MONASH University

MONASH BUSINESS SCHOOL

# High-dimensional time series analysis

Rob J Hyndman

5 December 2018

# Outline

1. Visualization
2. Forecasting
3. Anomaly detection
4. Forecast reconciliation

# Outline

# M3 competition

ELSEVIER

International Journal of Forecasting 16 (2000) 451–476

international journal of forecasting

www.elsevier.com/locate/ijforecast

## The M3-Competition: results, conclusions and implications

Spyros Makridakis, Michèle Hibon*

*INSEAD, Boulevard de Constance, 77305 Fontainebleau, France*

---

**Abstract**

This paper describes the M3-Competition, the latest of the M-Competitions. It explains the reasons for conducting the competition and summarizes its results and conclusions. In addition, the paper compares such results/conclusions with those of the previous two M-Competitions as well as with those of other major empirical studies. Finally, the implications of these results and conclusions are considered, their consequences for both the theory and practice of forecasting are explored and directions for future research are contemplated. 

*Keywords:* Comparative methods — time series: univariate; Forecasting competitions; M-Competition; Forecasting methods, Forecasting accuracy

# M3 competition

ELSEVIER

International Journal of Forecasting 16 (2000) 451–476

international journal of forecasting

www.elsevier.com/locate/ijforecast


etition: results, conclusions a s

Spyros Makridakis, Michèle Hibon*

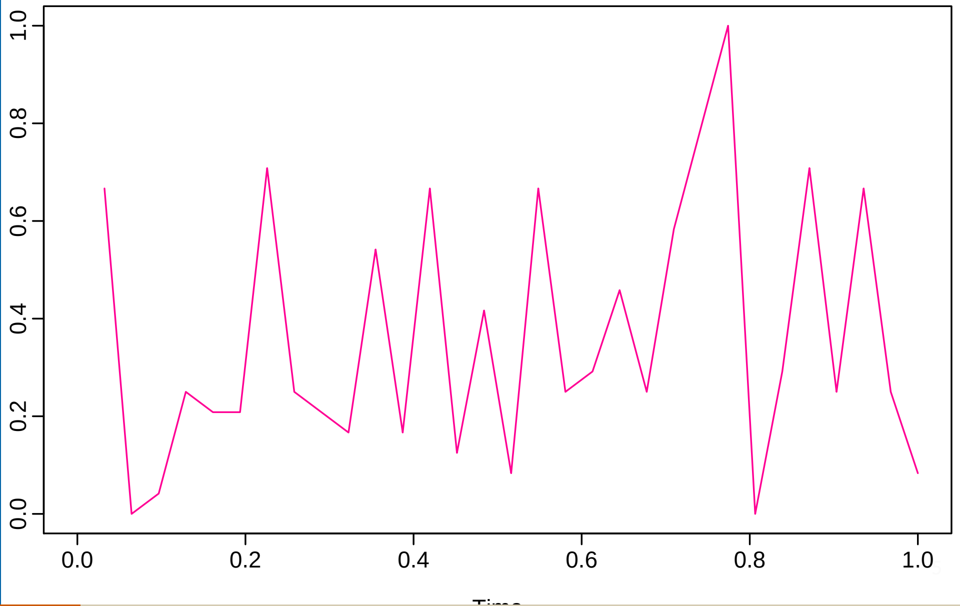
EAD, Boulevard de Constance, 77305 Fontainebleau, Fr

**Abst**

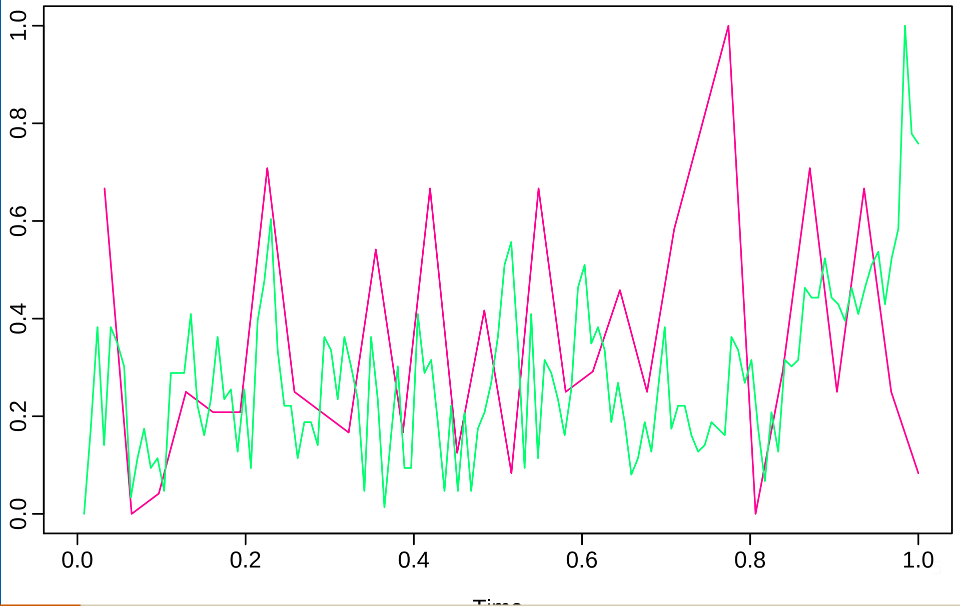
This paper describes the M3-Competition, the latest of the M-Competitions. It explains the reasons for conducting the competition and summarizes its results and conclusions. In addition, the paper compares such results/conclusions with those of the previous two M-Competitions as well as with those of other major empirical studies. Finally, the implications of these results and conclusions are considered, their consequences for both the theory and practice of forecasting are explored and directions for future research are contemplated. 

*Keywords:* Comparative methods — time series: univariate; Forecasting competitions; M-Competition; Forecasting methods, Forecasting accuracy

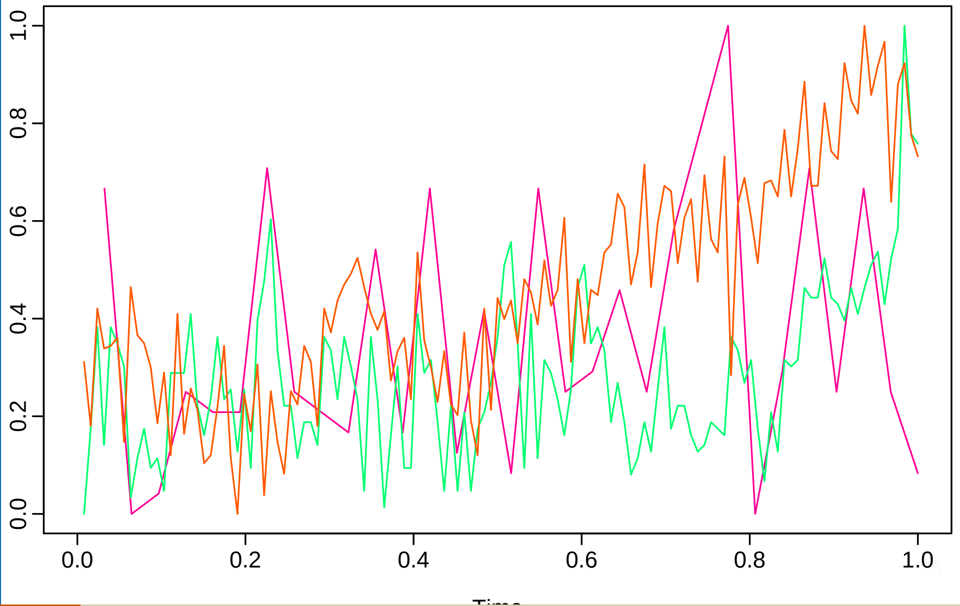
# How to plot lots of time series?



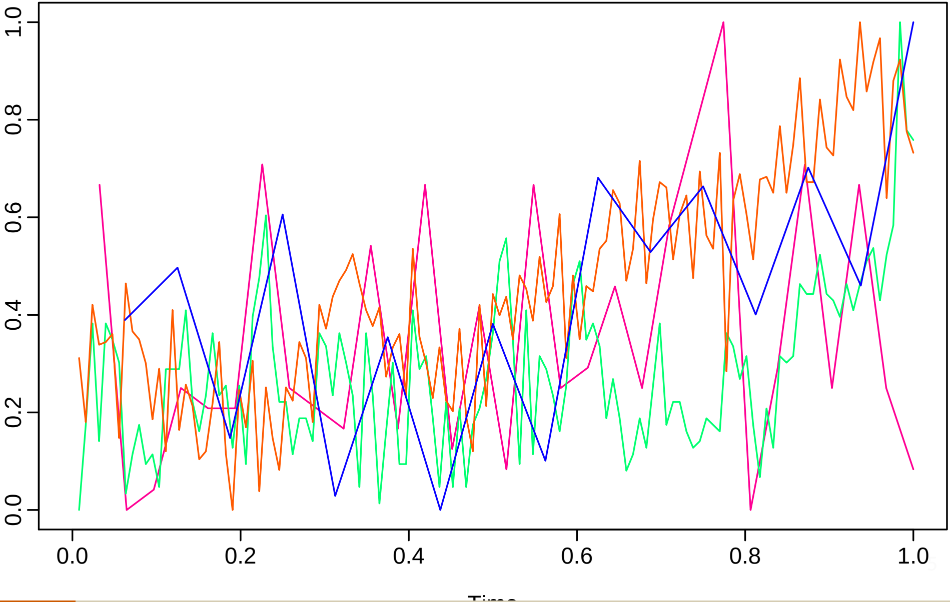
# How to plot lots of time series?



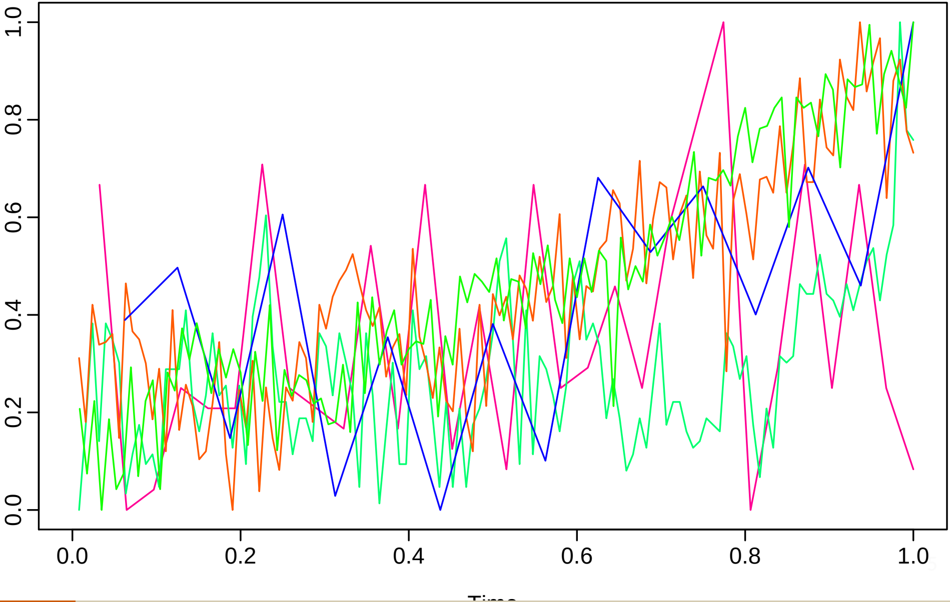
# How to plot lots of time series?



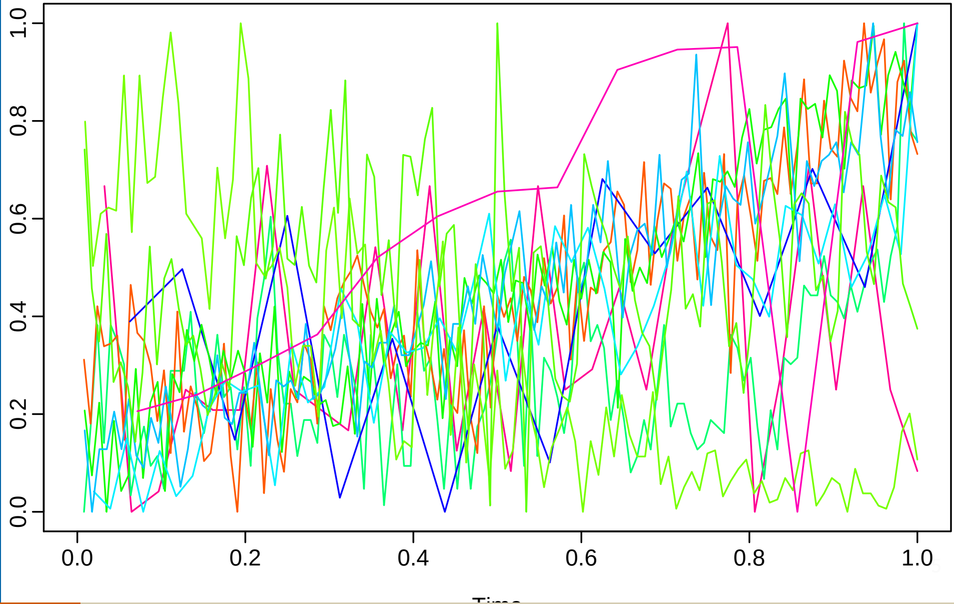
# How to plot lots of time series?



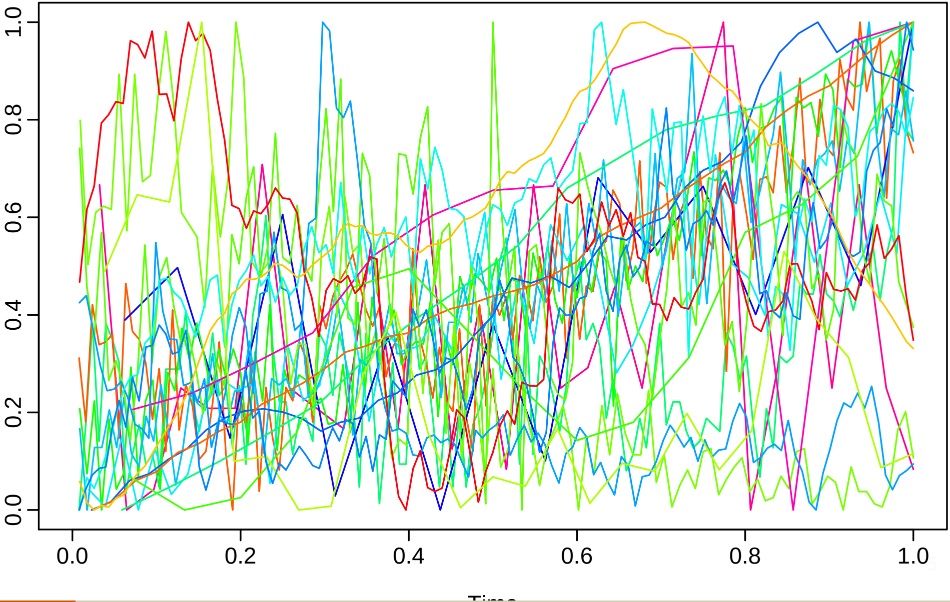
# How to plot lots of time series?



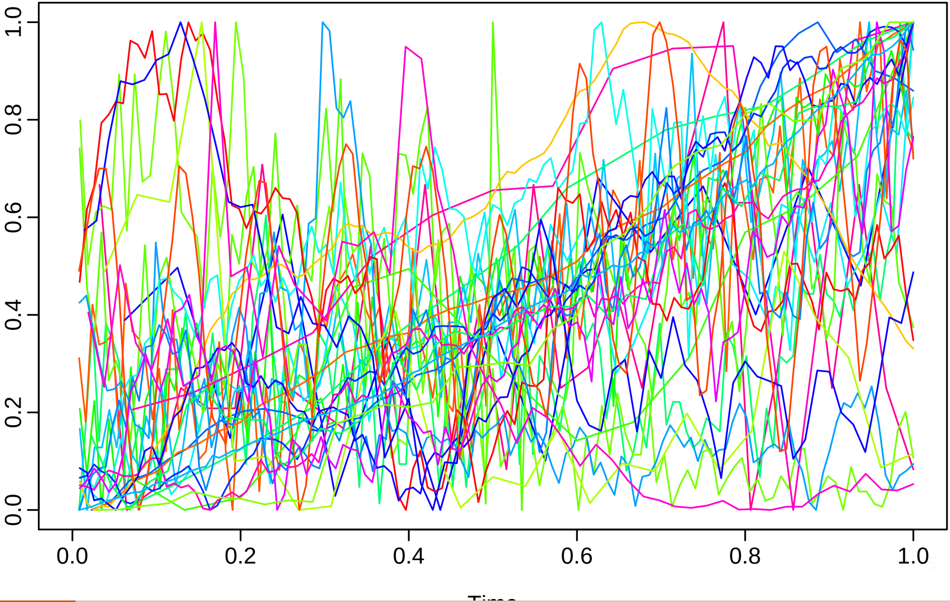
# How to plot lots of time series?



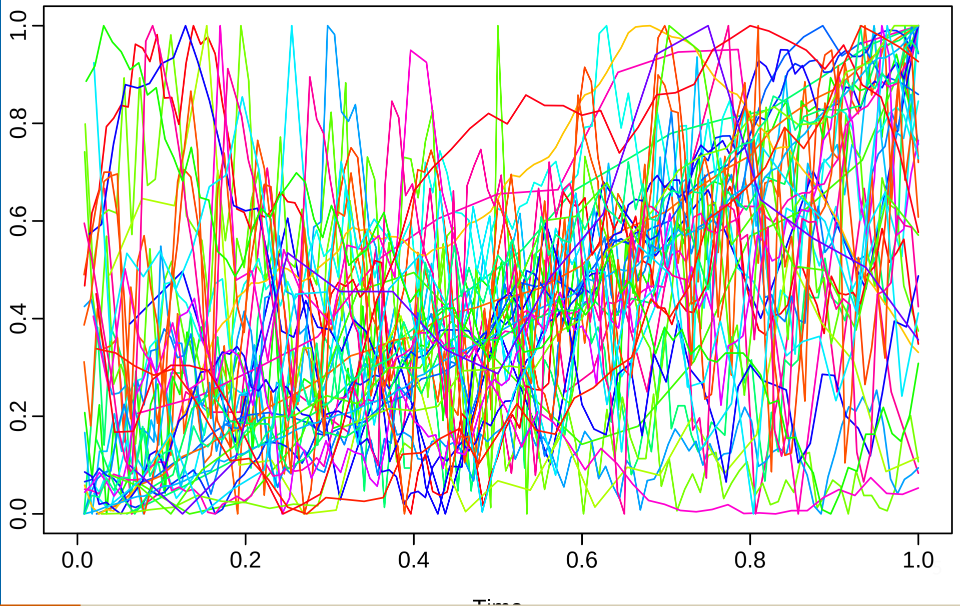
# How to plot lots of time series?



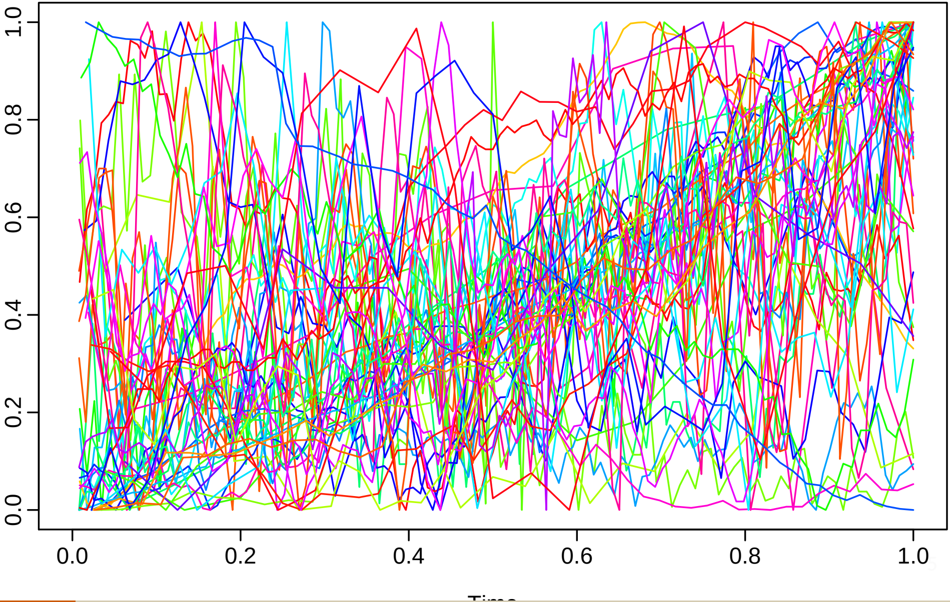
# How to plot lots of time series?



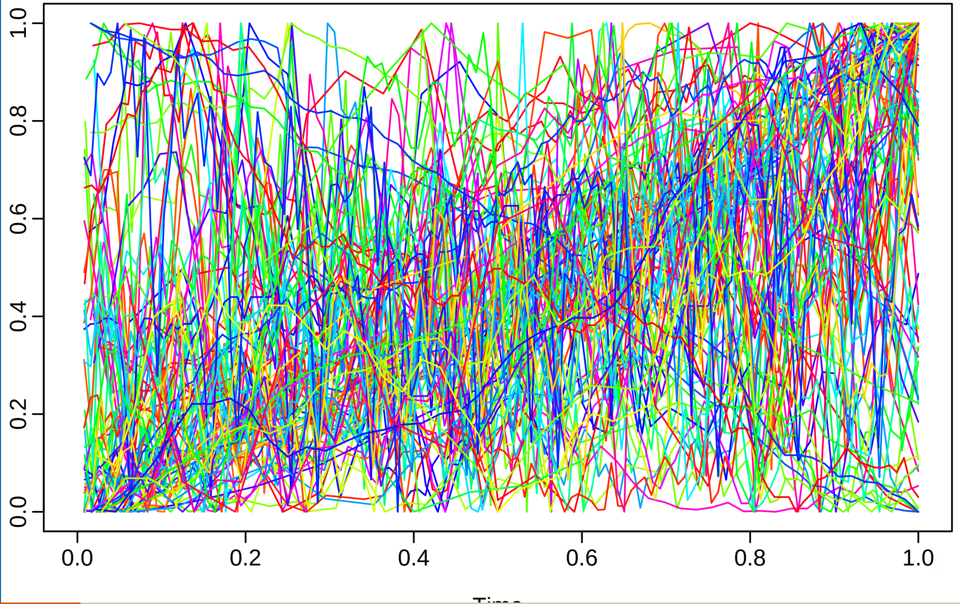
# How to plot lots of time series?



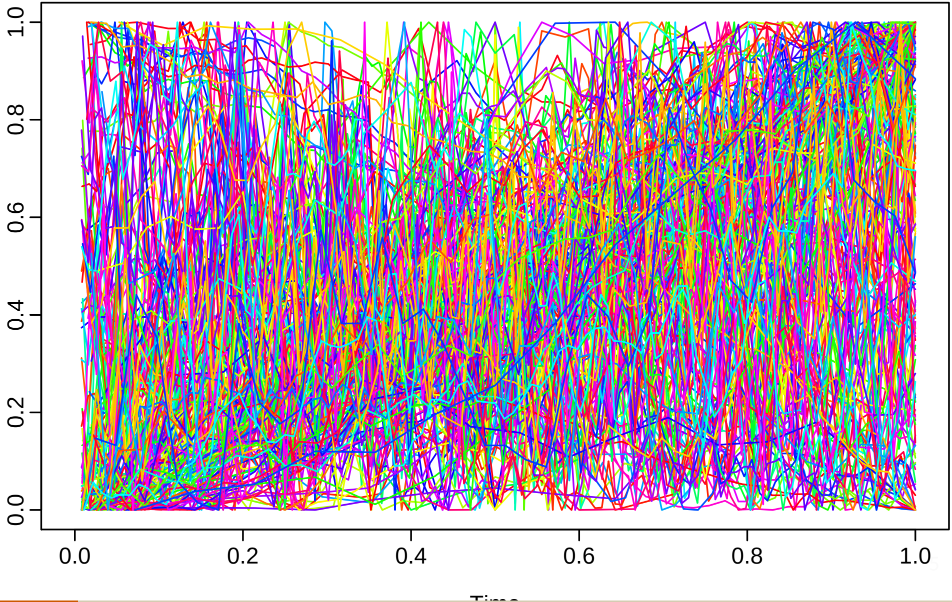
# How to plot lots of time series?



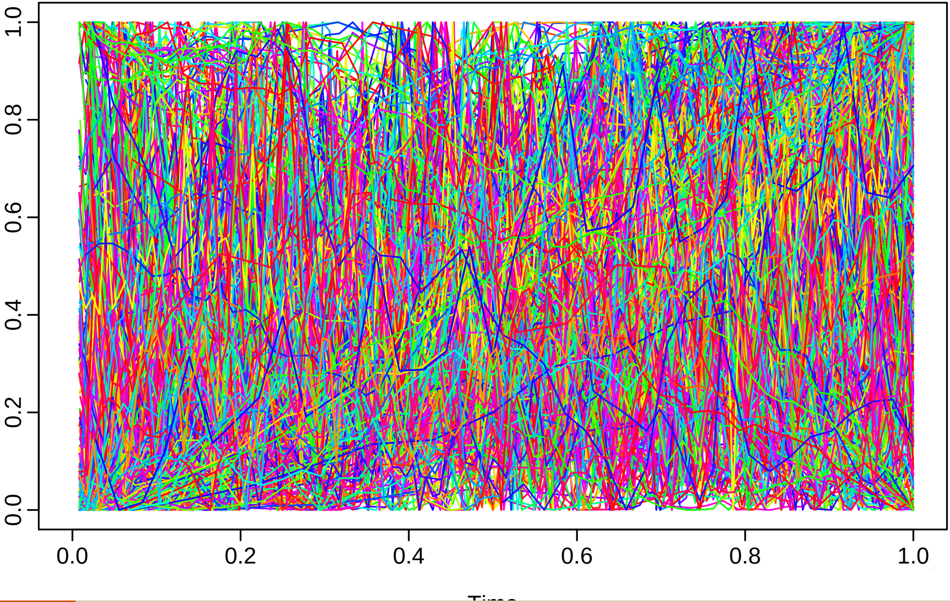
# How to plot lots of time series?



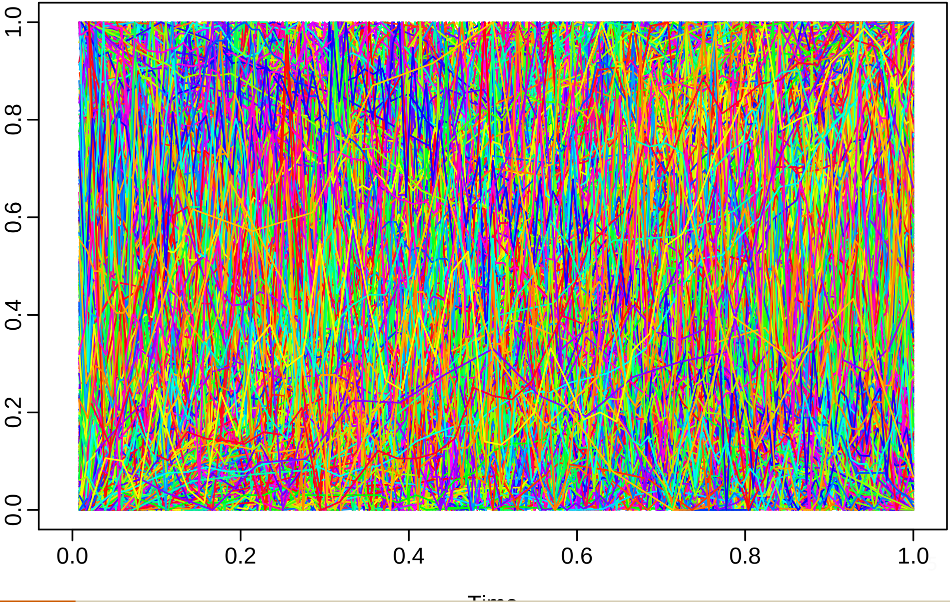
# How to plot lots of time series?



# How to plot lots of time series?



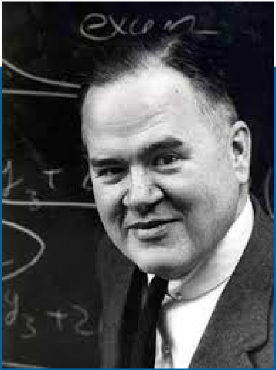
# How to plot lots of time series?



# Key idea

**Cognostics**

Computer-produced diagnostics (Tukey and Tukey, 1985).

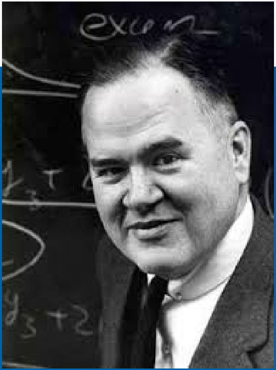


*John W Tukey*

# Key idea

**Cognostics**

Computer-produced diagnostics (Tukey and Tukey, 1985).




*John W Tukey*

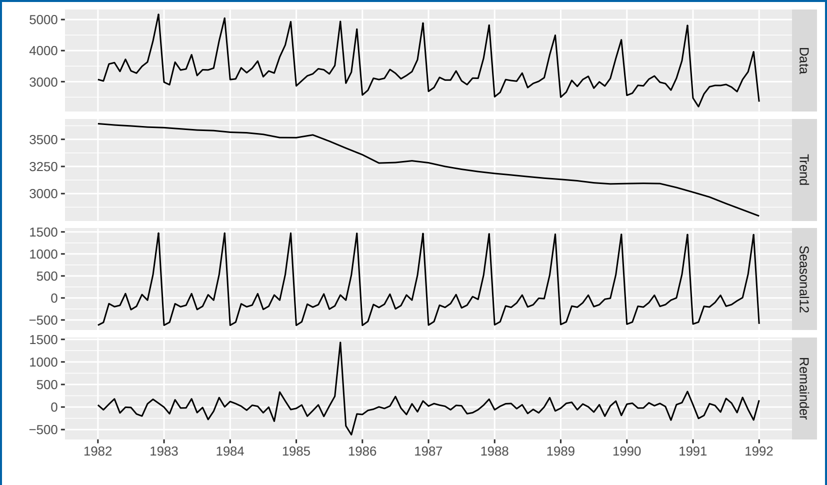
## Examples for time series

- lag correlation
- size and direction of trend
- strength of seasonality
- timing of peak seasonality
- spectral entropy

Called “features” in the machine learning literature.

# An STL decomposition: N2096

$Y_t = S_t + T_t + R_t$ $S_t$ is periodic with mean 0



# Candidate features

## STL decomposition

$$Y_t = S_t + T_t + R_t$$

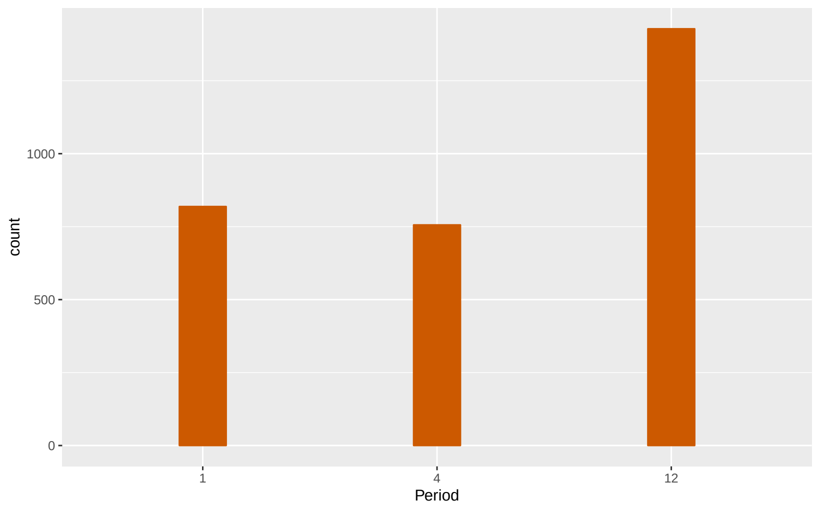
# Candidate features

**STL decomposition**

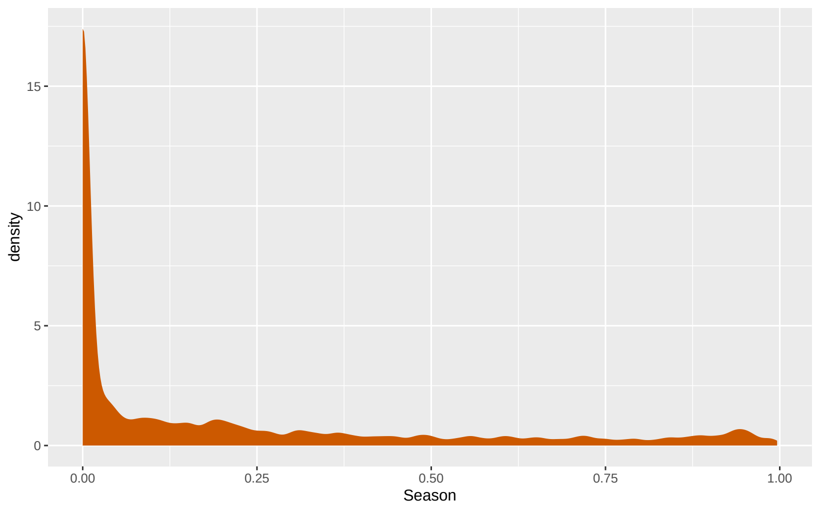
$$Y_t = S_t + T_t + R_t$$

- Seasonal period
- Autocorrelations of data $(Y_1, \dots, Y_T)$
- Autocorrelations of data $(R_1, \dots, R_T)$
- Strength of seasonality: $\max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(Y_t - T_t)}\right)$
- Strength of trend: $\max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(Y_t - S_t)}\right)$
- Spectral entropy: $H = -\int_{-\pi}^{\pi} f_y(\lambda) \log f_y(\lambda) d\lambda$, where $f_y(\lambda)$ is spectral density of $Y_t$. Low values of $H$ suggest a time series that is easier to forecast (more signal).
- Optimal Box-Cox transformation of data

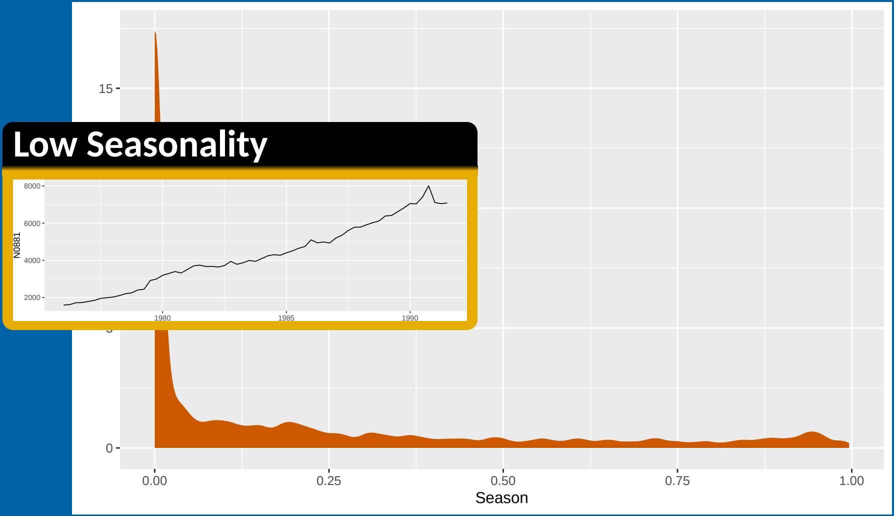
# Distribution of Period for M3



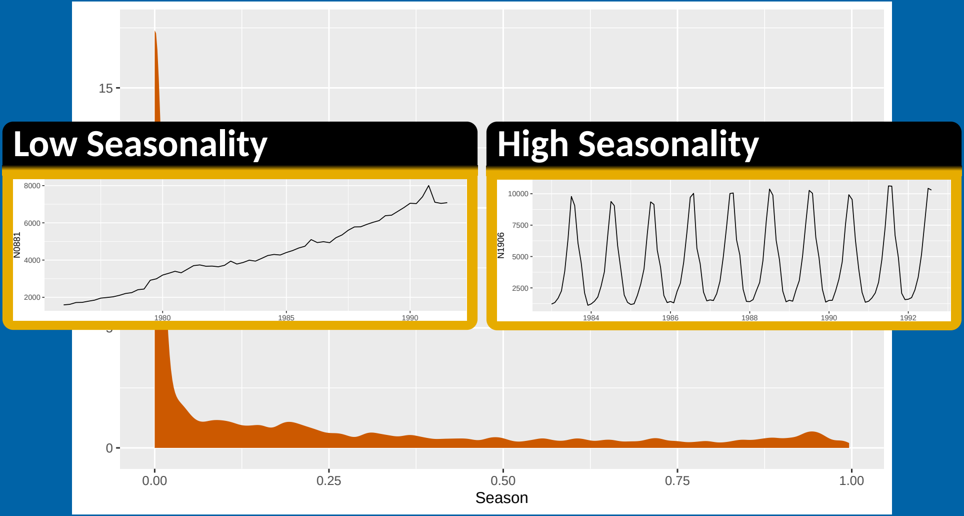
# Distribution of Seasonality for M3



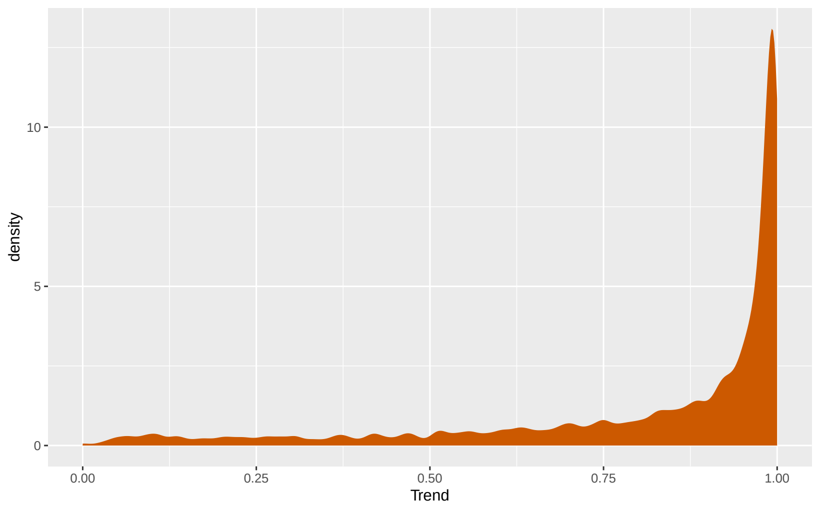
# Distribution of Seasonality for M3



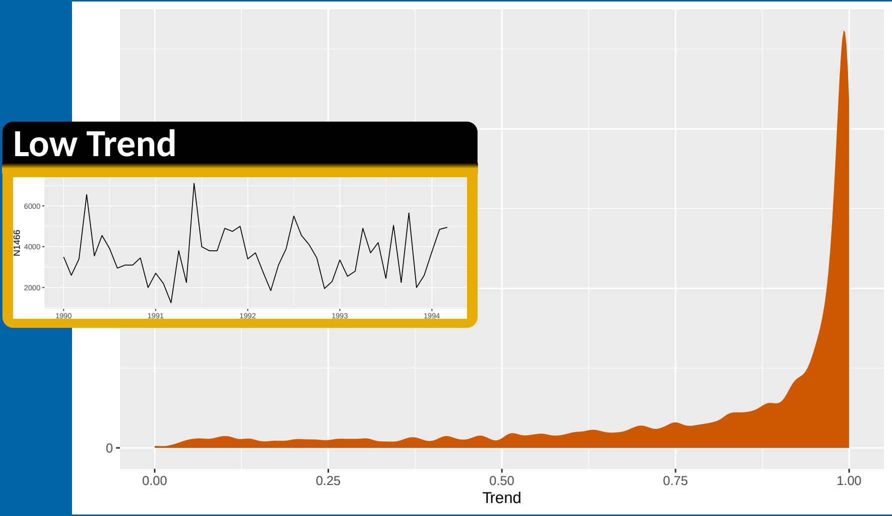
# Distribution of Seasonality for M3



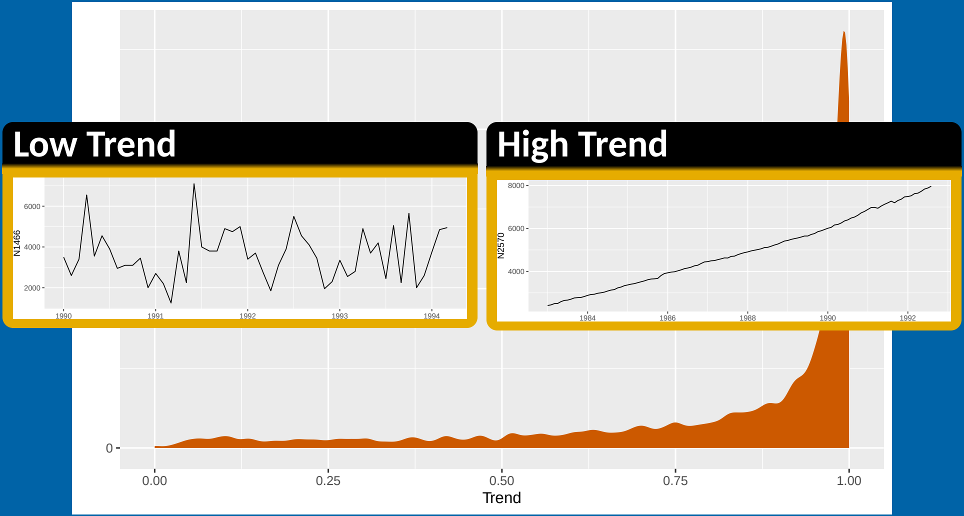
# Distribution of Trend for M3



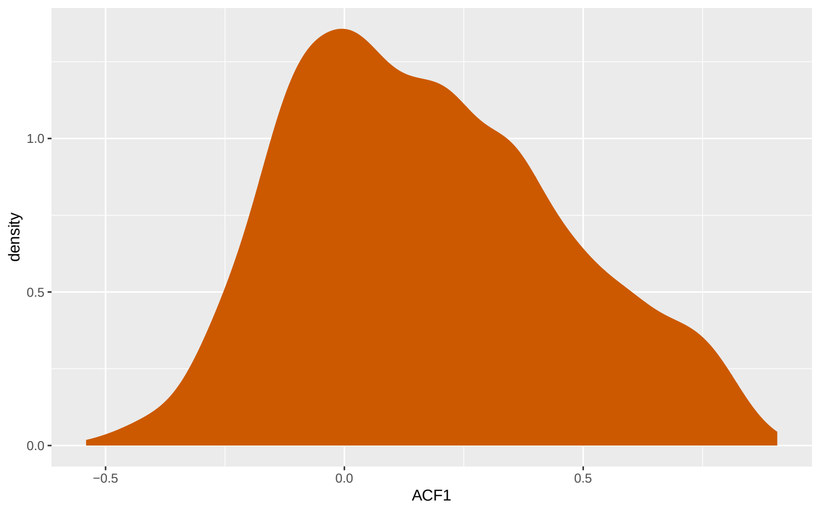
# Distribution of Trend for M3



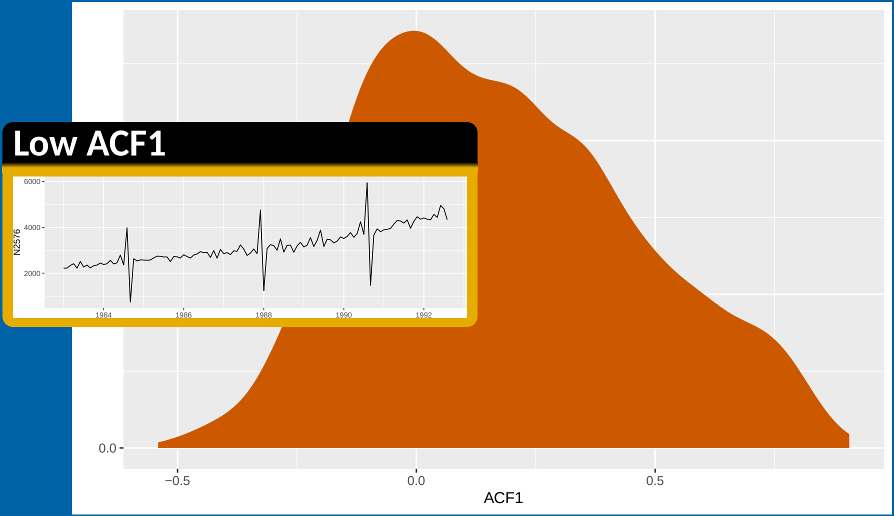
# Distribution of Trend for M3



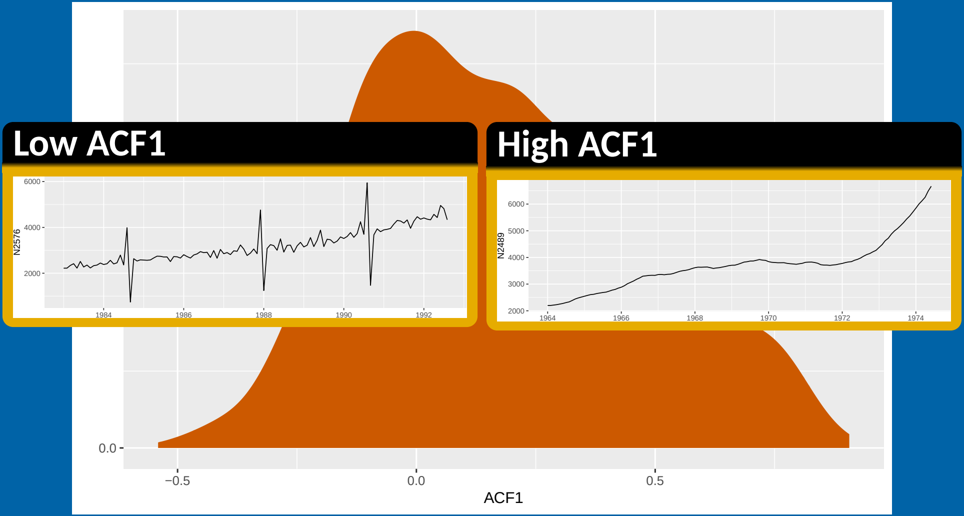
# Distribution of Residual ACF1 for M3



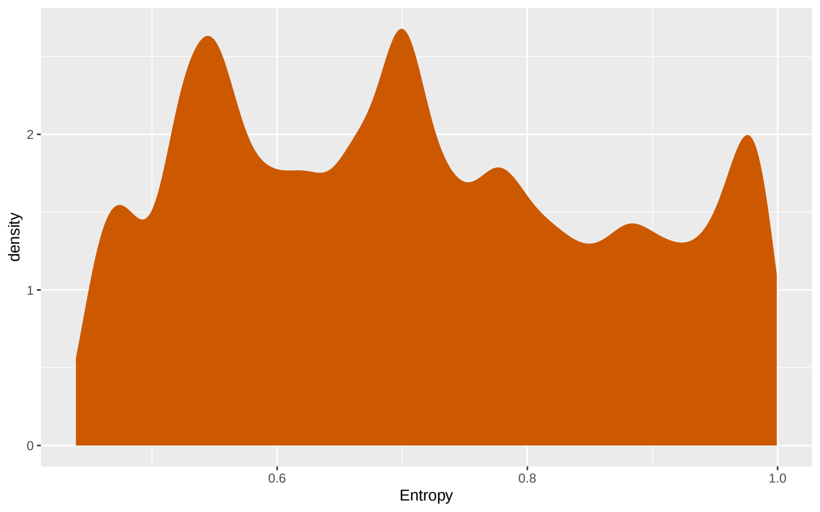
# Distribution of Residual ACF1 for M3



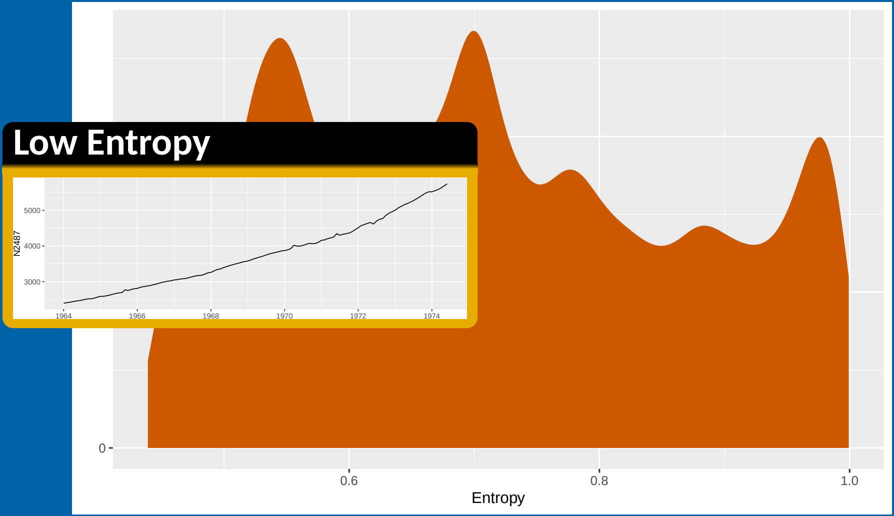
# Distribution of Residual ACF1 for M3



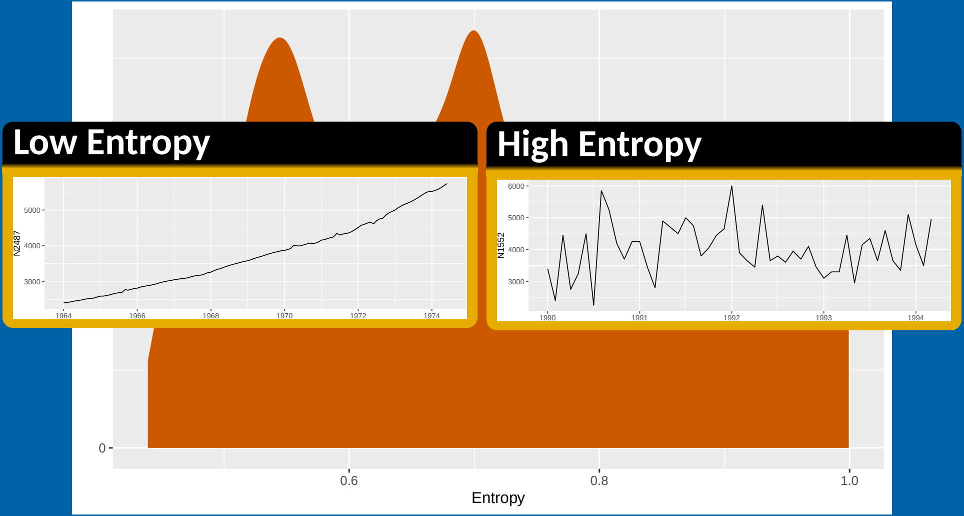
# Distribution of Spectral Entropy for M3



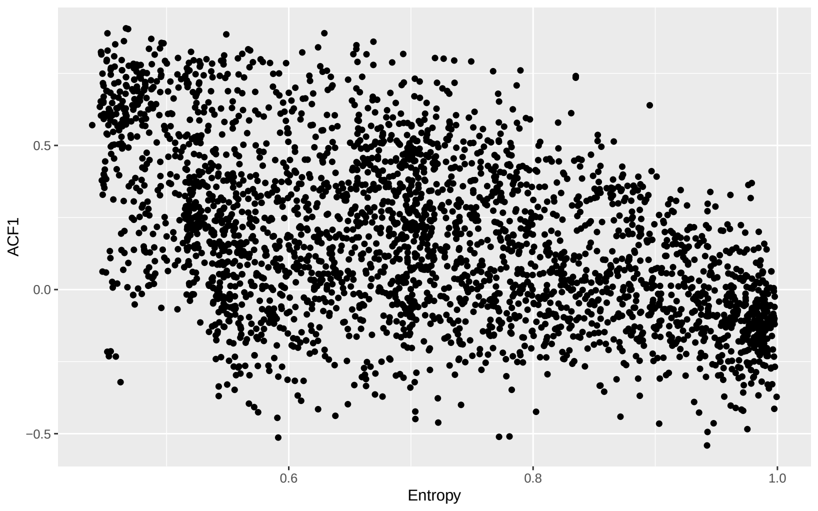
# Distribution of Spectral Entropy for M3



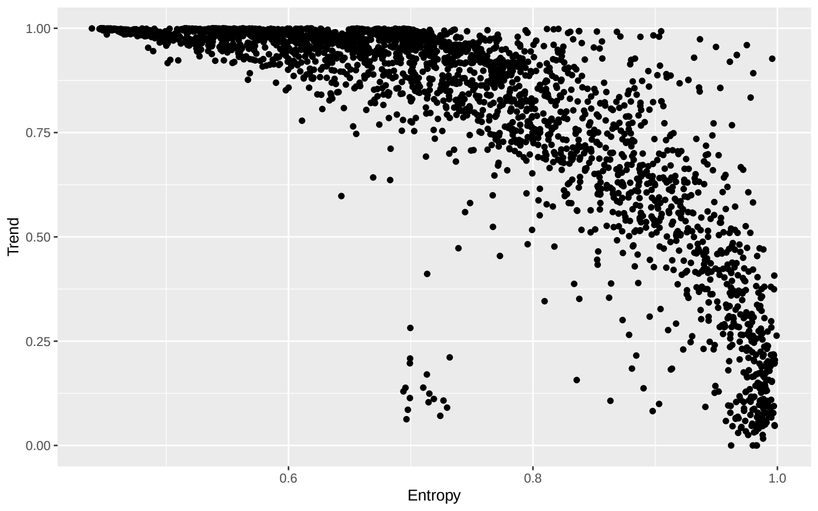
# Distribution of Spectral Entropy for M3



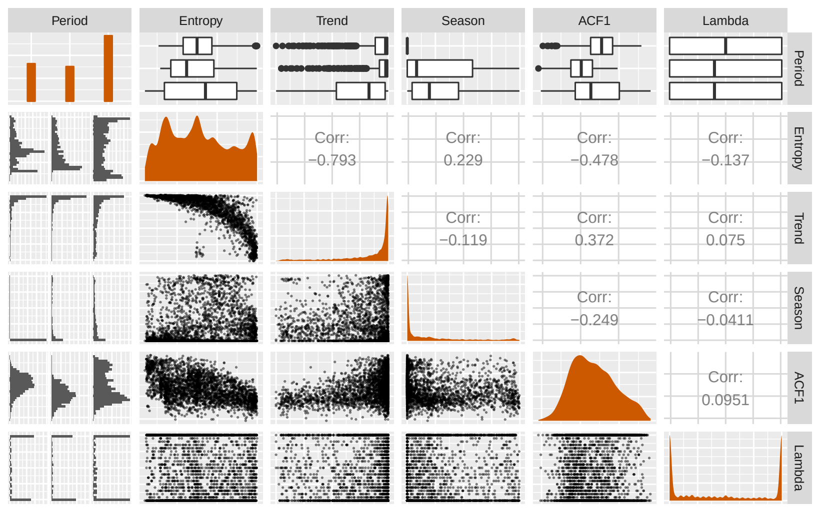
# Feature distributions



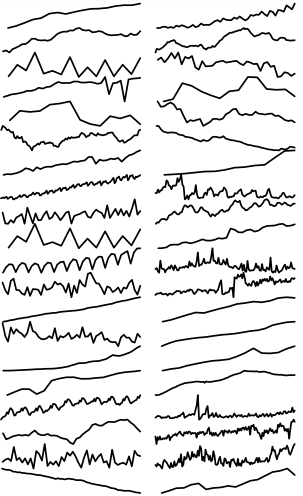
# Feature distributions



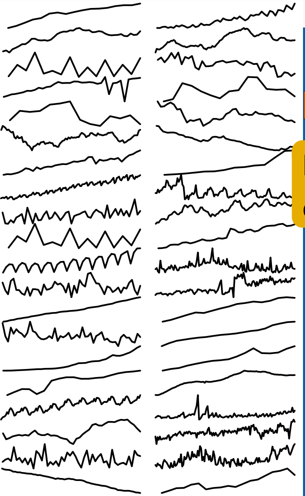
# Feature distributions



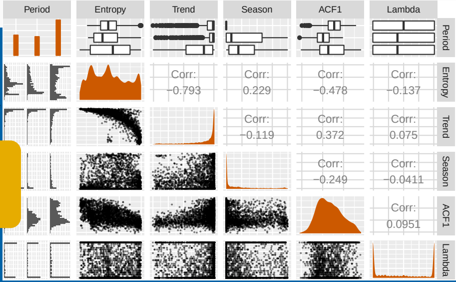
# Dimension reduction for time series



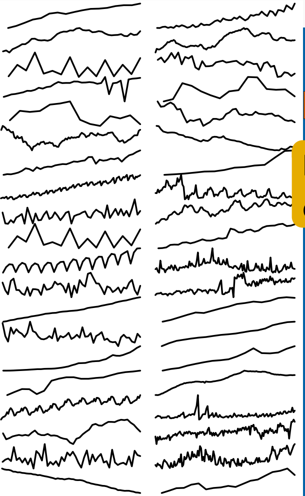
# Dimension reduction for time series



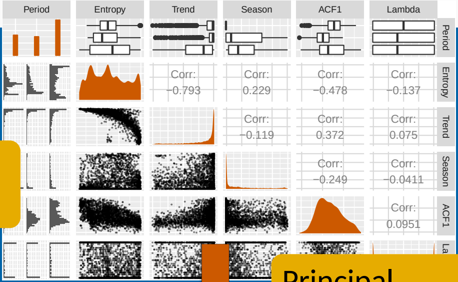
Feature calculation



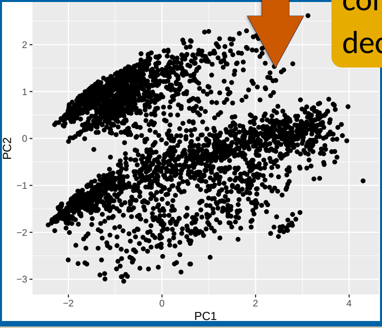
# Dimension reduction for time series



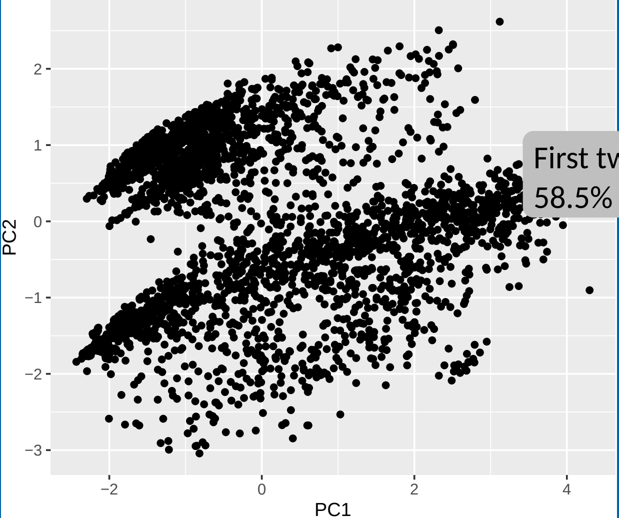
Feature calculation




Principal component decomposition

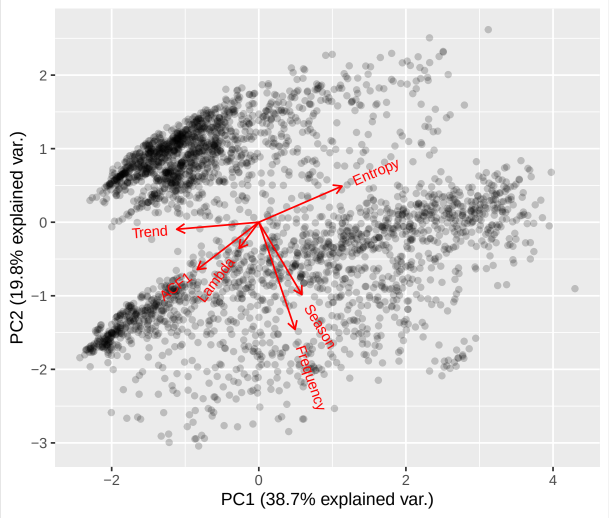


# M3 feature space

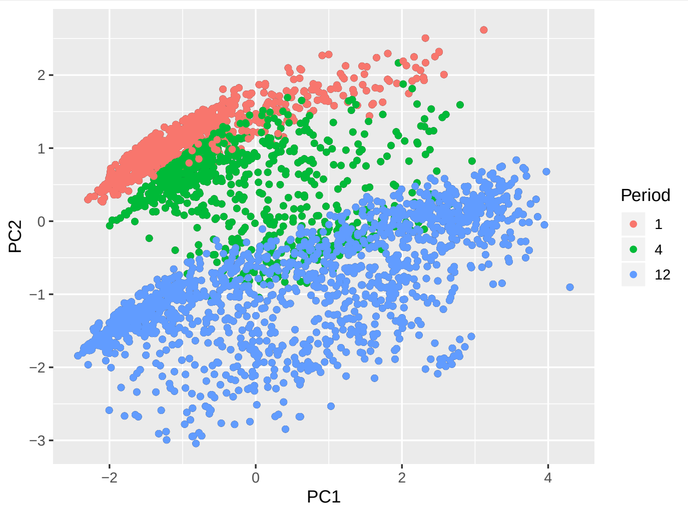



First two PCs explain 58.5% of the variance.

# M3 feature space



# M3 feature space



# Feature properties

In this analysis, we have restricted features to be

- ergodic
- scale-independent

For other analyses, it may be appropriate to have different requirements.

# Feature properties

In this analysis, we have restricted features to be

- ergodic
- scale-independent

For other analyses, it may be appropriate to have different requirements.

**R package**

**github.com/robjhyndman/tsfeatures**

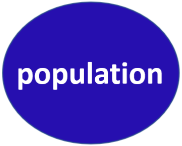
# Outline

# Forecast model selection

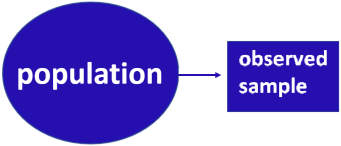
## Features used to select a forecasting model

- length
- strength of seasonality
- strength of trend
- linearity
- curvature
- spikiness
- stability
- lumpiness
- first ACF value of remainder series
- parameter estimates of Holt's linear trend method
- spectral entropy
- Hurst exponent
- nonlinearity
- parameter estimates of Holt-Winters' additive method
- unit root test statistics
- first ACF value of residual series of linear trend model
- ACF and PACF based features - calculated on both the raw and differenced series

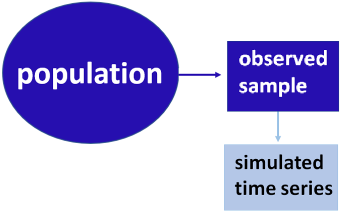
# FFORMS: Feature-based FORecast Model Selection



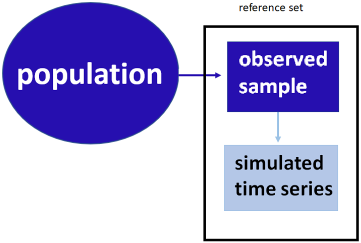
# FFORMS: Feature-based FORecast Model Selection



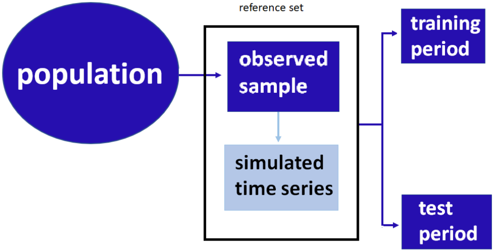
# FFORMS: Feature-based FORecast Model Selection



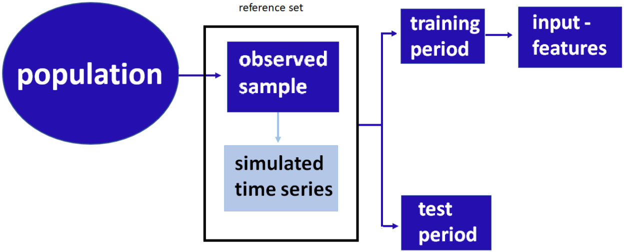
# FFORMS: Feature-based FORecast Model Selection



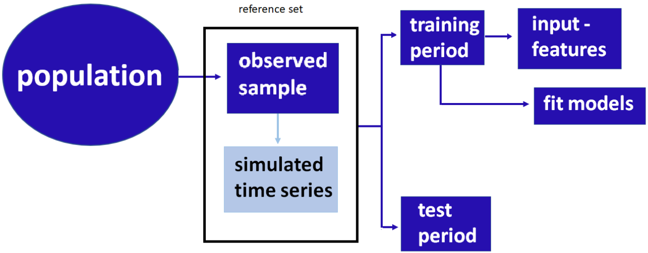
# FFORMS: Feature-based FORecast Model Selection



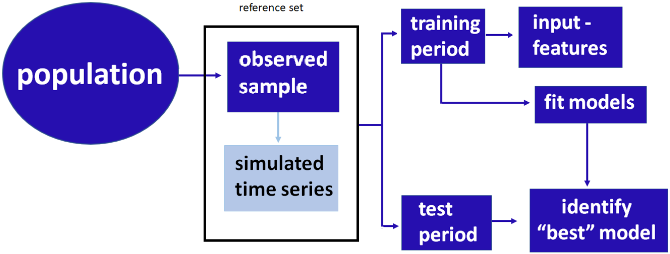
# FFORMS: Feature-based FORecast Model Selection



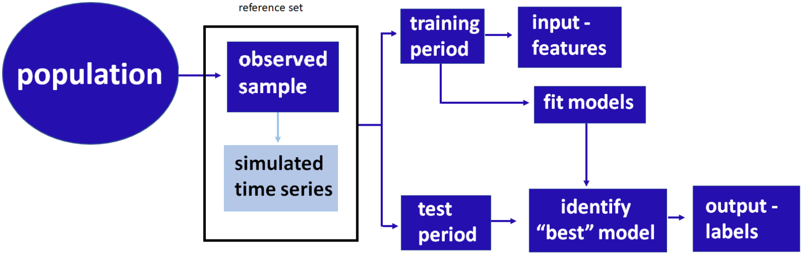
# FFORMS: Feature-based FORecast Model Selection



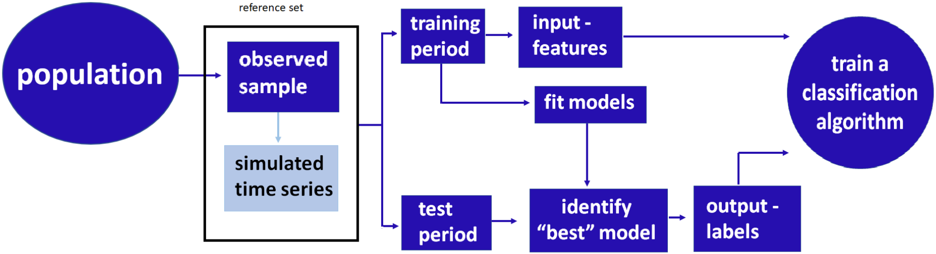
# FFORMS: Feature-based FORecast Model Selection



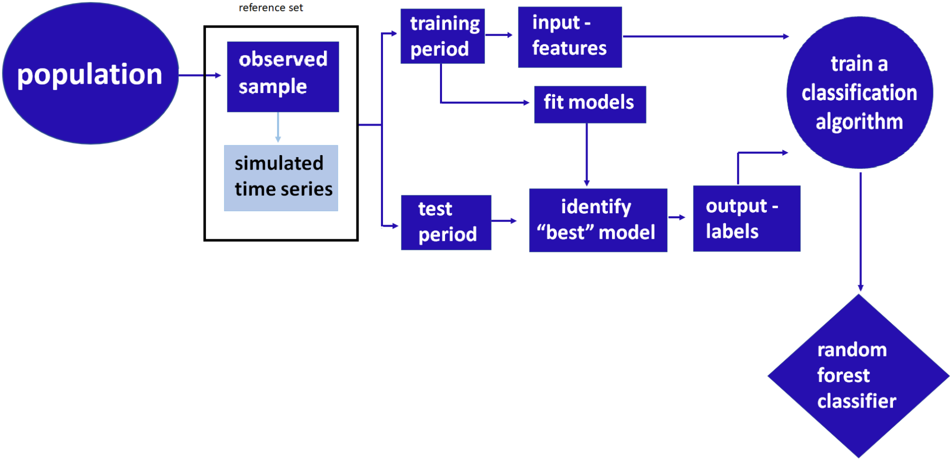
# FFORMS: Feature-based FORecast Model Selection



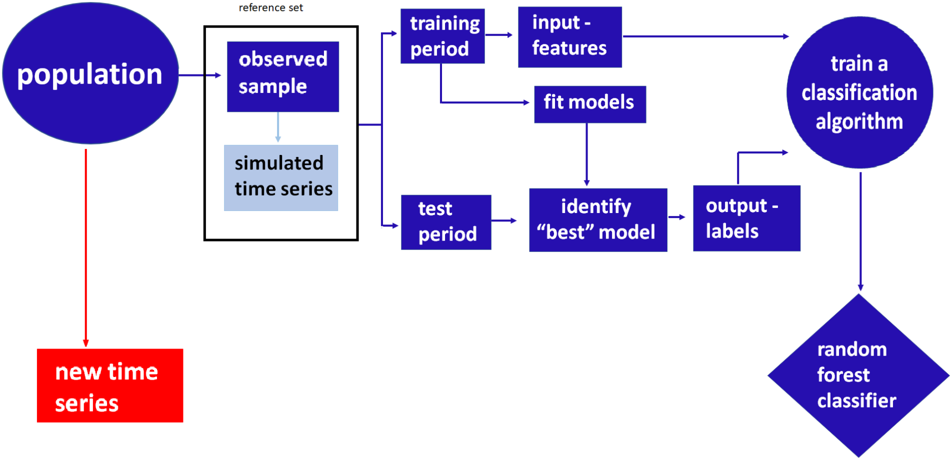
# FFORMS: Feature-based FORecast Model Selection



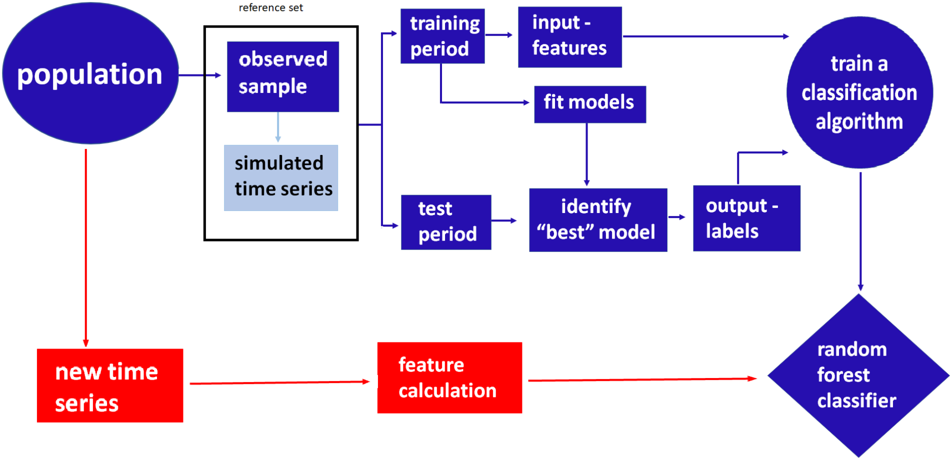
# FFORMS: Feature-based FORecast Model Selection



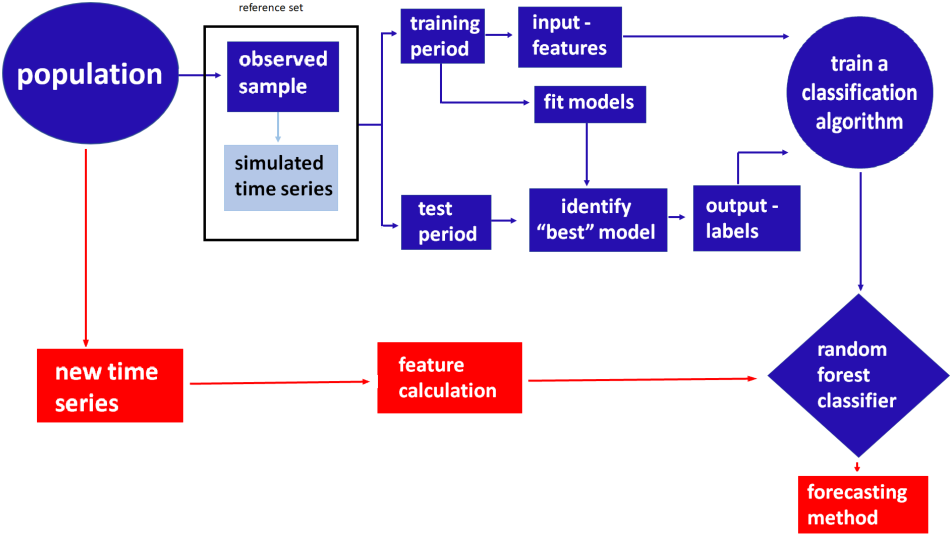
# FFORMS: Feature-based FORecast Model Selection



# FFORMS: Feature-based FORecast Model Selection



# FFORMS: Feature-based FORecast Model Selection



# Application to M competition data

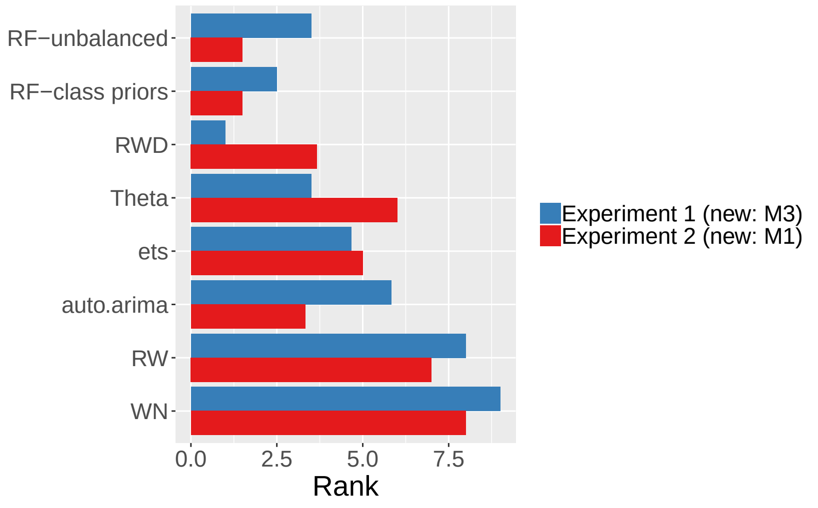
## Experiment 1

| | Source | Y | Q | M |
|---|---|---|---|---|
| Observed series | M1 | 181 | 203 | 617 |
| Simulated series | | 362000 | 406000 | 123400 |
| New series | M3 | 645 | 756 | 1428 |

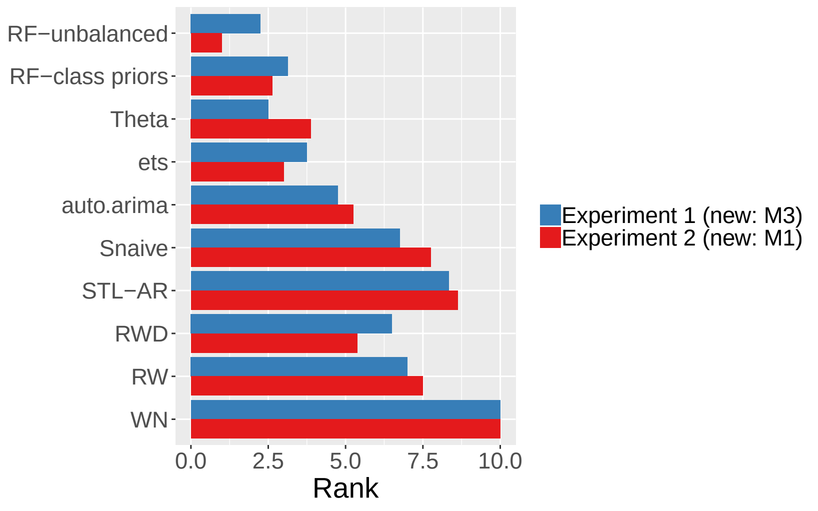
## Experiment 2

| | Source | Y | Q | M |
|---|---|---|---|---|
| Observed series | M3 | 645 | 756 | 1428 |
| Simulated series | | 1290000 | 1512000 | 285600 |
| New series | M1 | 181 | 203 | 617 |

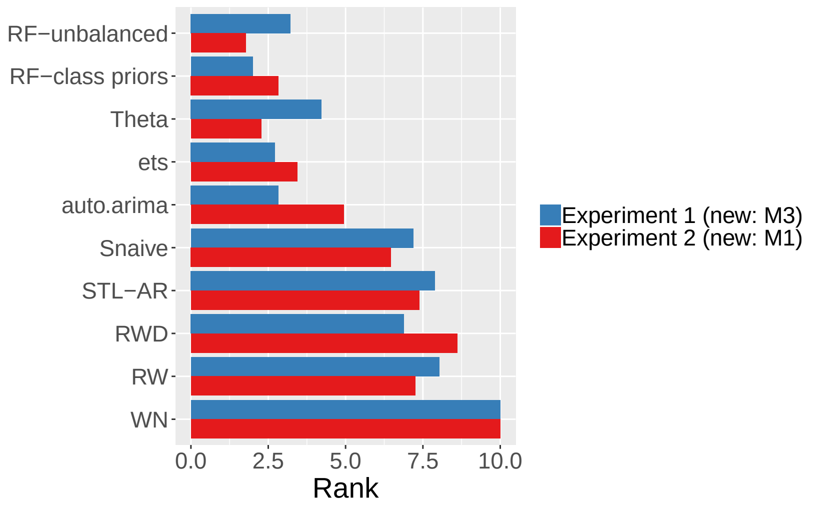
# Results: Yearly



# Results: Quarterly



# Results: Monthly



# FFORMA: Feature-based FORecast Model Averaging

- Like FFORMS but we use gradient boosted trees rather than a random forest.
- The optimization criterion is forecast accuracy not classification accuracy.
- The probability of each model being best is used to construct a model weight.
- A combination forecast is produced using these weights.
- **Came second in the M4 forecasting competition**

# FFORMA: Feature-based FORecast Model Averaging

## Models included

1. Naive
2. Seasonal naive
3. Random walk with drift
4. Theta method
5. ARIMA
6. ETS
7. TBATS
8. STLM-AR

# R Packages

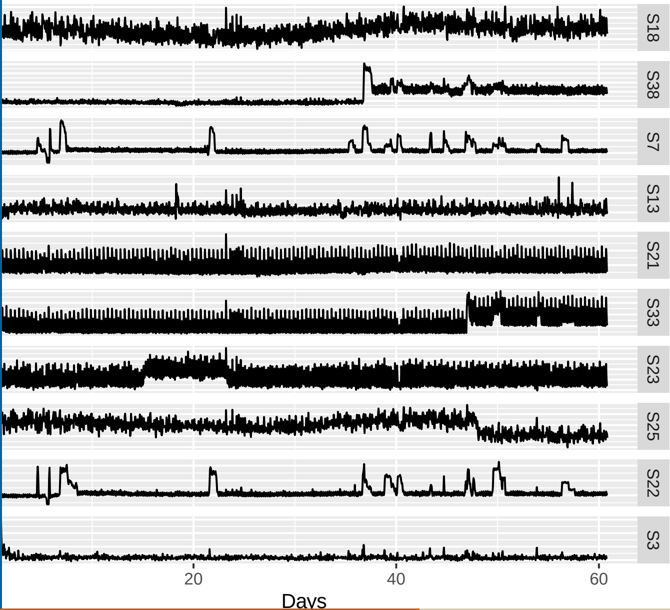
- **seer**: FFORMS — selecting forecasting model using features.
  `github.com/thiyangt/seer`
- **M4metalearning**: FFORMA – forecast combinations using features to choose weights.
  `github.com/robjhyndman/M4metalearning`

# Outline

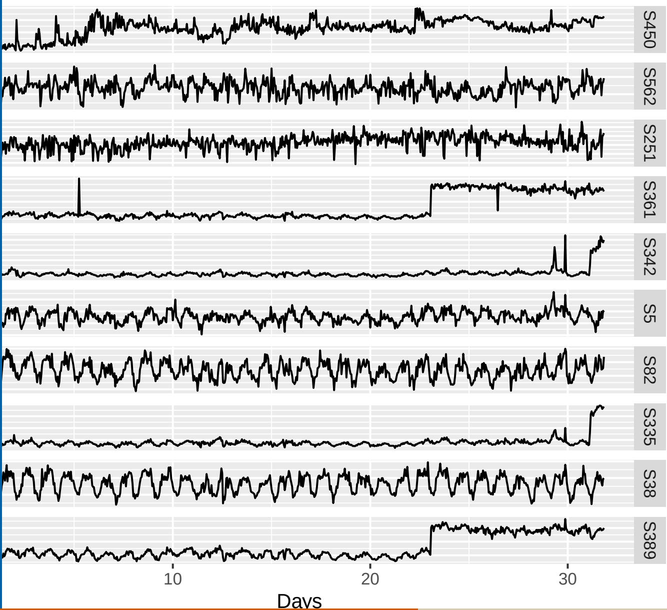
# Yahoo server metrics

- Tens of thousands of time series collected at one-hour intervals over 1–2 months.
- Consisting of several server metrics (e.g. CPU usage and paging views) from many server farms globally.
- Aim: find unusual (anomalous) time series.

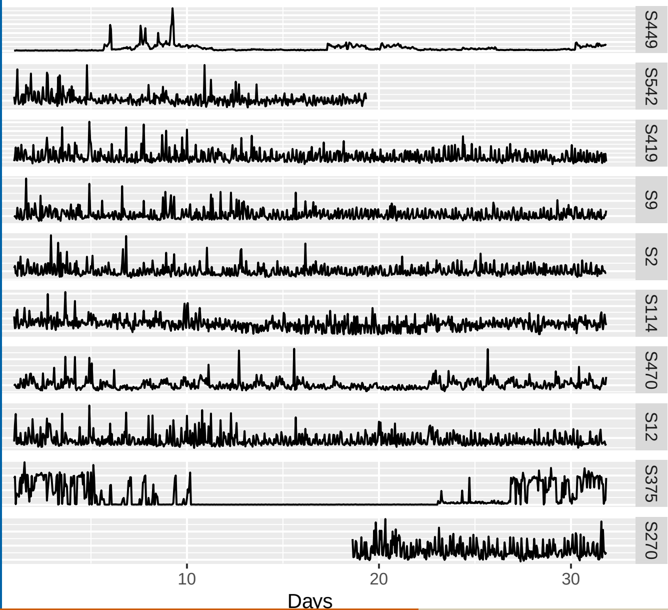
# Yahoo server metrics



# Yahoo server metrics



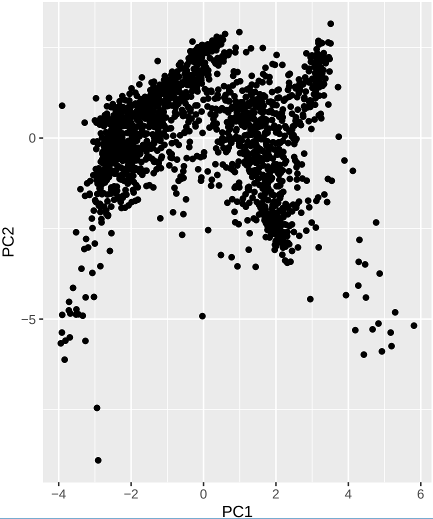
# Yahoo server metrics



# Yahoo server metrics

- **ACF1**: first order autocorrelation = $\text{Corr}(Y_t, Y_{t-1})$
- Strength of **trend** and **seasonality** based on STL
- Size of seasonal **peak** and **trough**
- Spectral **entropy**
- **Lumpiness**: variance of block variances (block size 24).
- **Spikiness**: variances of leave-one-out variances of STL remainders.
- **Level shift**: Maximum difference in trimmed means of consecutive moving windows of size 24.
- **Variance change**: Max difference in variances of consecutive moving windows of size 24.
- **Flat spots**: Discretize sample space into 10 equal-sized intervals. Find max run length in any interval.
- Number of **crossing points** of mean line.
- **Kullback-Leibler score**: Maximum of $D_{KL}(P\|Q) = \int P(x) \ln P(x)/Q(x) dx$ where $P$ and $Q$ are estimated by kernel density estimators applied to consecutive windows

# Feature space



# Feature space

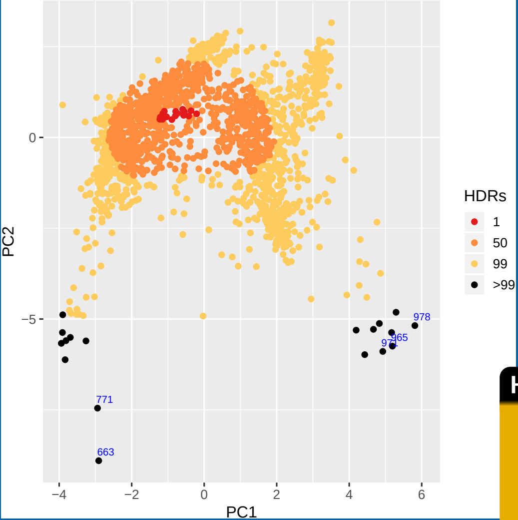
# Feature space


**What is "anomalous"?**

- We need a measure of the "anomalousness" of a time series.
- Rank points based on their local density using a bivariate kernel density estimate.

# Finding weird time series

```r
hdrcde::hdrscatterplot(pc[,1], pc[,2], noutliers=5)
```




**Highest Density Regions**

- Estimate using `hdrcde` package
- Highlight outlying points as those with lowest density.

# Packages

- **hdrcde**: scatterplots with bivariate HDRs.
  CRAN | `github.com/robjhyndman/hdrcde`
- **stray**: finding outliers in high dimensions.
  `github.com/pridiltal/stray`
- **oddstream**: finding outliers in streaming data.
  `github.com/pridiltal/oddstream`
- **anomalous**: yahoo data.
  `github.com/robjhyndman/anomalous`

# Outline

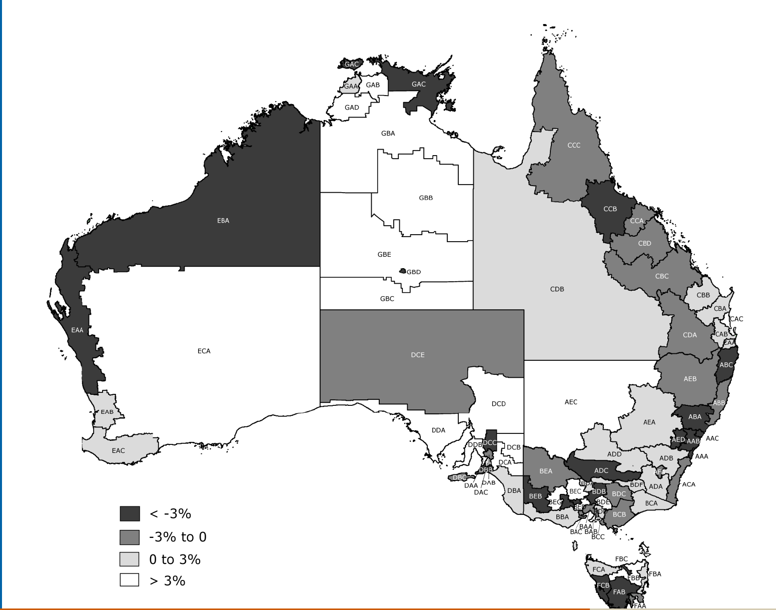
1 Visualization

2 Forecasting

3 Anomaly detection

4 Forecast reconciliation

# Australian tourism



# Australian tourism

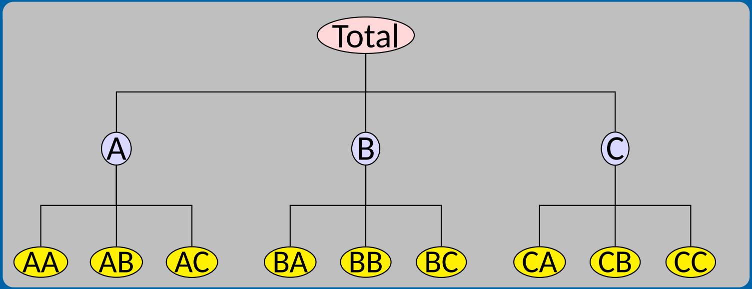
- Quarterly data on visitor night from 1998:Q1 – 2013:Q4
- From: *National Visitor Survey*, based on annual interviews of 120,000 Australians aged 15+, collected by Tourism Research Australia.
- Split by 7 states, 27 zones and 76 regions (a geographical hierarchy)
- Also split by purpose of travel
  - Holiday
  - Visiting friends and relatives (VFR)
  - Business
  - Other
- 304 bottom-level series

# Spectacle sales

- Monthly UK sales data from 2000 – 2014
- Provided by a large spectacle manufacturer
- Split by brand (26), gender (3), price range (6), materials (4), and stores (600)
- About 1 million bottom-level series

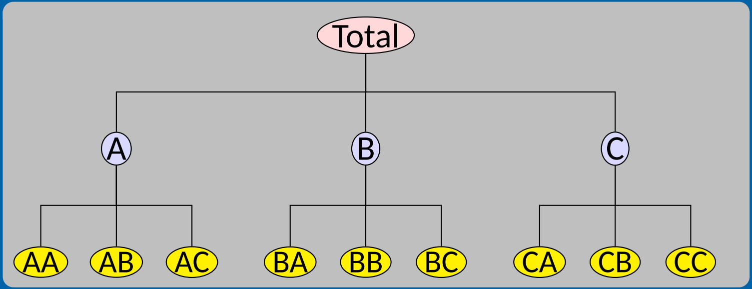
# Hierarchical time series

A **hierarchical time series** is a collection of several time series that are linked together in a hierarchical structure.



# Hierarchical time series

A **hierarchical time series** is a collection of several time series that are linked together in a hierarchical structure.

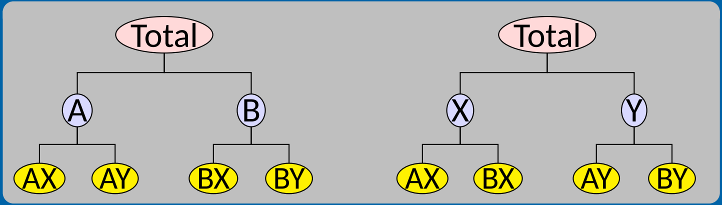



## Examples

- Tourism demand by state and region

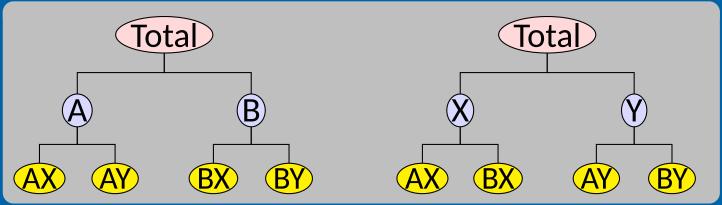
# Grouped time series

A **grouped time series** is a collection of time series that can be grouped together in a number of non-hierarchical ways.



# Grouped time series

A **grouped time series** is a collection of time series that can be grouped together in a number of non-hierarchical ways.




## Examples

- Spectacle sales by brand, gender, stores, etc.
- Tourism by state and purpose of travel

# The problem

1. How to forecast time series at all nodes such that the forecasts add up in the same way as the original data?
2. Can we exploit relationships between the series to improve the forecasts?

# The problem

1. How to forecast time series at all nodes such that the forecasts add up in the same way as the original data?
2. Can we exploit relationships between the series to improve the forecasts?

## The solution

1. Forecast all series at all levels of aggregation using an automatic forecasting algorithm. (e.g., `ets`, `auto.arima`, ...)
2. Reconcile the resulting forecasts so they add up correctly using least squares optimization (i.e., find closest reconciled forecasts to the original forecasts).
3. This is available in the **hts** package in R.

# Hierarchical and grouped time series

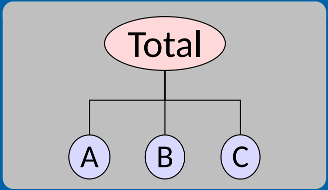
Every collection of time series with aggregation constraints can be written as

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$$

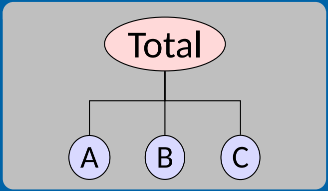
where

- $\boldsymbol{y}_t$ is a vector of all series at time $t$
- $\boldsymbol{b}_t$ is a vector of the most disaggregated series at time $t$
- $\boldsymbol{S}$ is a "summing matrix" containing the aggregation constraints.

# Hierarchical time series



# Hierarchical time series

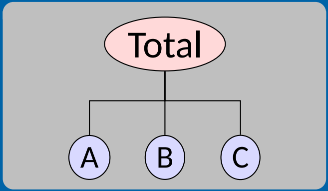



$y_t$ : observed aggregate of all series at time $t$.

$y_{X,t}$ : observation on series $X$ at time $t$.

$\boldsymbol{b}_t$ : vector of all series at bottom level in time $t$.

# Hierarchical time series



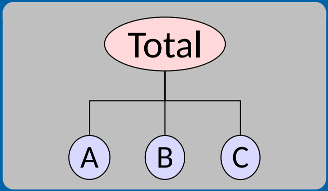

$y_t$ : observed aggregate of all series at time $t$.

$y_{X,t}$ : observation on series $X$ at time $t$.

$\boldsymbol{b}_t$ : vector of all series at bottom level in time $t$.

$$\boldsymbol{y}_t = \begin{pmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{C,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} Y_{A,t} \\ y_{B,t} \\ y_{C,t} \end{pmatrix}$$

# Hierarchical time series




$y_t$ : observed aggregate of all series at time $t$.

$y_{X,t}$ : observation on series $X$ at time $t$.

$\boldsymbol{b}_t$ : vector of all series at bottom level in time $t$.

$$\boldsymbol{y}_t = \begin{pmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{C,t} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{S}} \underbrace{\begin{pmatrix} y_{A,t} \\ y_{B,t} \\ y_{C,t} \end{pmatrix}}_{\boldsymbol{b}_t}$$

$$\boxed{\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t}$$

# Forecasting notation

Let $\hat{\boldsymbol{y}}_n(h)$ be vector of initial $h$-step forecasts, made at time $n$, stacked in same order as $\boldsymbol{y}_t$.

# Forecasting notation

Let $\hat{\boldsymbol{y}}_n(h)$ be vector of initial $h$-step forecasts, made at time $n$, stacked in same order as $\boldsymbol{y}_t$.
(In general, they will not “add up”.)

# Forecasting notation

Let $\hat{\mathbf{y}}_n(h)$ be vector of initial $h$-step forecasts, made at time $n$, stacked in same order as $\mathbf{y}_t$.
(In general, they will not "add up".)

Reconciled forecasts must be of the form:

$$\tilde{\mathbf{y}}_n(h) = \mathbf{S}\mathbf{G}\hat{\mathbf{y}}_n(h)$$

for some matrix $\mathbf{G}$.

# Forecasting notation

Let $\hat{\boldsymbol{y}}_n(h)$ be vector of initial $h$-step forecasts, made at time $n$, stacked in same order as $\boldsymbol{y}_t$.
(In general, they will not "add up".)

Reconciled forecasts must be of the form:

$$\tilde{\boldsymbol{y}}_n(h) = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_n(h)$$

for some matrix $\boldsymbol{G}$.

- $\boldsymbol{G}$ extracts and combines base forecasts $\hat{\boldsymbol{y}}_n(h)$ to get bottom-level forecasts.
- $\boldsymbol{S}$ adds them up

# Optimal combination forecasts

**Main result**

The best (minimum sum of variances) unbiased forecasts are obtained when $\boldsymbol{G} = (\boldsymbol{S}'\Sigma_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\Sigma_h^{-1}$, where $\Sigma_h$ is the $h$-step base forecast error covariance matrix.

# Optimal combination forecasts

**Main result**

The best (minimum sum of variances) unbiased forecasts are obtained when $\boldsymbol{G} = (\boldsymbol{S}'\Sigma_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\Sigma_h^{-1}$, where $\Sigma_h$ is the $h$-step base forecast error covariance matrix.

$$\tilde{\boldsymbol{y}}_n(h) = \boldsymbol{S}(\boldsymbol{S}'\Sigma_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\Sigma_h^{-1}\hat{\boldsymbol{y}}_n(h)$$

**Problem:** $\Sigma_h$ hard to estimate, especially for $h > 1$.

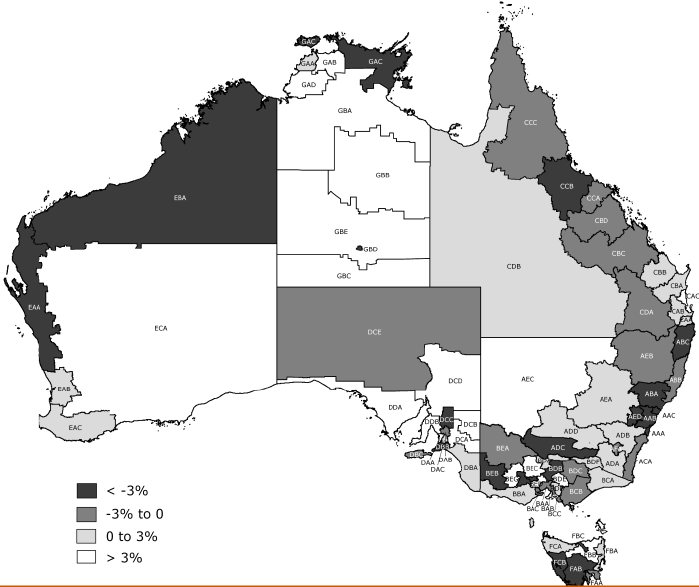
**Solutions:**

- Ignore $\Sigma_h$ (OLS)
- Assume $\Sigma_h$ diagonal (WLS) [Default in `hts`]
- Try to estimate $\Sigma_h$ (GLS)

# Features

- Covariates can be included in initial forecasts.
- Adjustments can be made to initial forecasts at any level.
- Very simple and flexible method. Can work with *any* hierarchical or grouped time series.
- Conceptually easy to implement: regression of base forecasts on structure matrix.

# Australian tourism

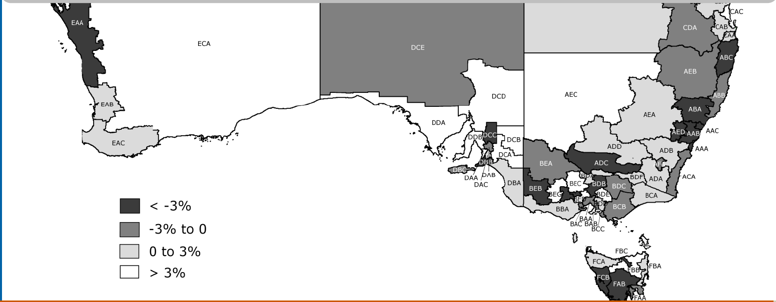


# Australian tourism

## Domestic visitor nights

Quarterly data: 1998 – 2006.

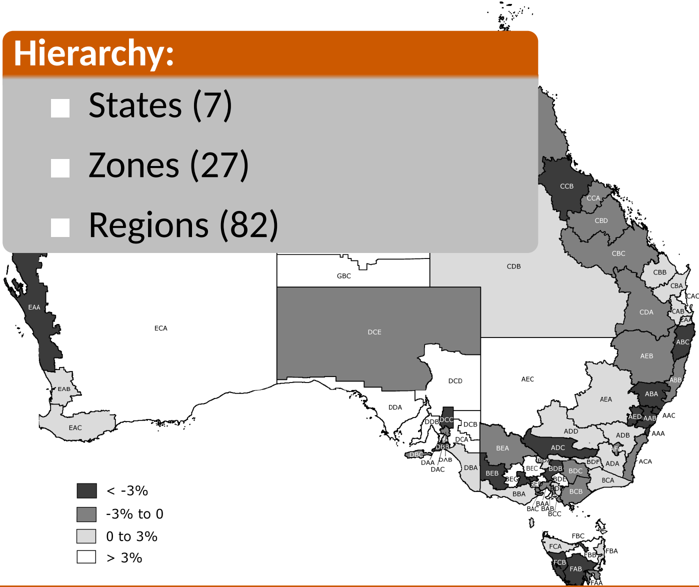
From: *National Visitor Survey*, based on annual interviews of 120,000 Australians aged 15+, collected by Tourism Research Australia.



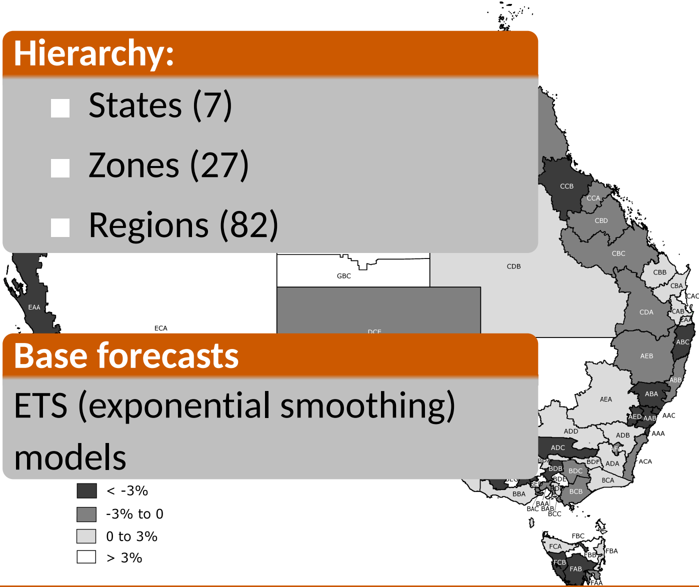
# Australian tourism

## Hierarchy:

- States (7)
- Zones (27)
- Regions (82)



# Australian tourism



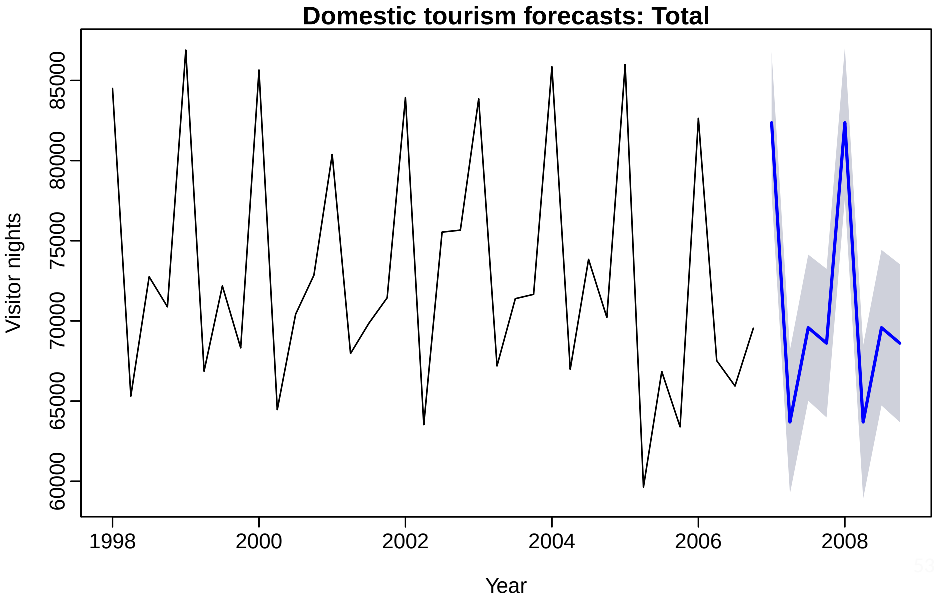

## Hierarchy:

- States (7)
- Zones (27)
- Regions (82)

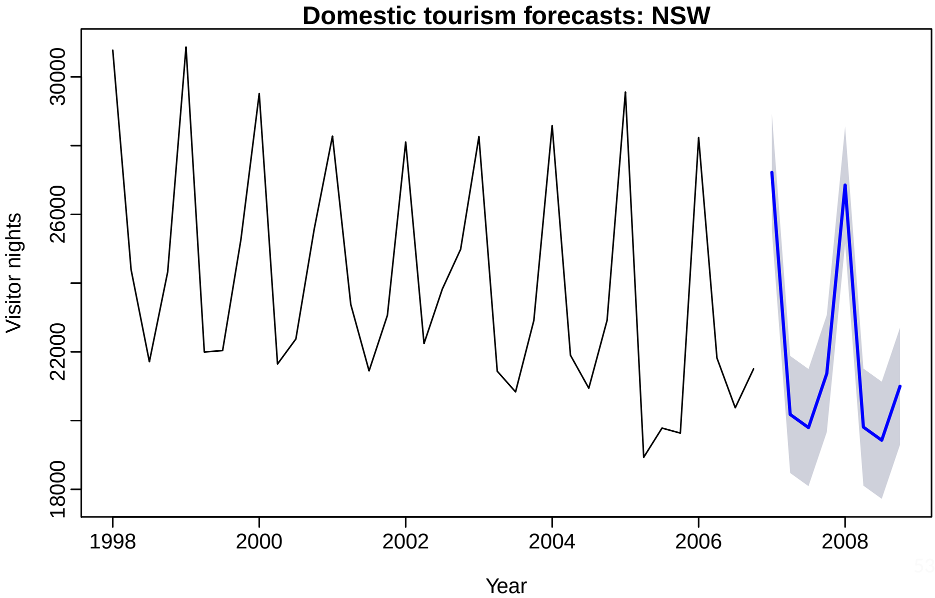
## Base forecasts

ETS (exponential smoothing) models

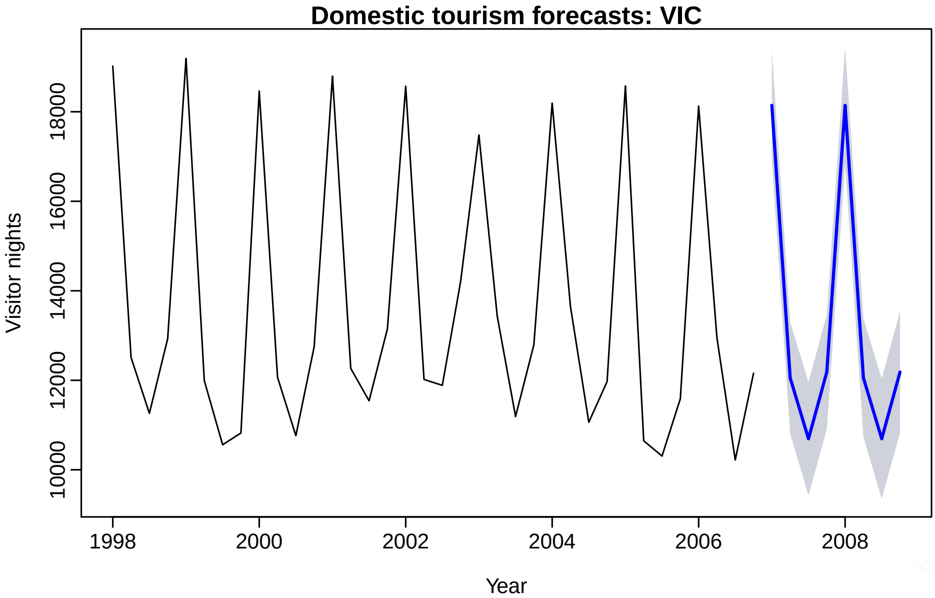
# Base forecasts



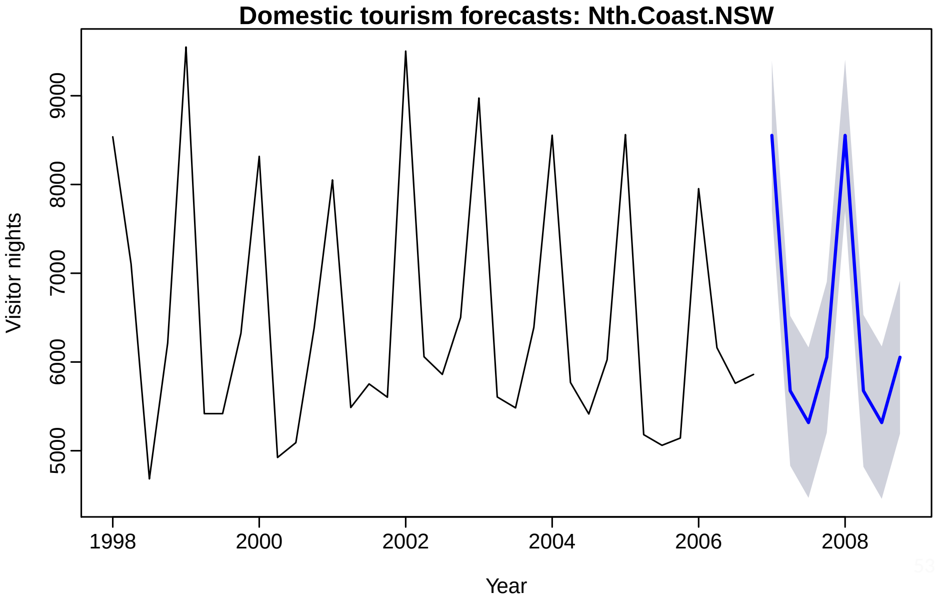
# Base forecasts



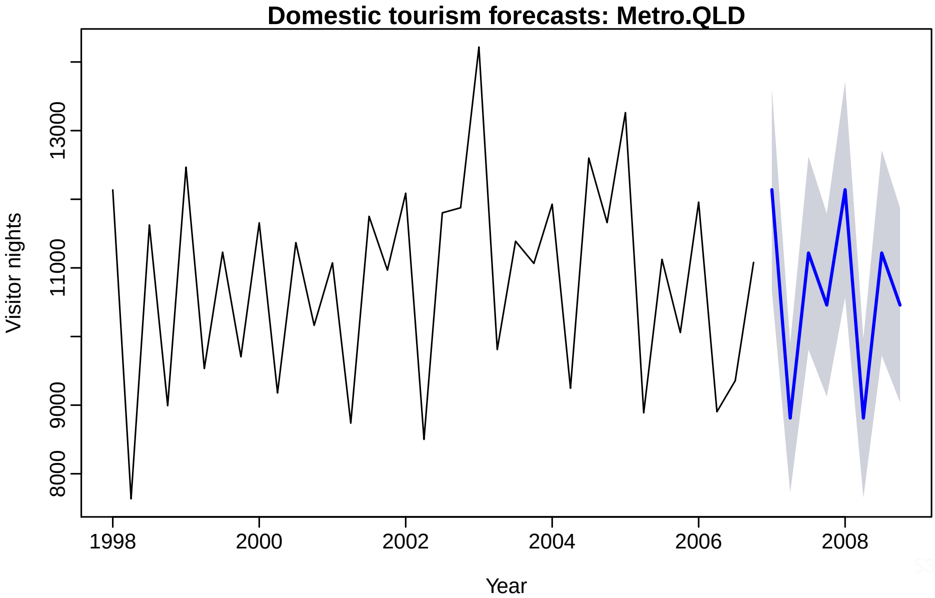
# Base forecasts



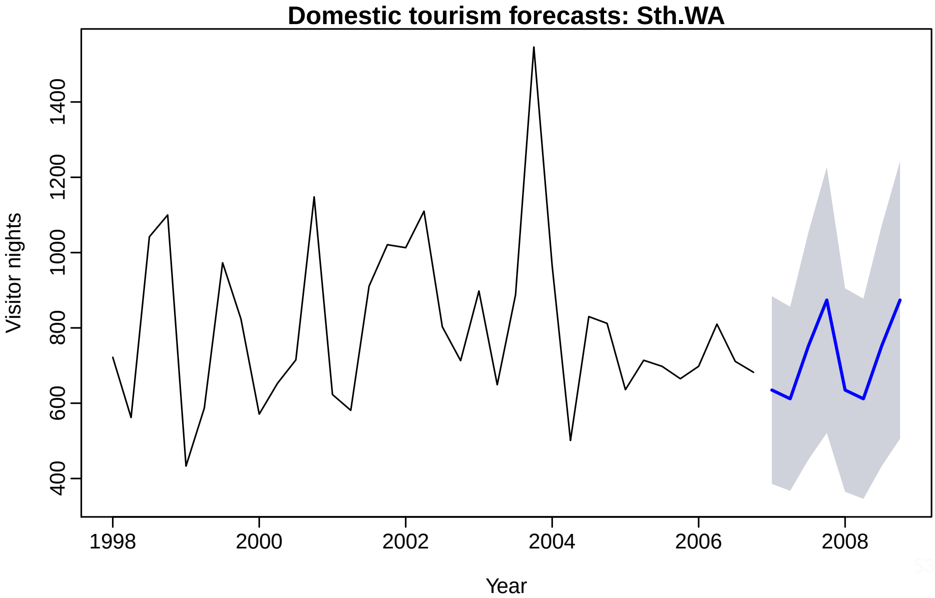
# Base forecasts



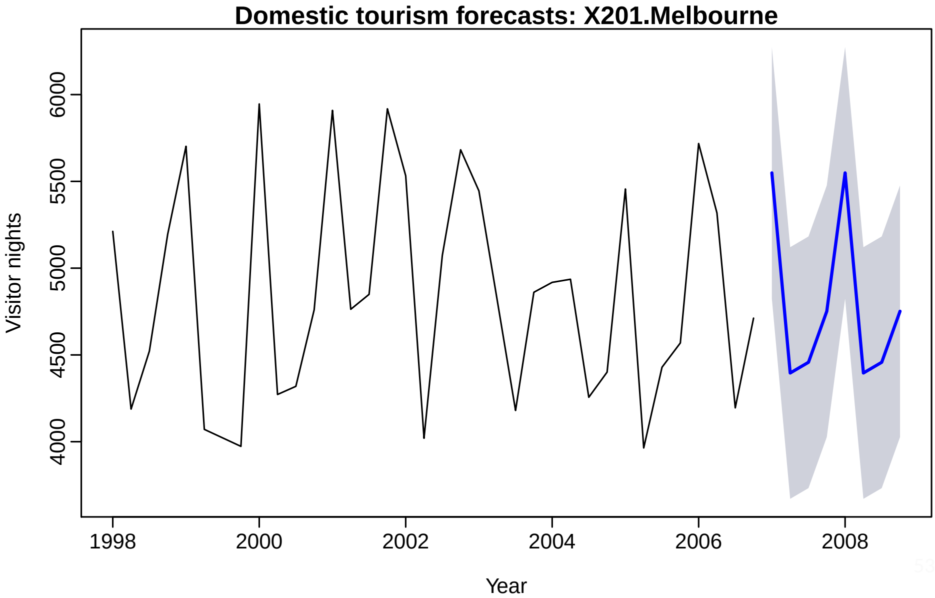
# Base forecasts



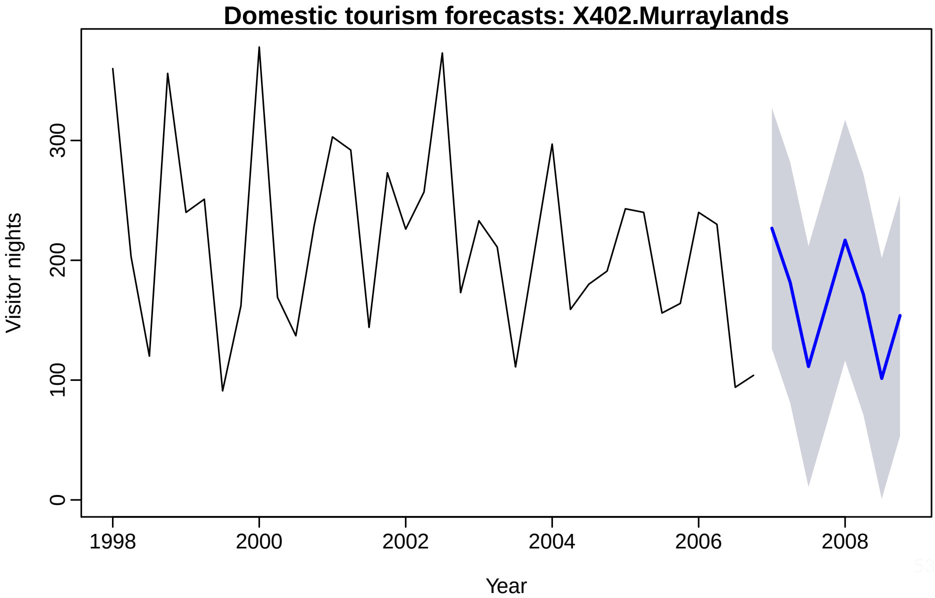
# Base forecasts



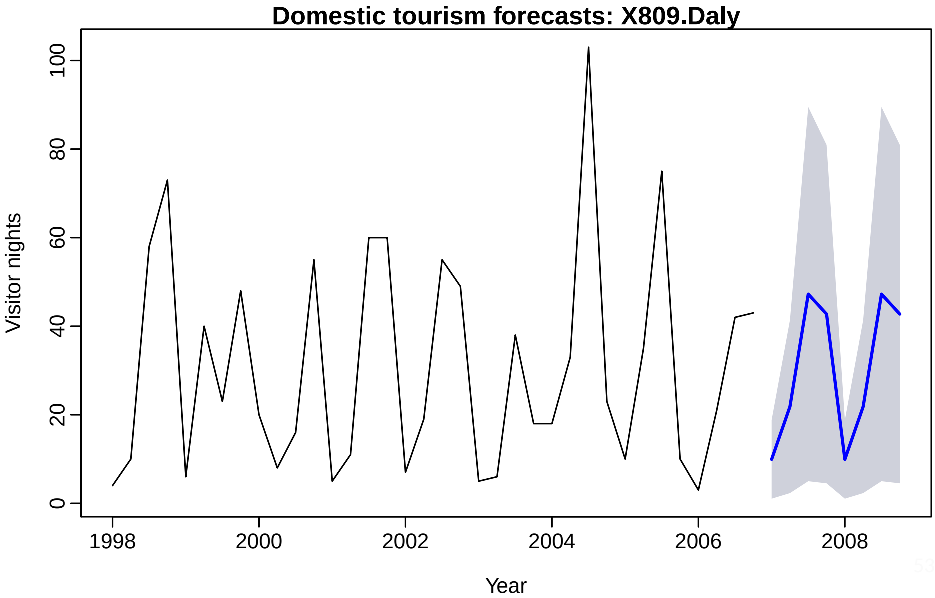
# Base forecasts



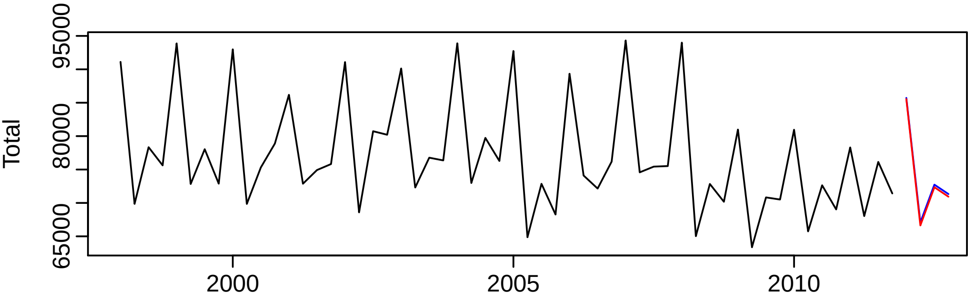
# Base forecasts



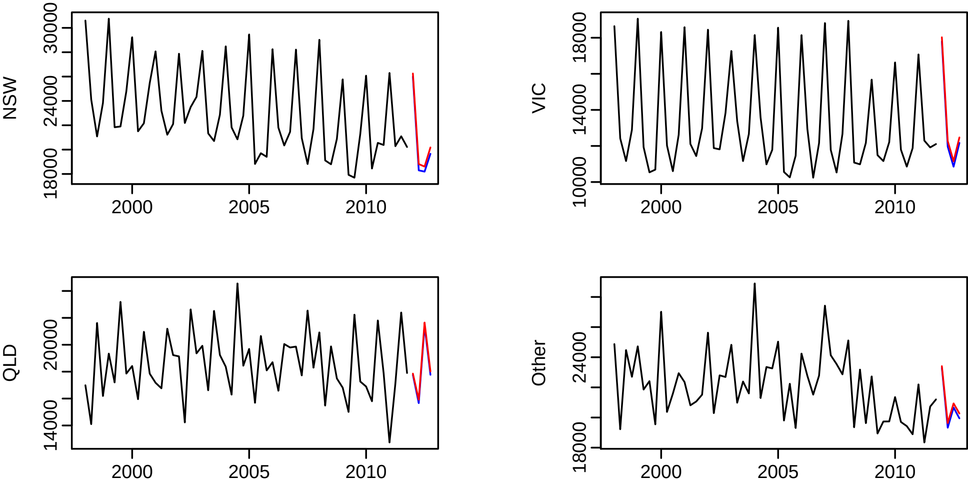
# Base forecasts



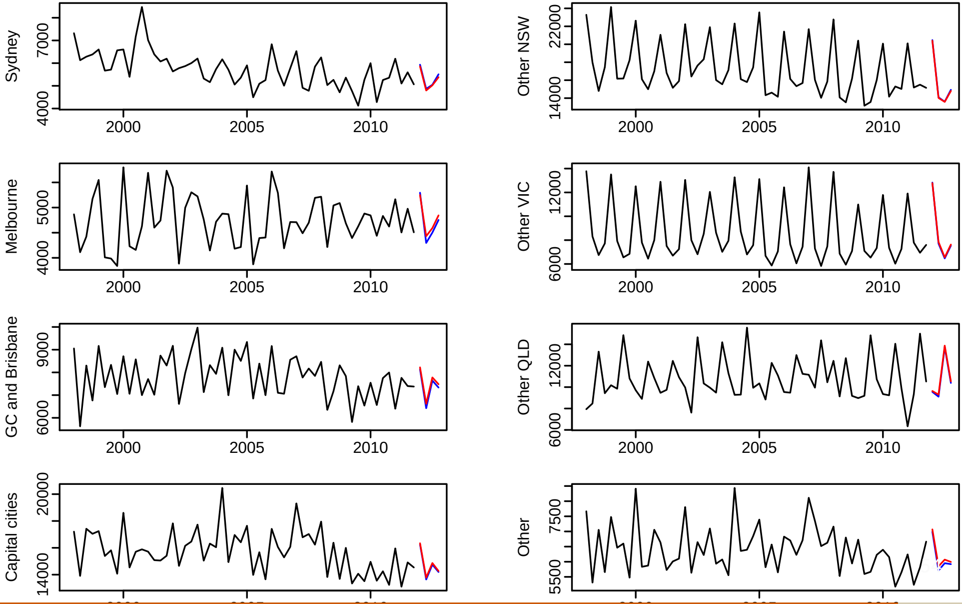
# Reconciled forecasts



# Reconciled forecasts



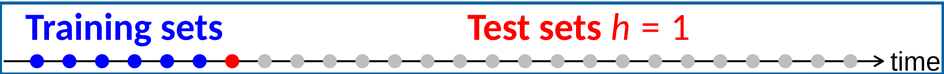
# Reconciled forecasts



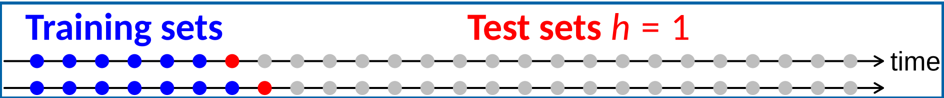
# Forecast evaluation

- Select models using all observations;
- Re-estimate models using first 12 observations and generate 1- to 8-step-ahead forecasts;
- Increase sample size one observation at a time, re-estimate models, generate forecasts until the end of the sample;
- In total 24 1-step-ahead, 23 2-steps-ahead, up to 17 8-steps-ahead for forecast evaluation.

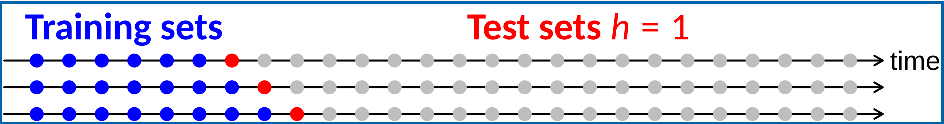
# Forecast evaluation



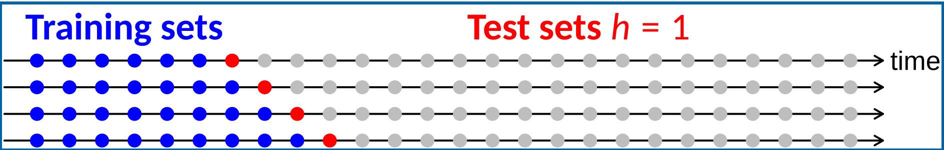
# Forecast evaluation



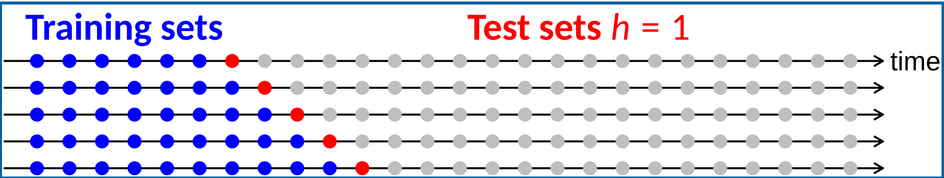
# Forecast evaluation



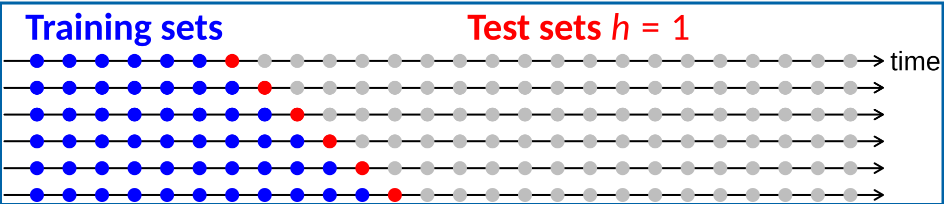
# Forecast evaluation



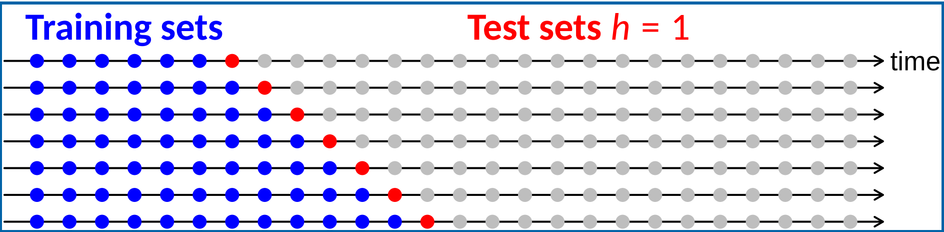
# Forecast evaluation



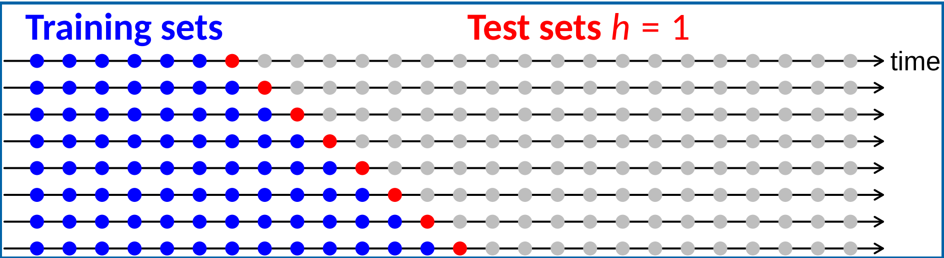
# Forecast evaluation



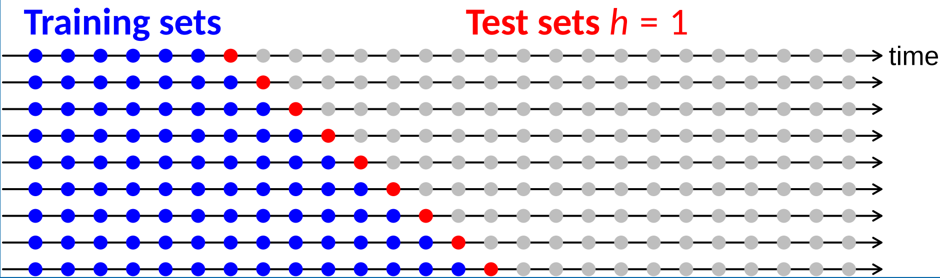
# Forecast evaluation



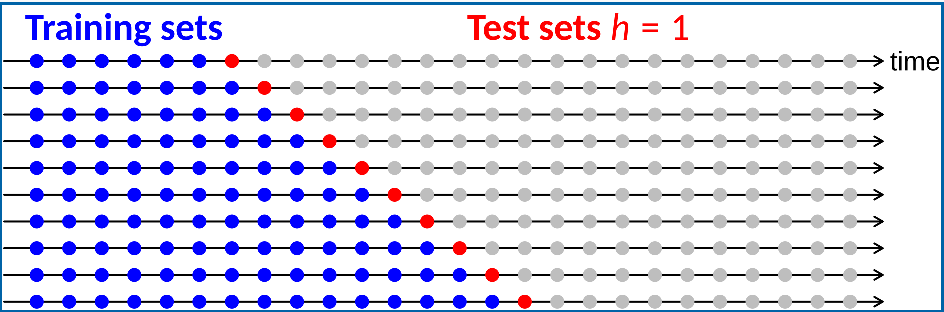
# Forecast evaluation



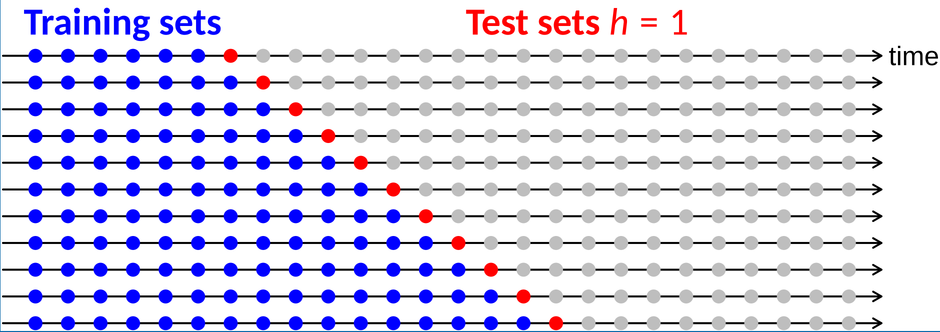
# Forecast evaluation



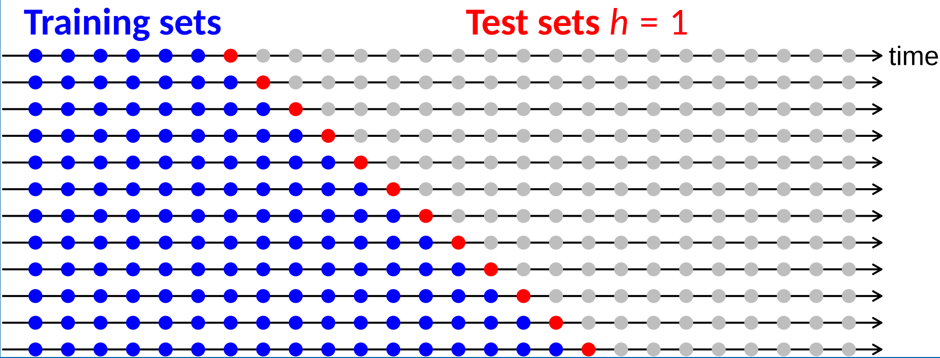
# Forecast evaluation



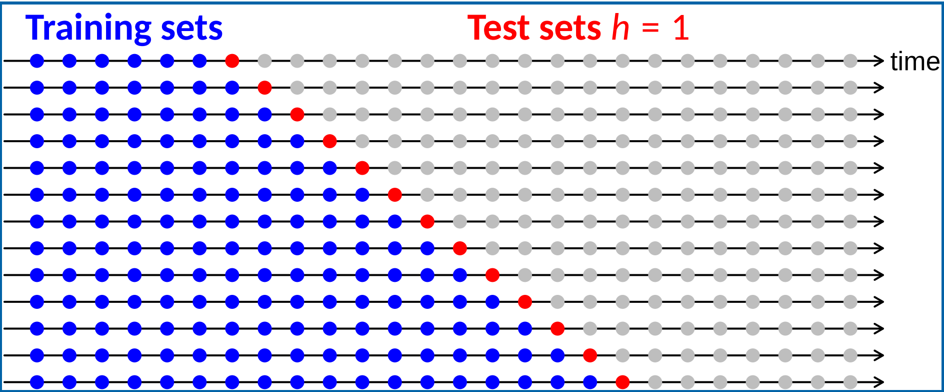
# Forecast evaluation



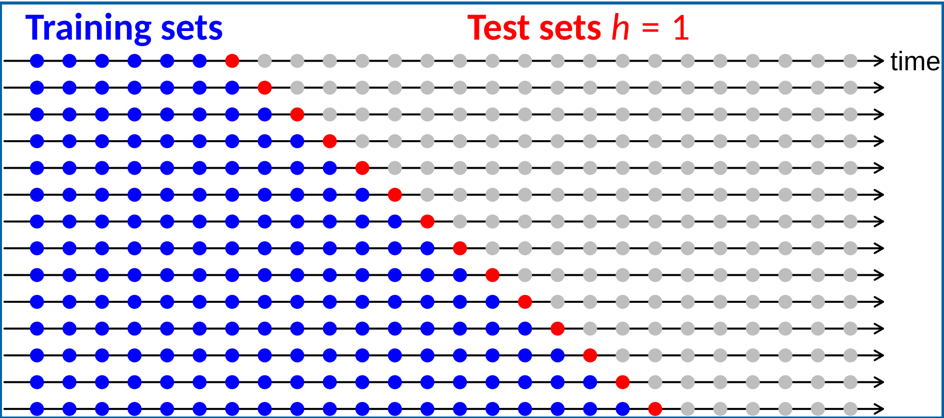
# Forecast evaluation



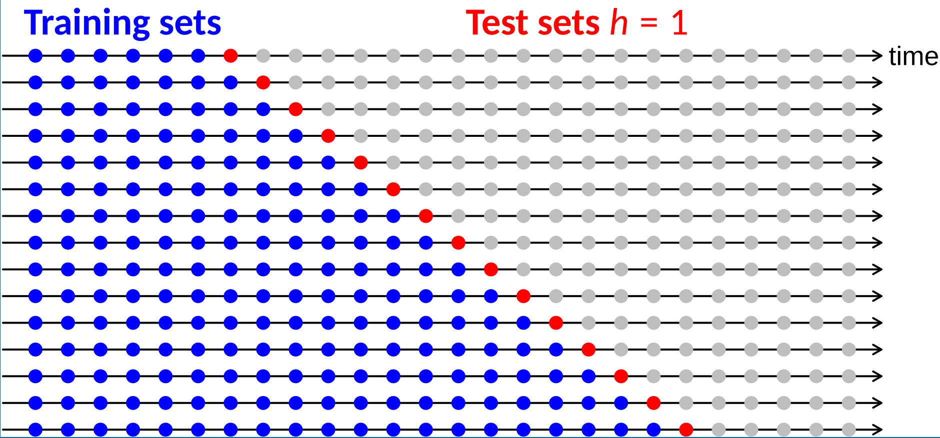
# Forecast evaluation



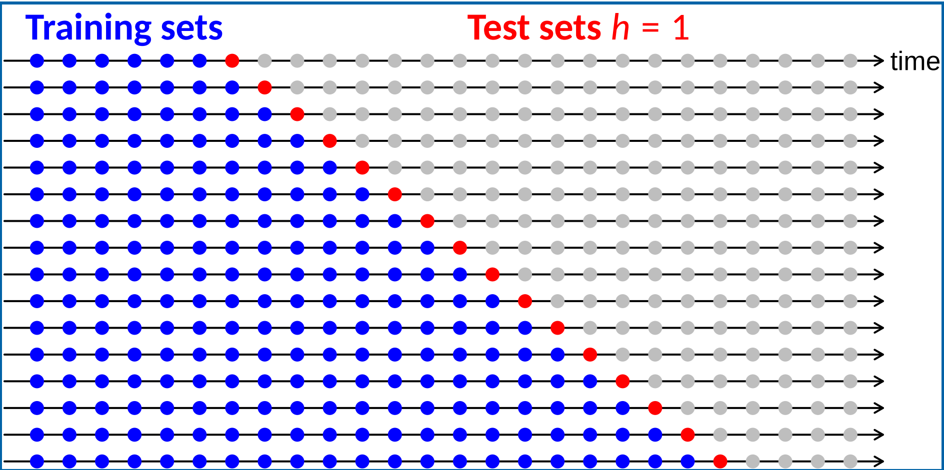
# Forecast evaluation



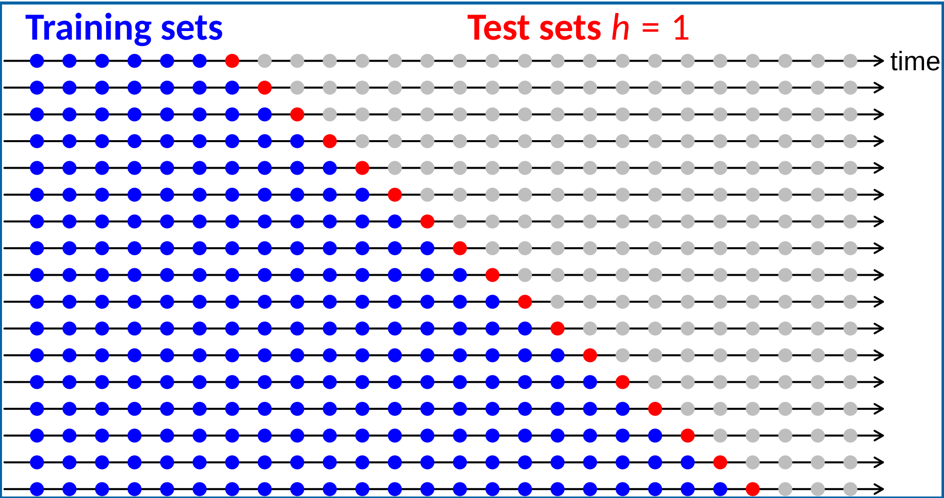
# Forecast evaluation



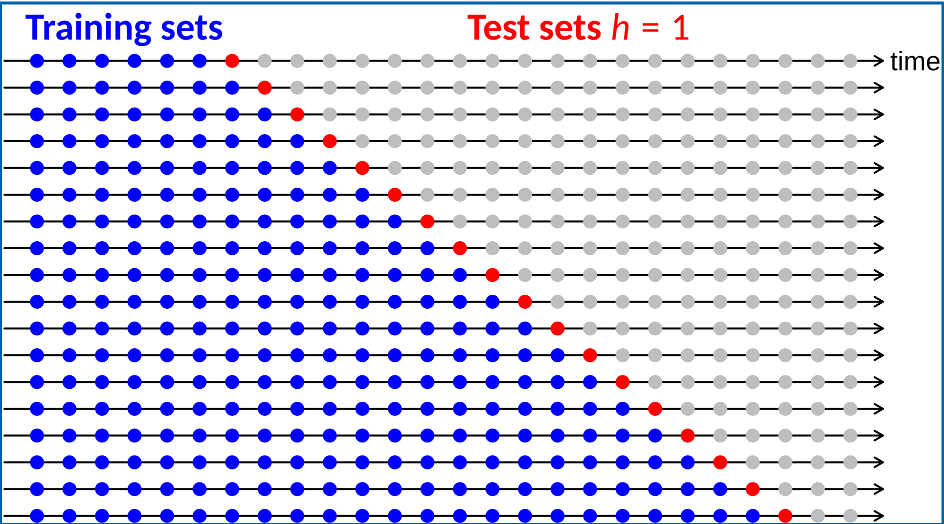
# Forecast evaluation



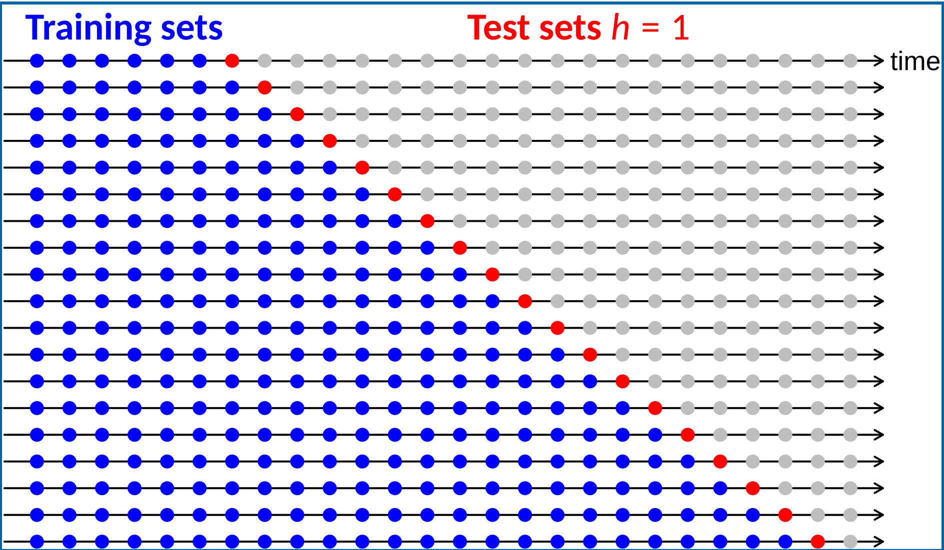
# Forecast evaluation



# Forecast evaluation

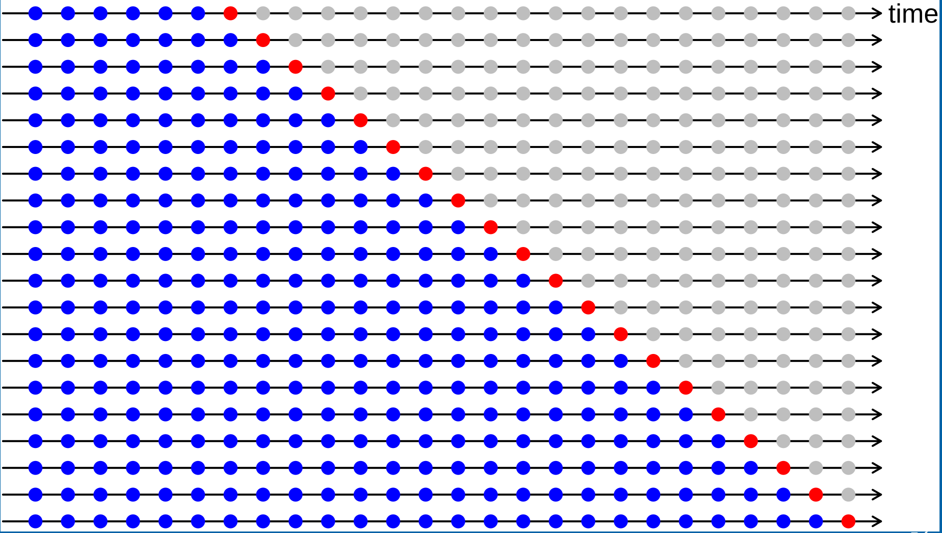


# Forecast evaluation

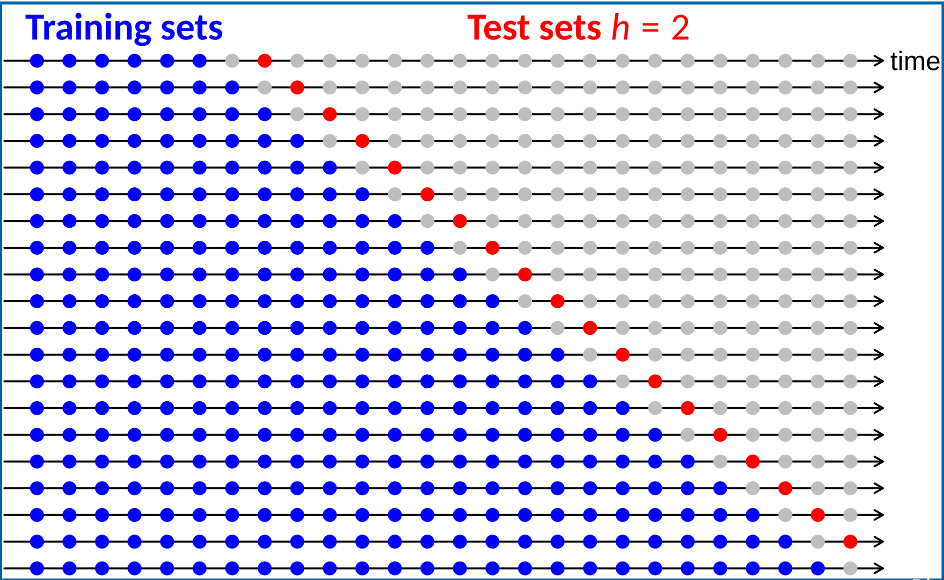


# Forecast evaluation

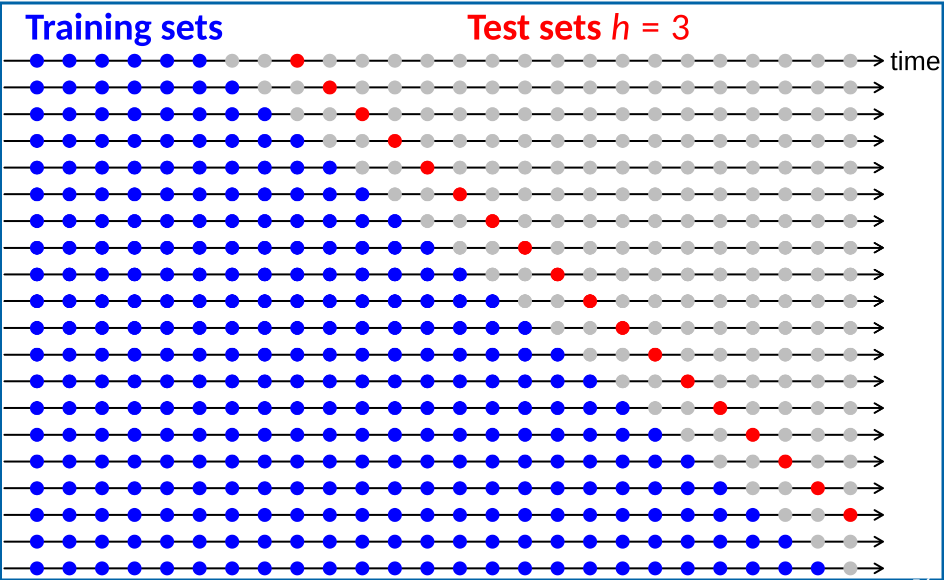
**Training sets** **Test sets $h$ = 1**



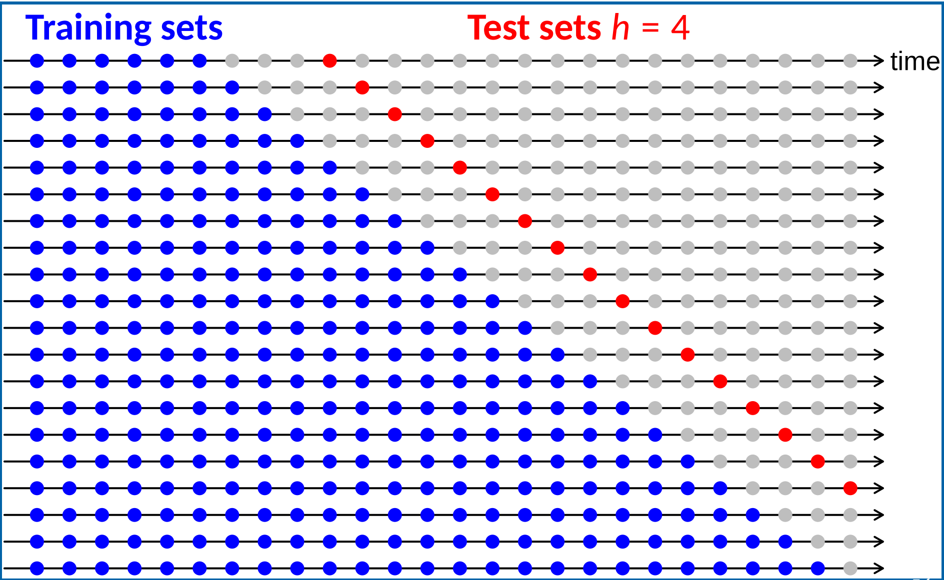
# Forecast evaluation



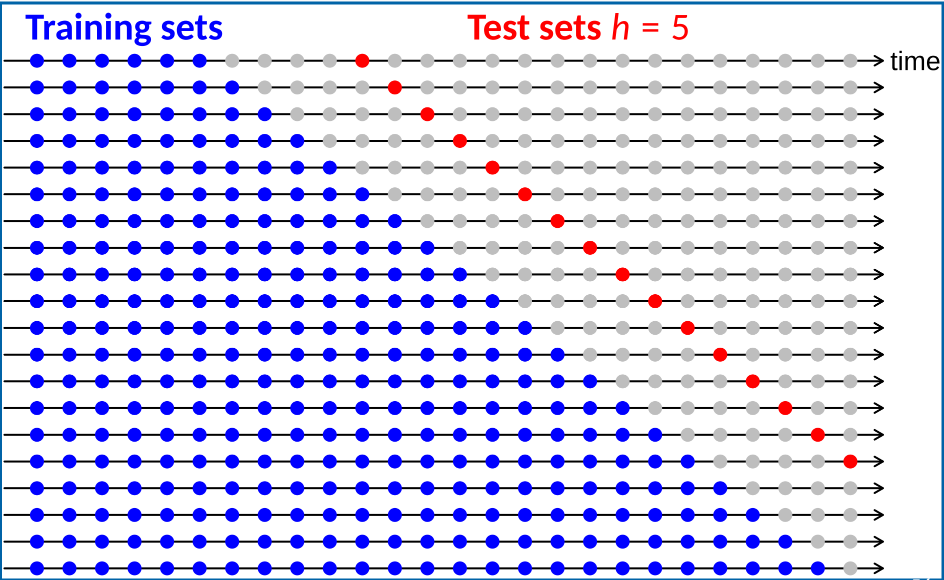
# Forecast evaluation



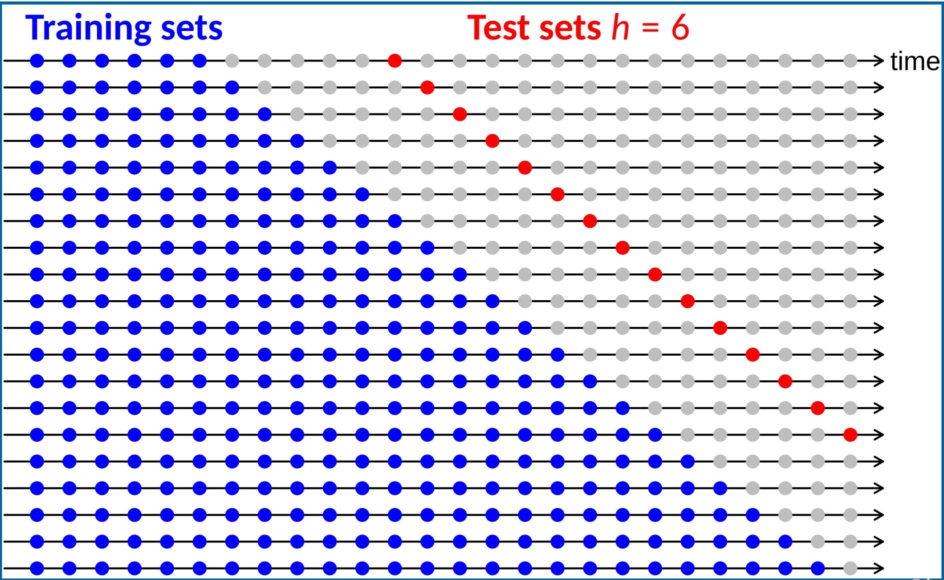
# Forecast evaluation



# Forecast evaluation



# Forecast evaluation



# Hierarchy: states, zones, regions

| RMSE | $h = 1$ | $h = 2$ | $h = 3$ | $h = 4$ | $h = 5$ | $h = 6$ | Ave |
|---|---|---|---|---|---|---|---|
| | | | Forecast horizon | | | | |
| **Australia** | | | | | | | |
| Base | 1762.04 | 1770.29 | 1766.02 | 1818.82 | 1705.35 | 1721.17 | **1757.28** |
| Bottom | 1736.92 | 1742.69 | 1722.79 | 1752.74 | 1666.73 | 1687.43 | **1718.22** |
| WLS | 1705.21 | 1715.87 | 1703.75 | 1729.56 | 1627.79 | 1661.24 | **1690.57** |
| GLS | 1704.64 | 1715.60 | 1705.31 | 1729.04 | 1626.36 | 1661.64 | **1690.43** |
| **States** | | | | | | | |
| Base | 399.77 | 404.16 | 401.92 | 407.26 | 395.38 | 401.17 | **401.61** |
| Bottom | 404.29 | 406.95 | 404.96 | 409.02 | 399.80 | 401.55 | **404.43** |
| WLS | 398.84 | 402.12 | 400.71 | 405.03 | 394.76 | 398.23 | **399.95** |
| GLS | 398.84 | 402.16 | 400.86 | 405.03 | 394.59 | 398.22 | **399.95** |
| **Regions** | | | | | | | |
| Base | 93.15 | 93.38 | 93.45 | 93.79 | 93.50 | 93.56 | **93.47** |
| Bottom | 93.15 | 93.38 | 93.45 | 93.79 | 93.50 | 93.56 | **93.47** |
| WLS | 93.02 | 93.32 | 93.38 | 93.72 | 93.39 | 93.53 | **93.39** |
| GLS | 92.98 | 93.27 | 93.34 | 93.66 | 93.34 | 93.46 | **93.34** |

# Acknowledgments

Dilini Talagala

Kate Smith-Miles

George Athanasopoulos

Earo Wang

Thiyanga Talagala

Pablo Montero-Manso

Mitchell O’Hara-Wild

Yanfei Kang

Shanika Wickramasuriya