. \([\hat{\theta}^*_3(\lambda); \hat{\theta}^*_3(\lambda)]_{b=B} \)

5. Select \( \lambda' \) such that estimate \( \hat{\theta} \) is in \((1-2\alpha) \times B \) cases in \( \{ \} \)

6. Set confidence level to \((1-2\lambda')\)
Grid of confidence levels

\[ \hat{\theta}_1^b(\lambda), \quad \lambda = 0.01, \ldots, 0.99. \]

Calibration curve

\[ \hat{p}(\lambda) = \frac{\# \left\{ \hat{\theta}_1^b(\lambda) < \hat{\theta} < \hat{\theta}_u^b(\lambda) \right\}}{B}, \]
1. If persistence exists, then we can recover correlation information.
2. Binning is better than interpolation.