MONASH University

MONASH BUSINESS SCHOOL

Feature-based time series analysis

Rob J Hyndman

21 June 2018

Outline

1. Time series features
2. Finding anomalies
3. Irish smart metre data
4. Finding typical and unusual households
5. Visualization via embedding

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

petition: results, conclusions a ns

Spyros Makridakis, Michèle Hibon*

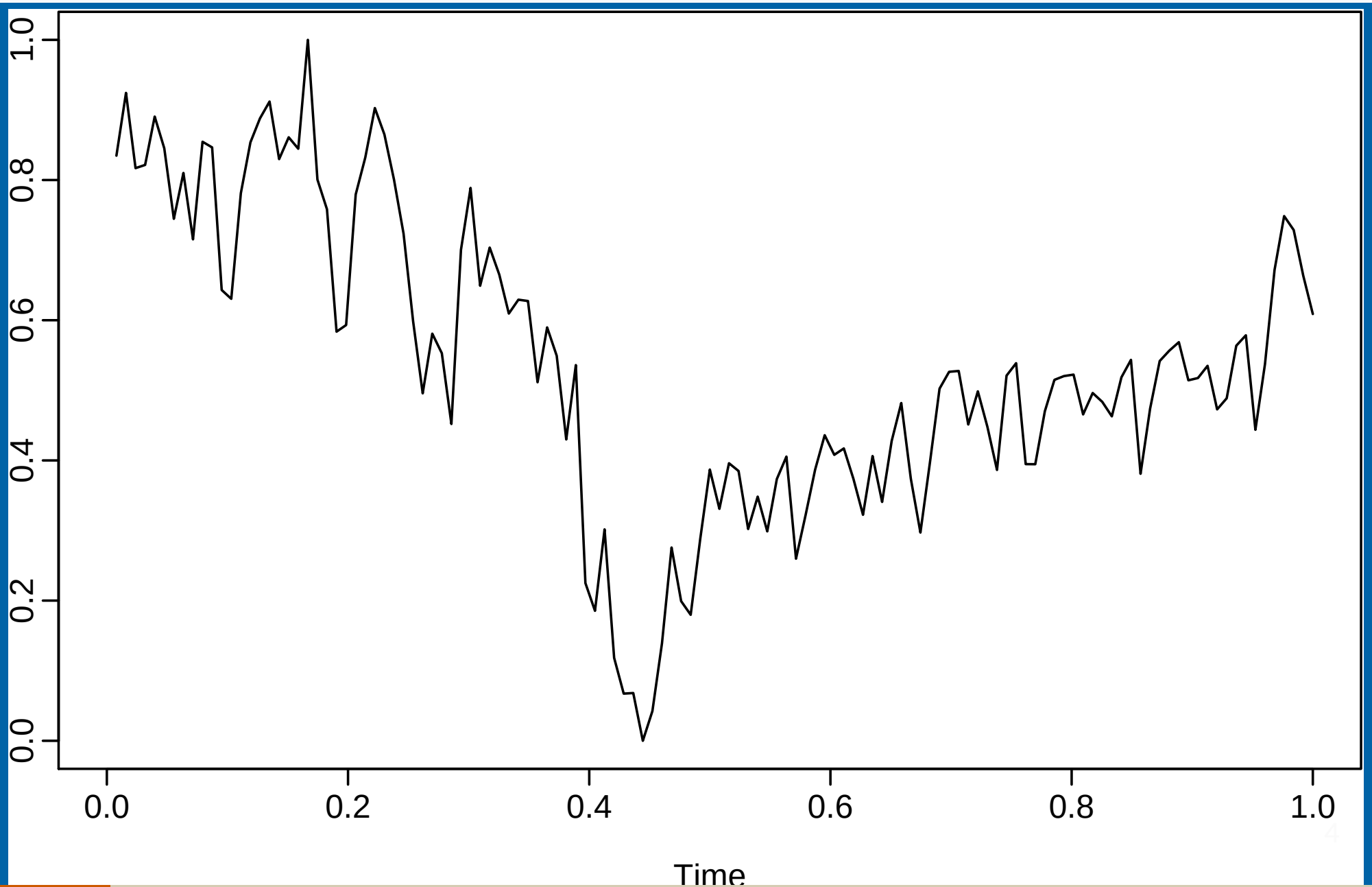
SEAD, Boulevard de Constance, 77305 Fontainebleau, Fr

Abstr

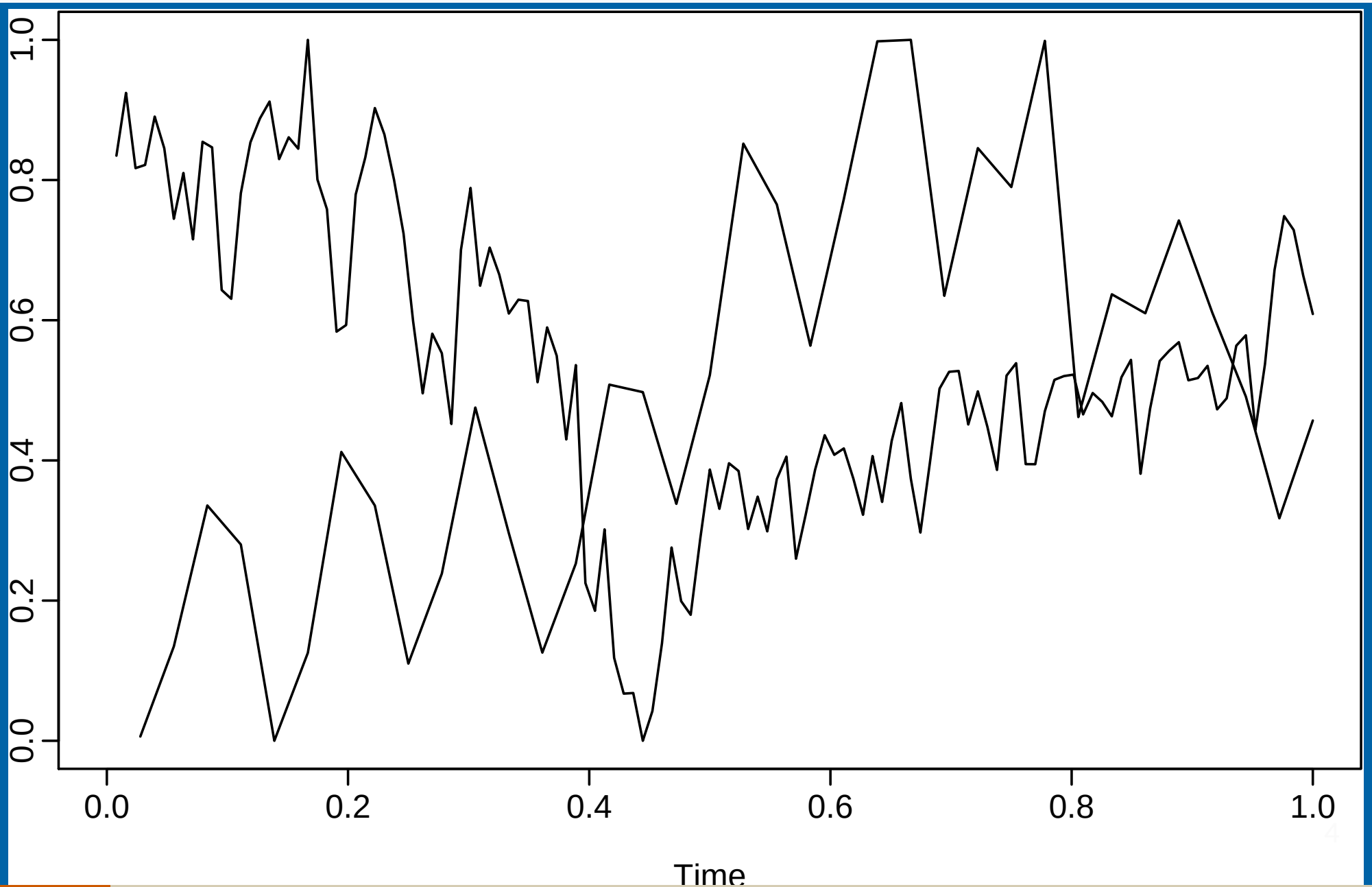
This paper describes the M3-Competition, the latest of the M-Competitions. It explains the reasons for conducting the competition and summarizes its results and conclusions. In addition, the paper compares such results/conclusions with those of the previous two M-Competitions as well as with those of 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

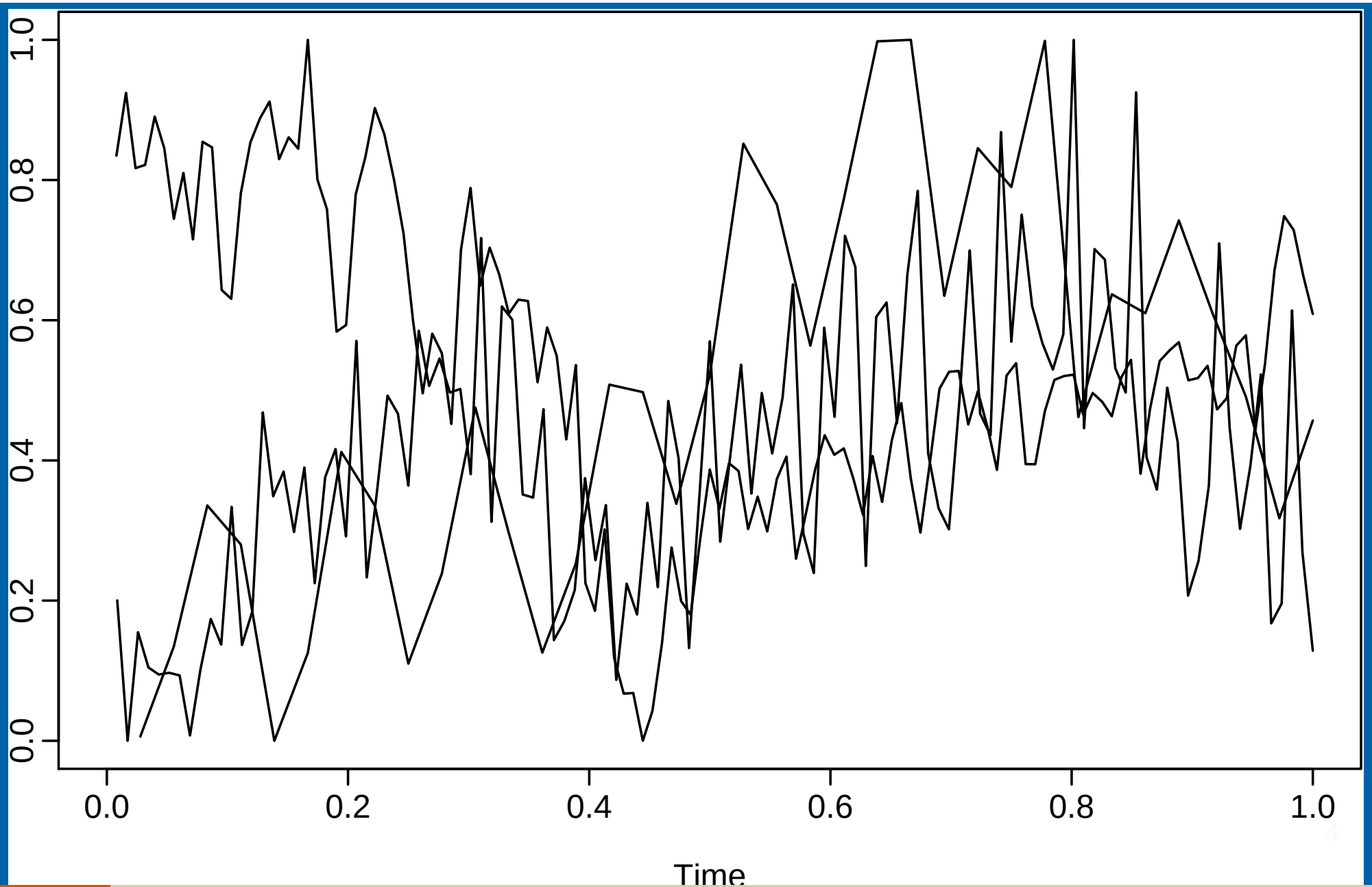
How to plot lots of time series?



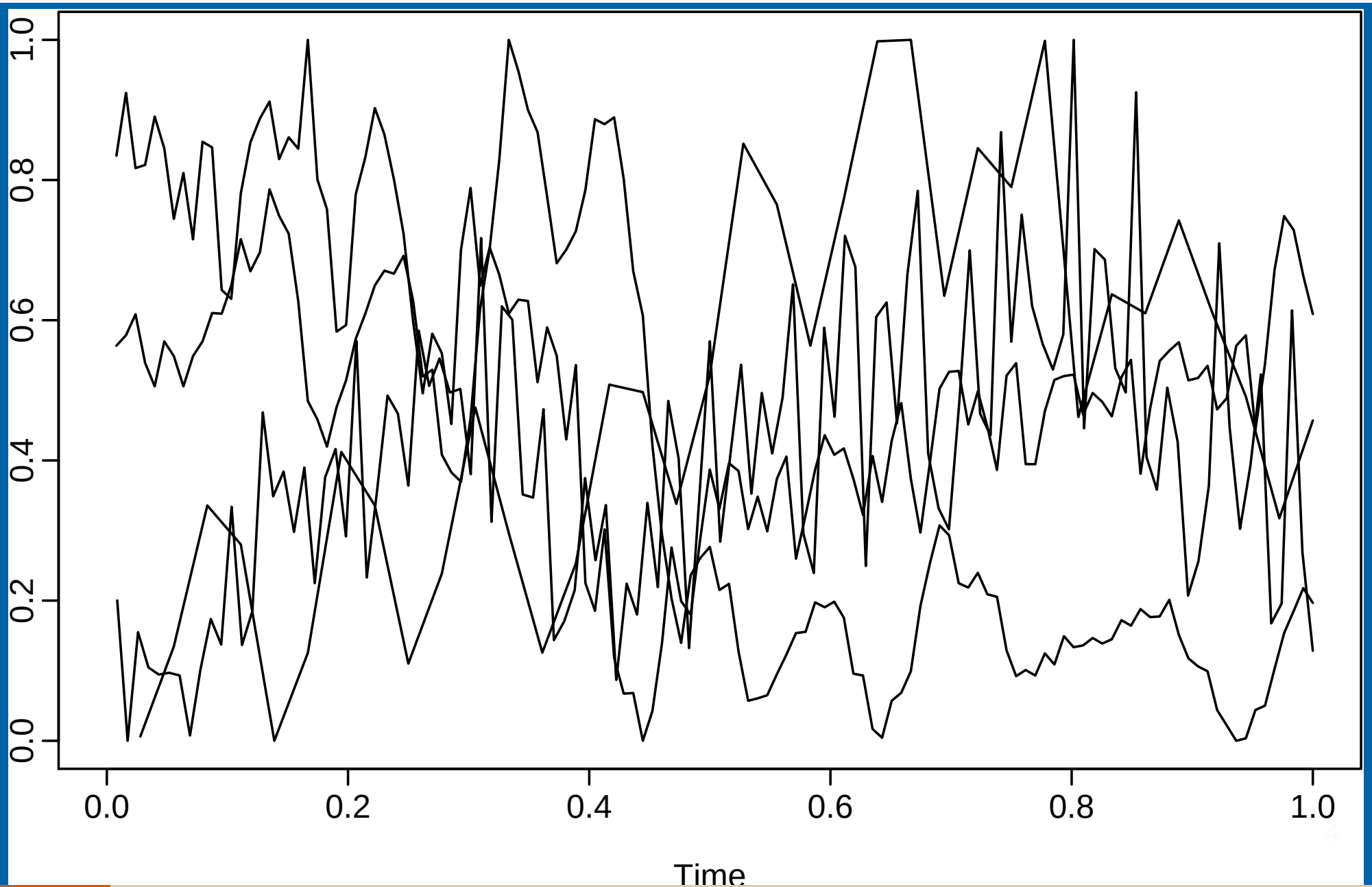
How to plot lots of time series?



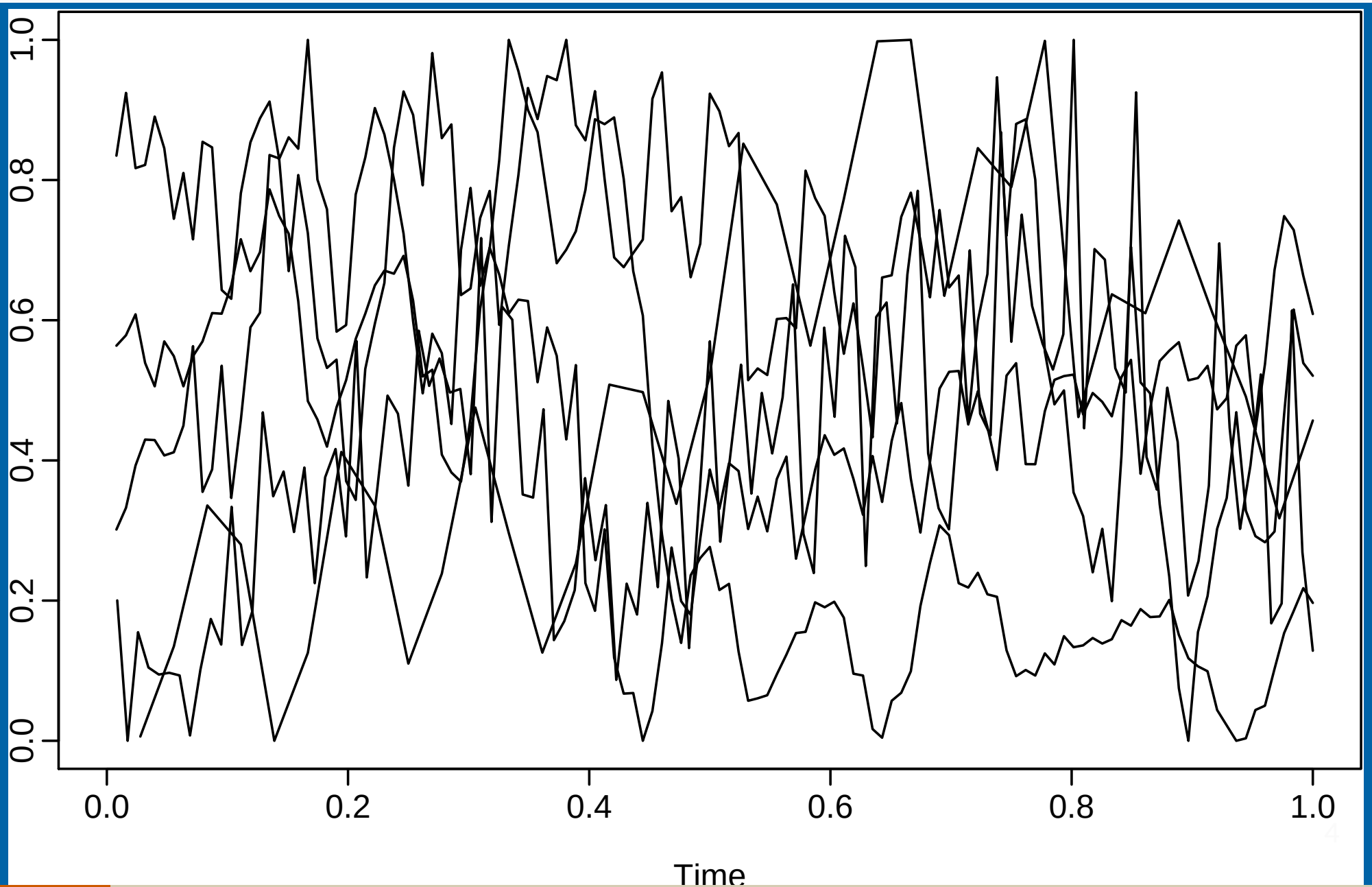
How to plot lots of time series?



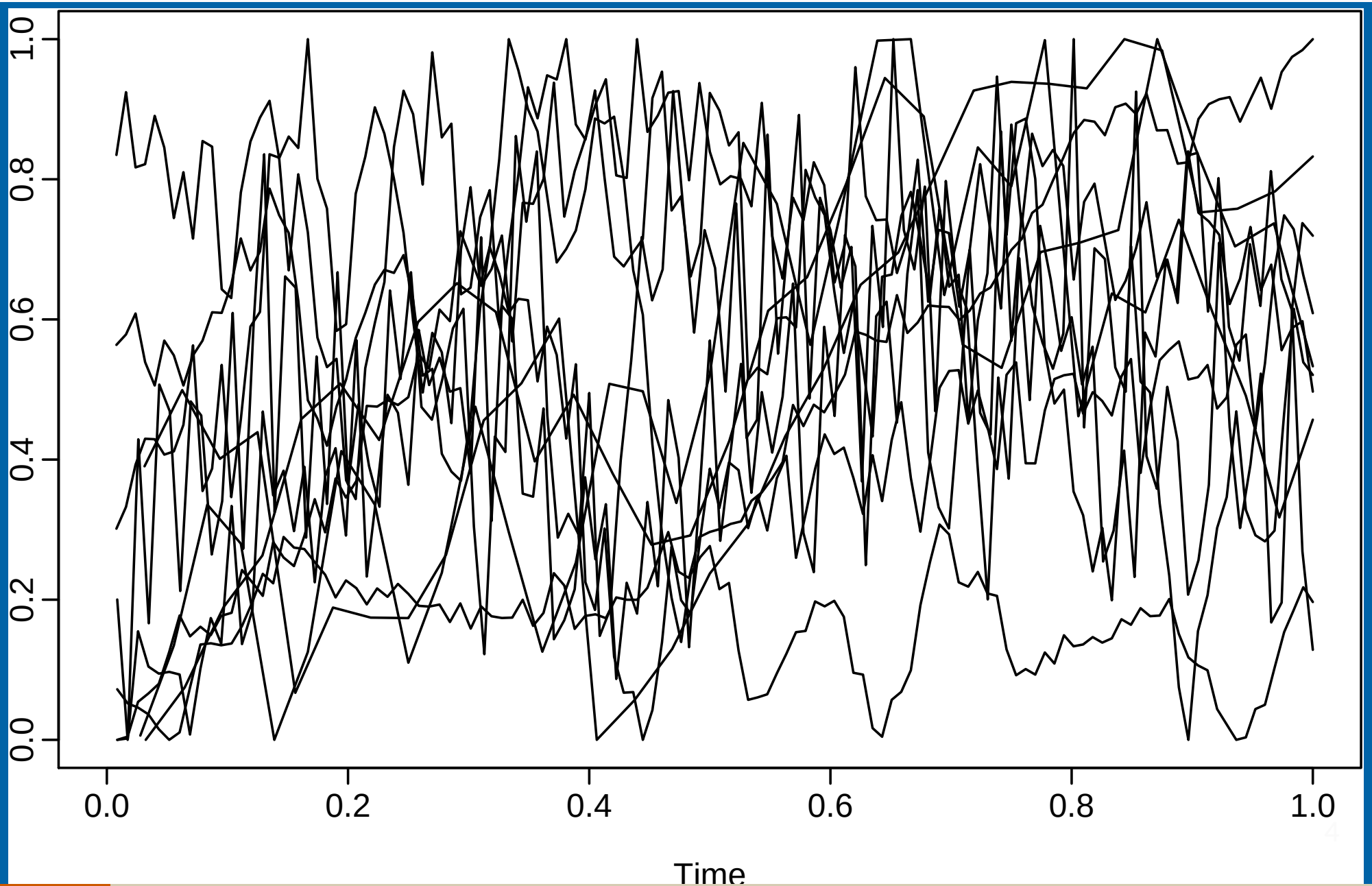
How to plot lots of time series?



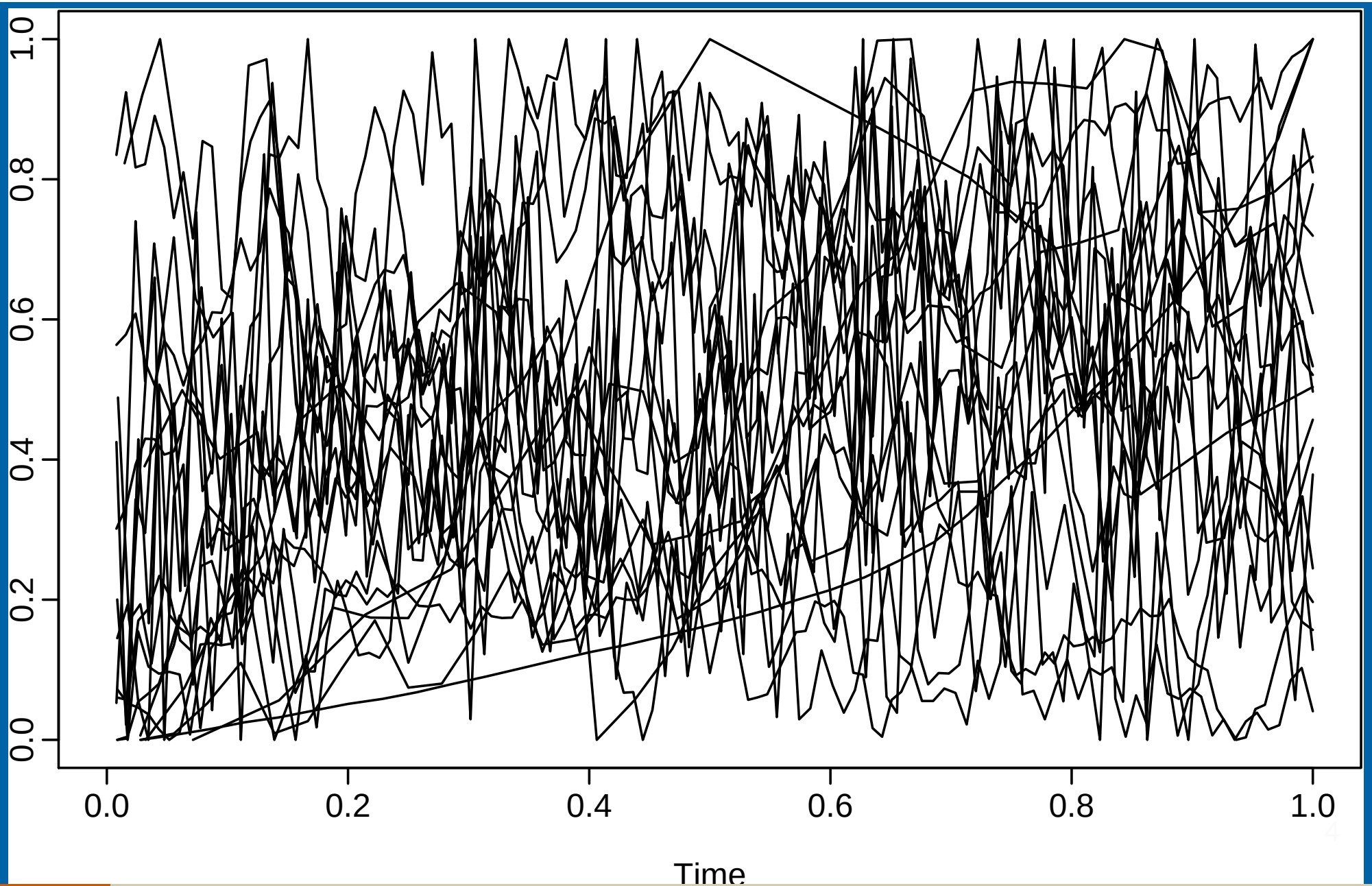
How to plot lots of time series?



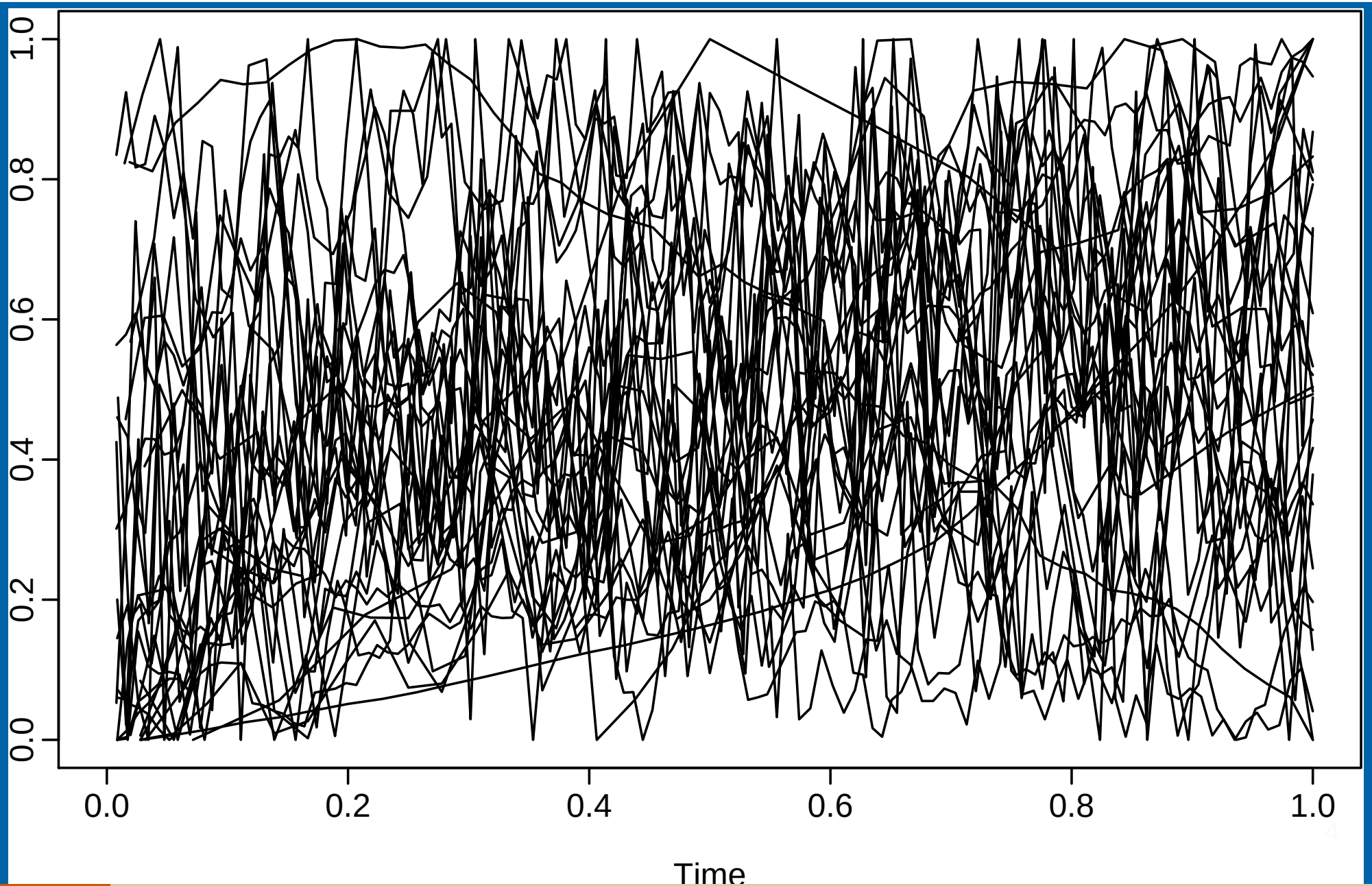
How to plot lots of time series?



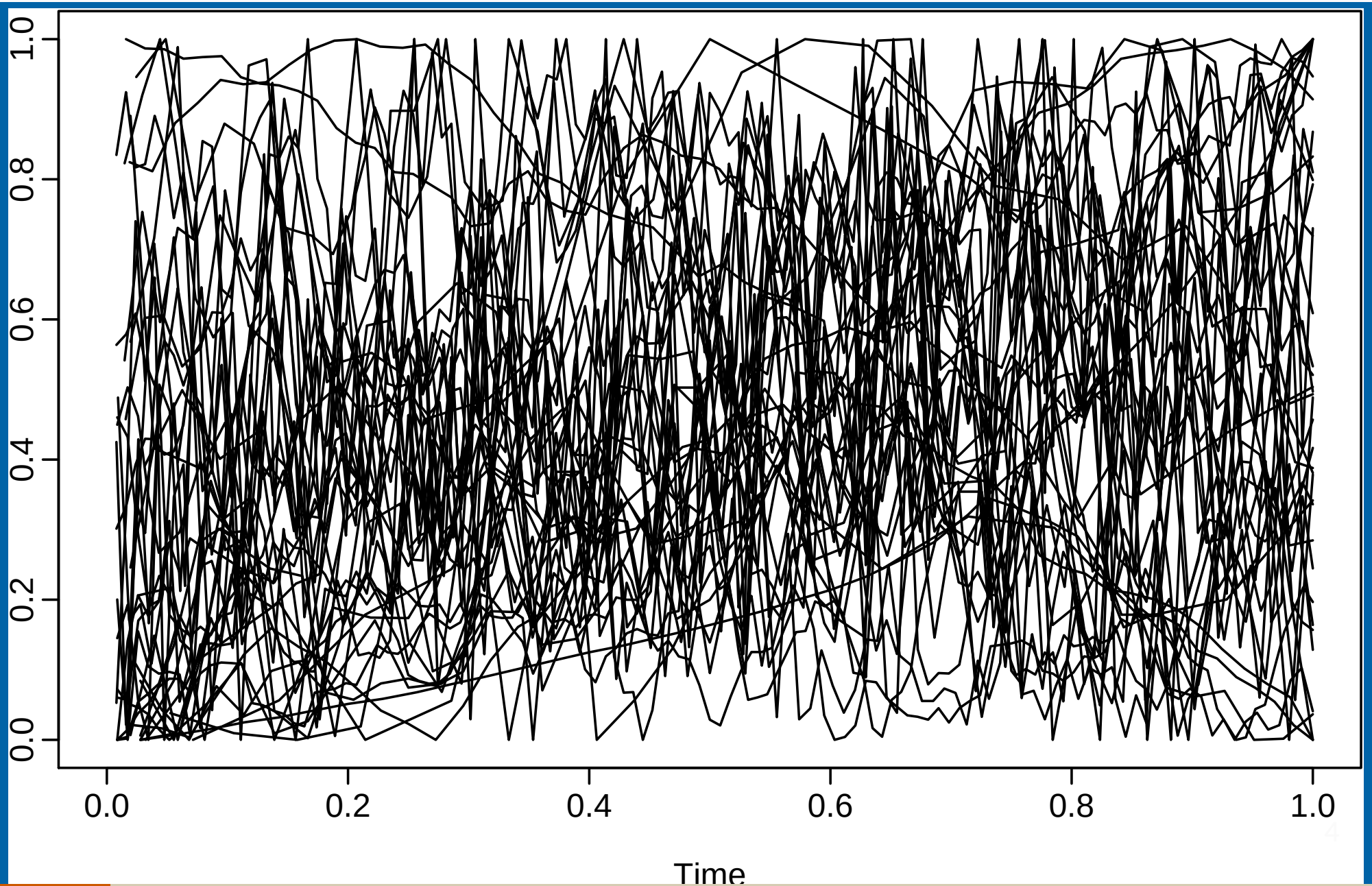
How to plot lots of time series?



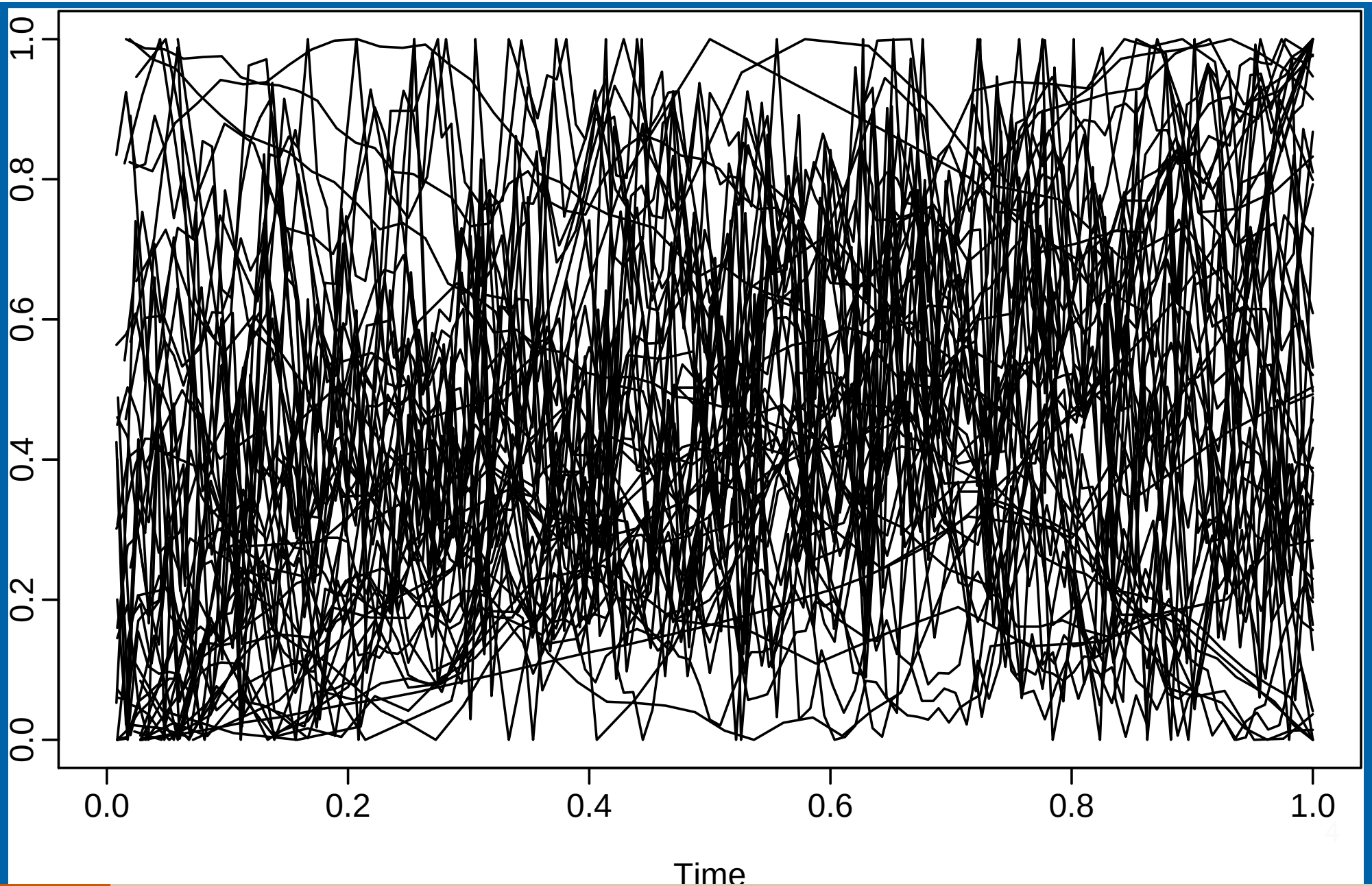
How to plot lots of time series?



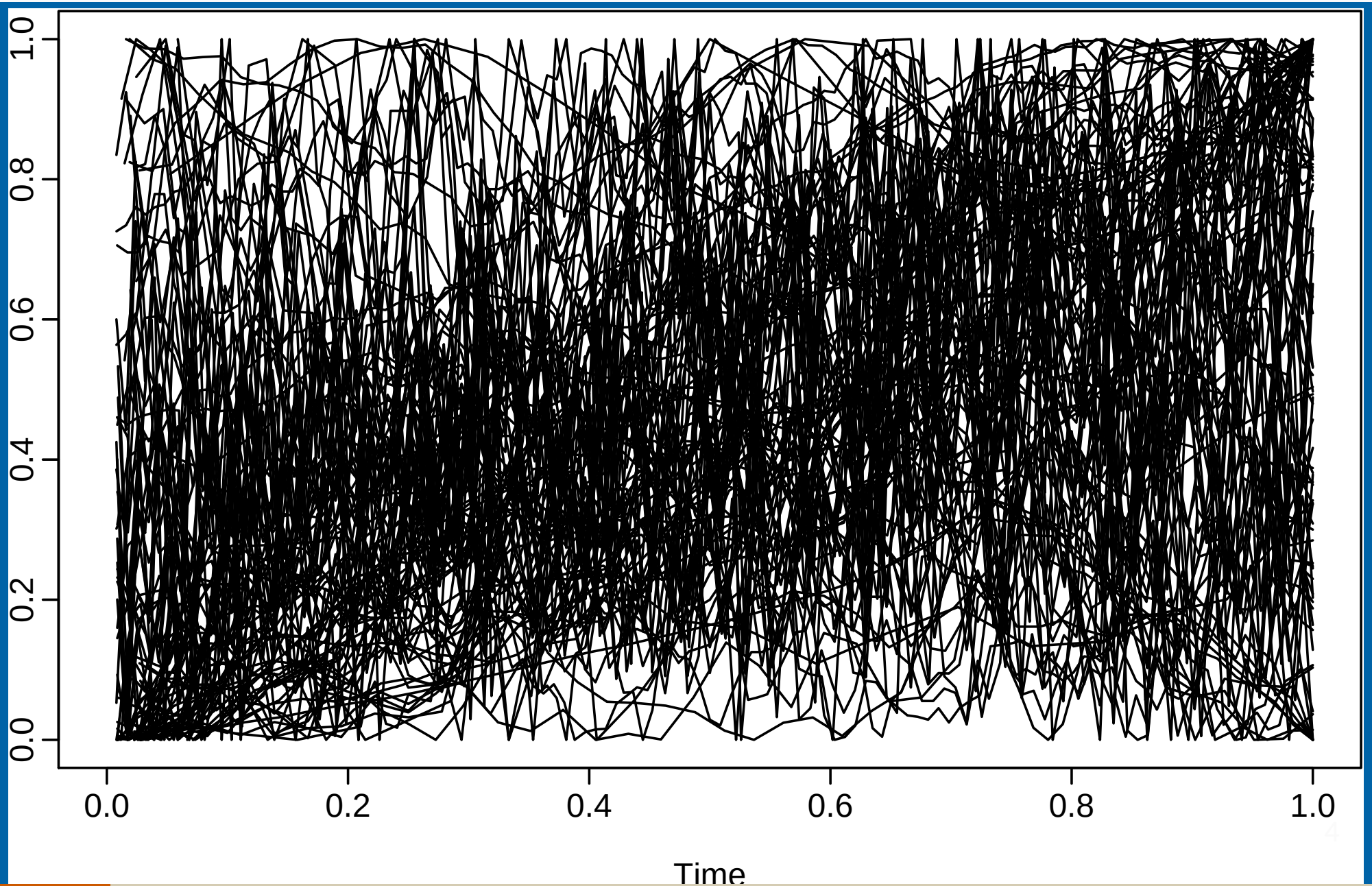
How to plot lots of time series?



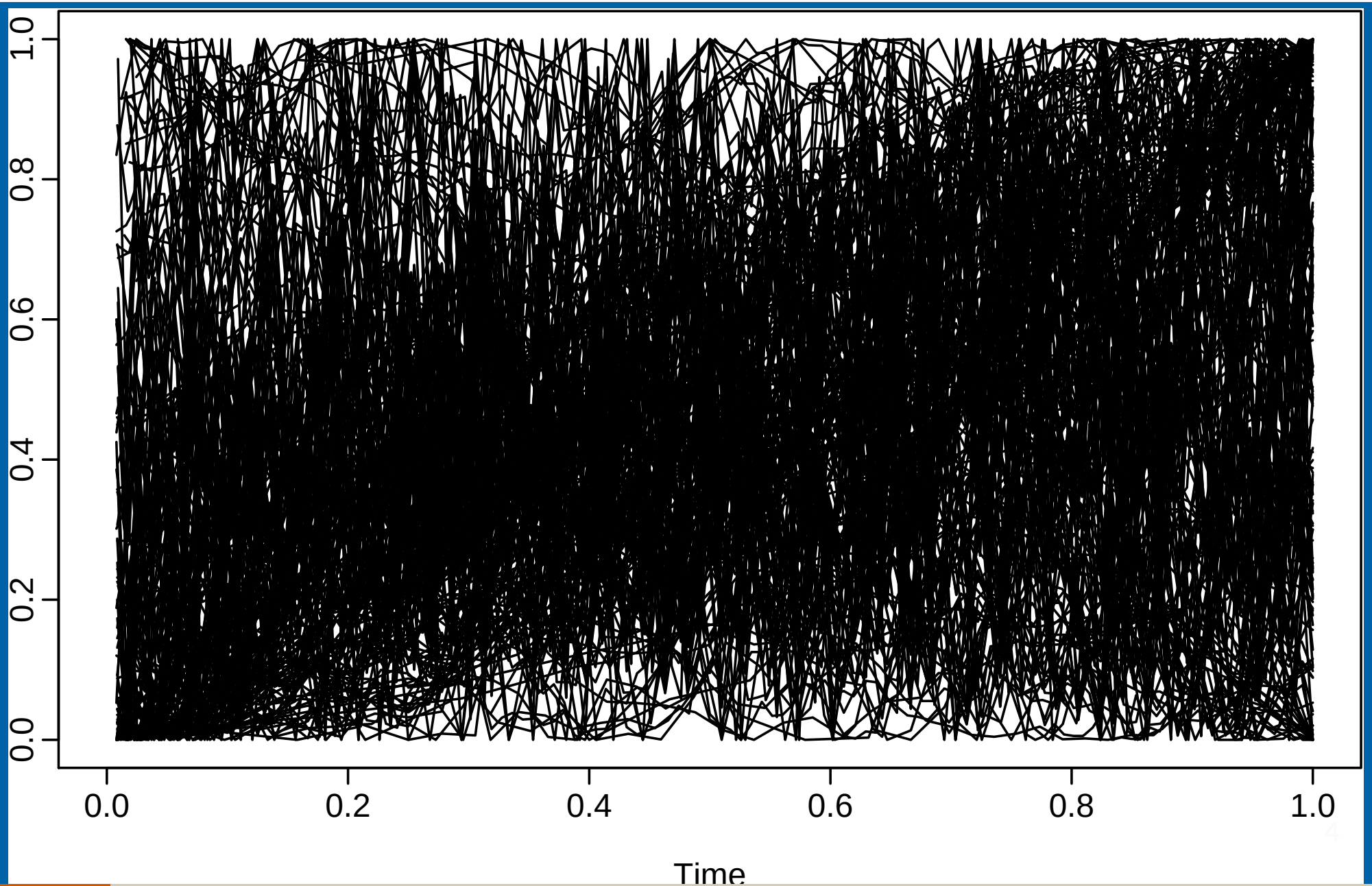
How to plot lots of time series?



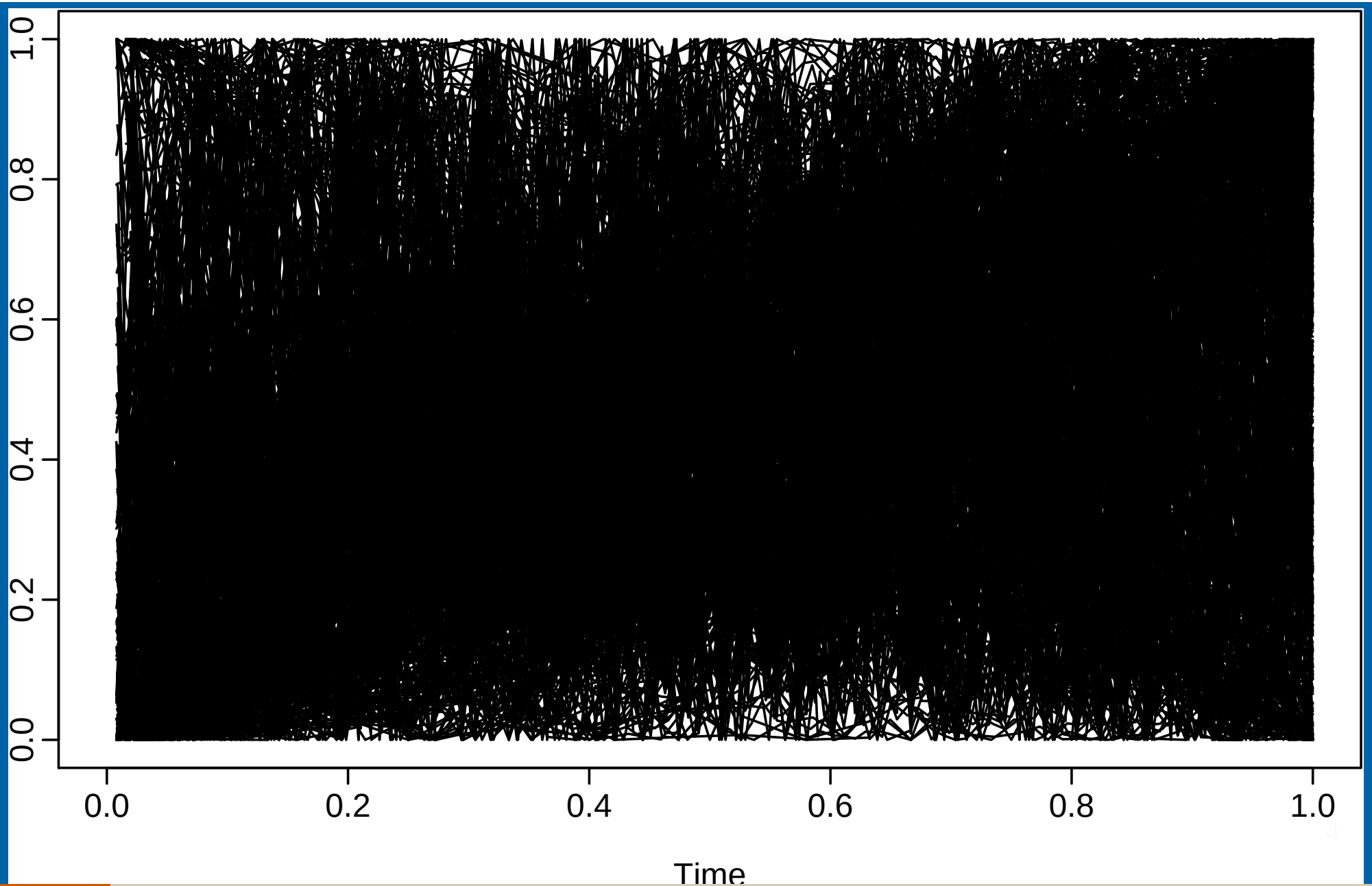
How to plot lots of time series?



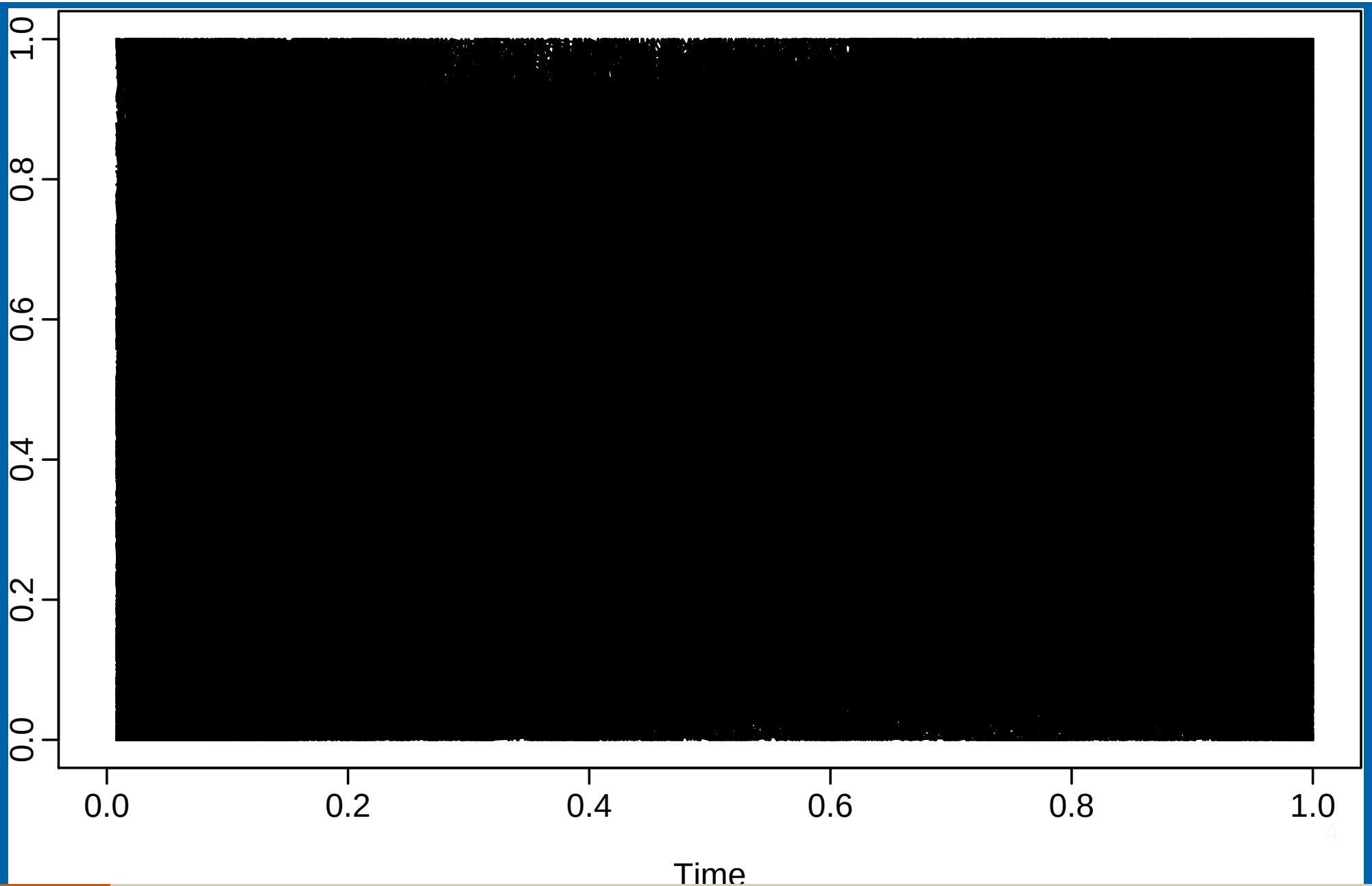
How to plot lots of time series?



How to plot lots of time series?



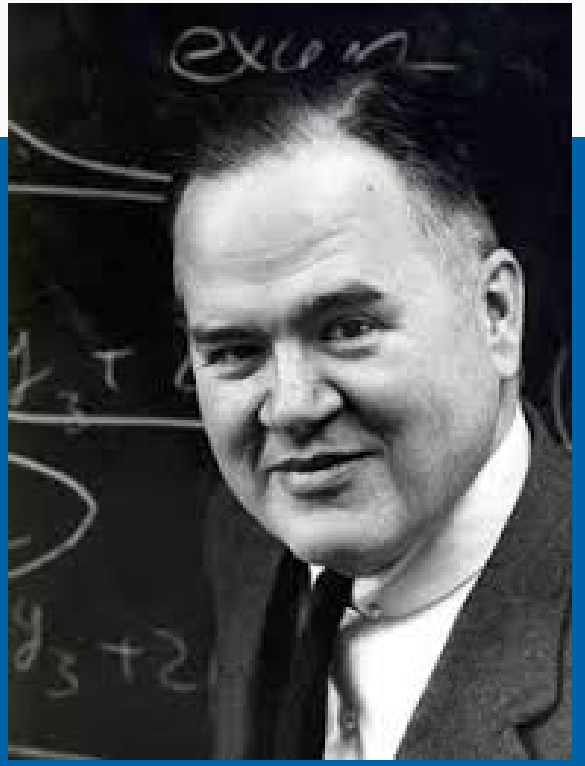
How to plot lots of time series?



Key idea

Cognostics

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



John W Tukey

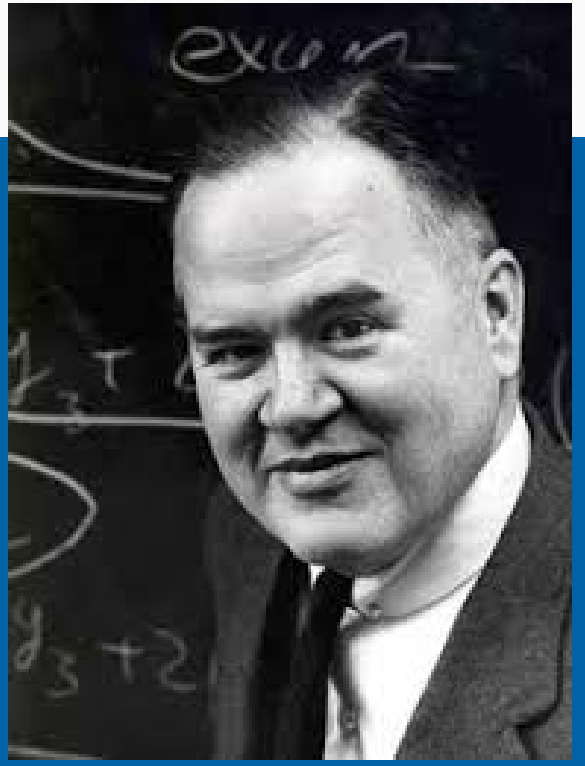
Key idea

Cognostics

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

Examples for time series

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

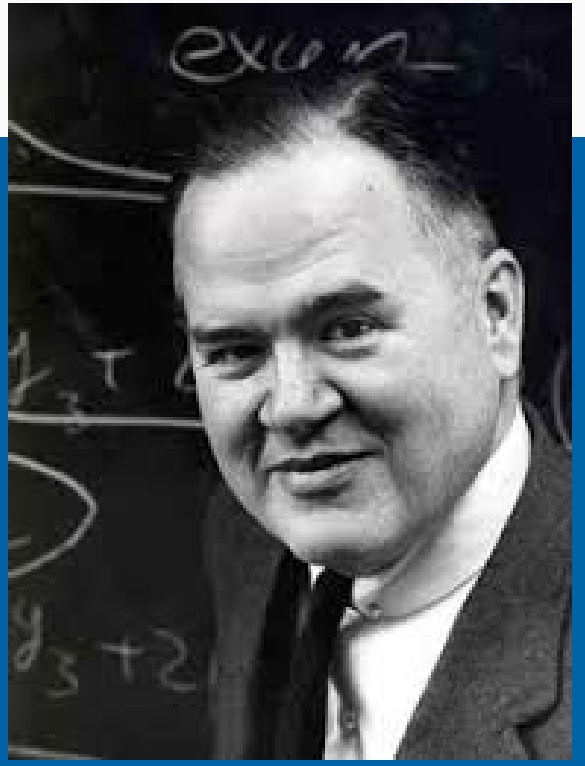


John W Tukey

Key idea

Cognostics

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



John W Tukey

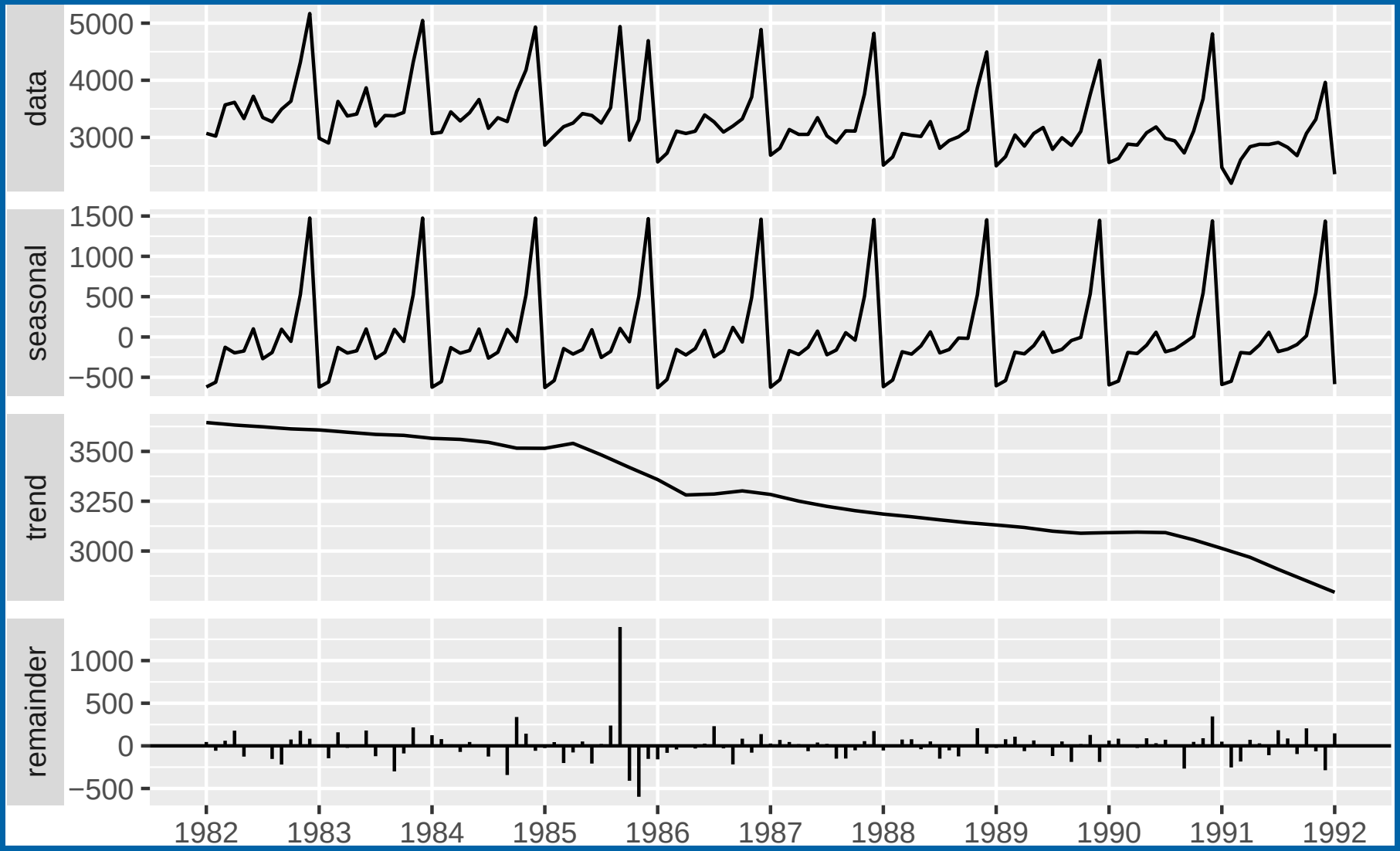
Examples for time series

- 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

$$Y_t = S_t + T_t + R_t$$

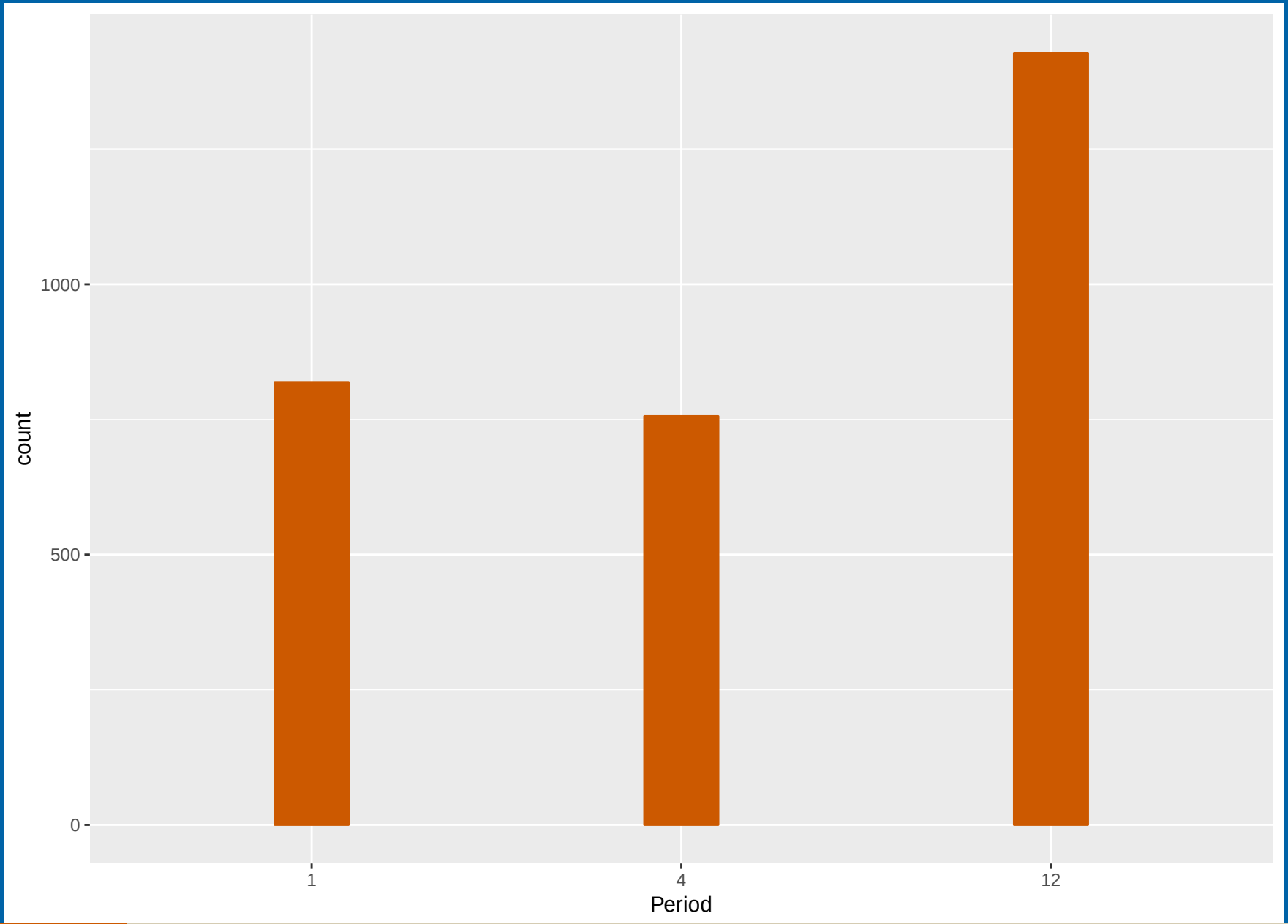
Candidate features

STL decomposition

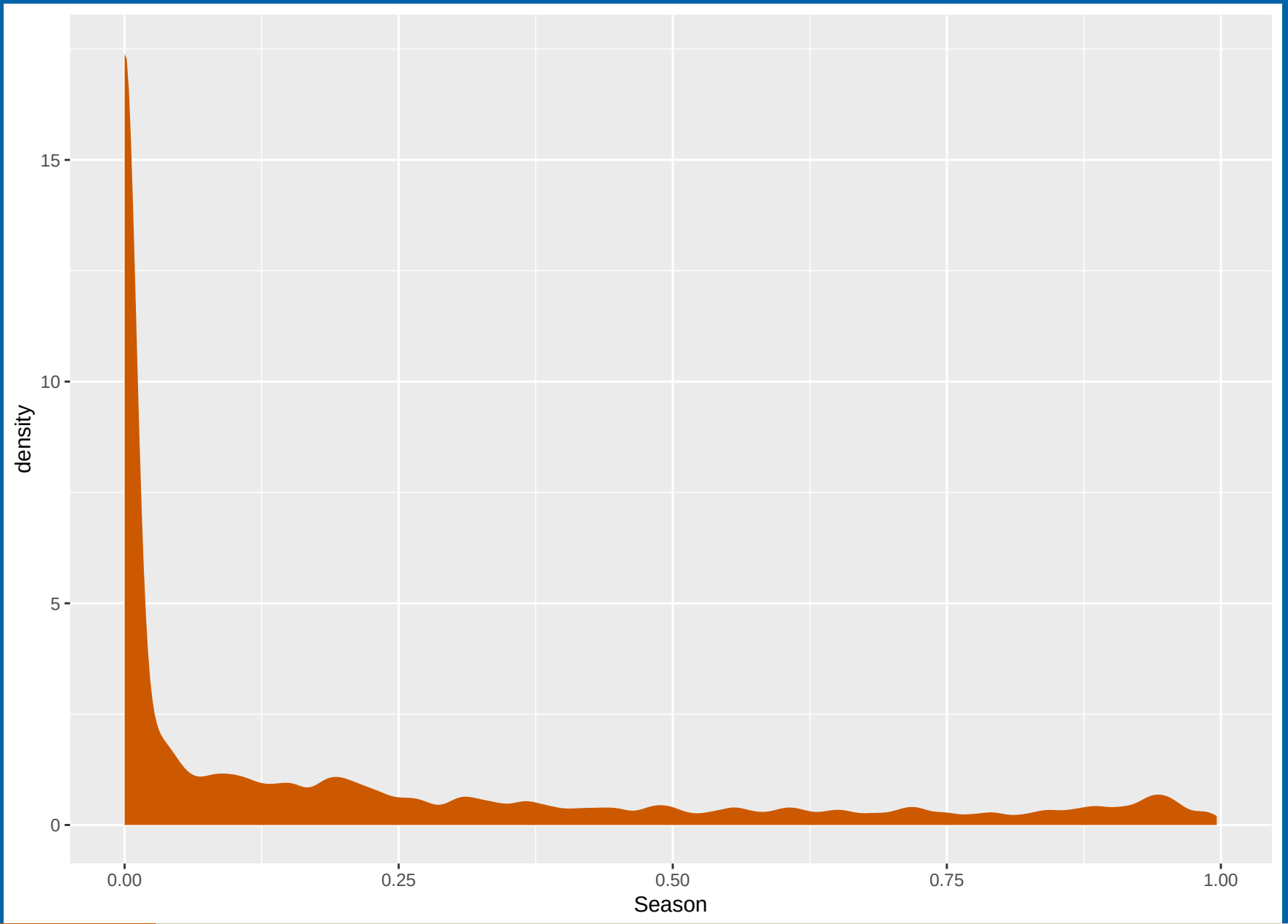
$$Y_t = S_t + T_t + R_t$$

- Seasonal period
- Autocorrelations of data $(Y_1, \ldots, Y_T)$
- Autocorrelations of residuals $(R_1, \ldots, R_T)$
- Strength of seasonality: $\max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t+R_t)}\right)$
- Strength of trend: $\max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t+R_t)}\right)$
- 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

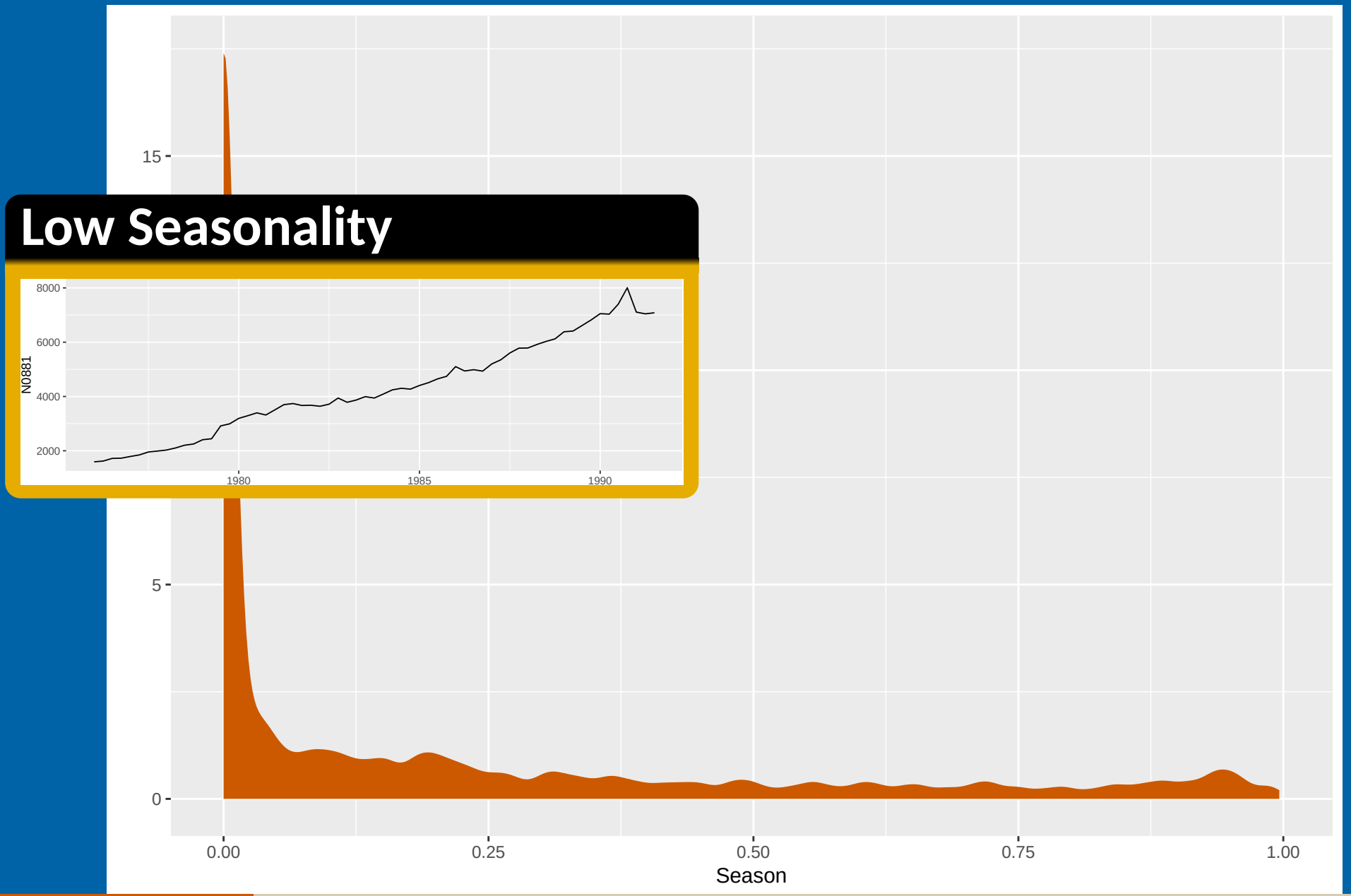
Distribution of Period for M3



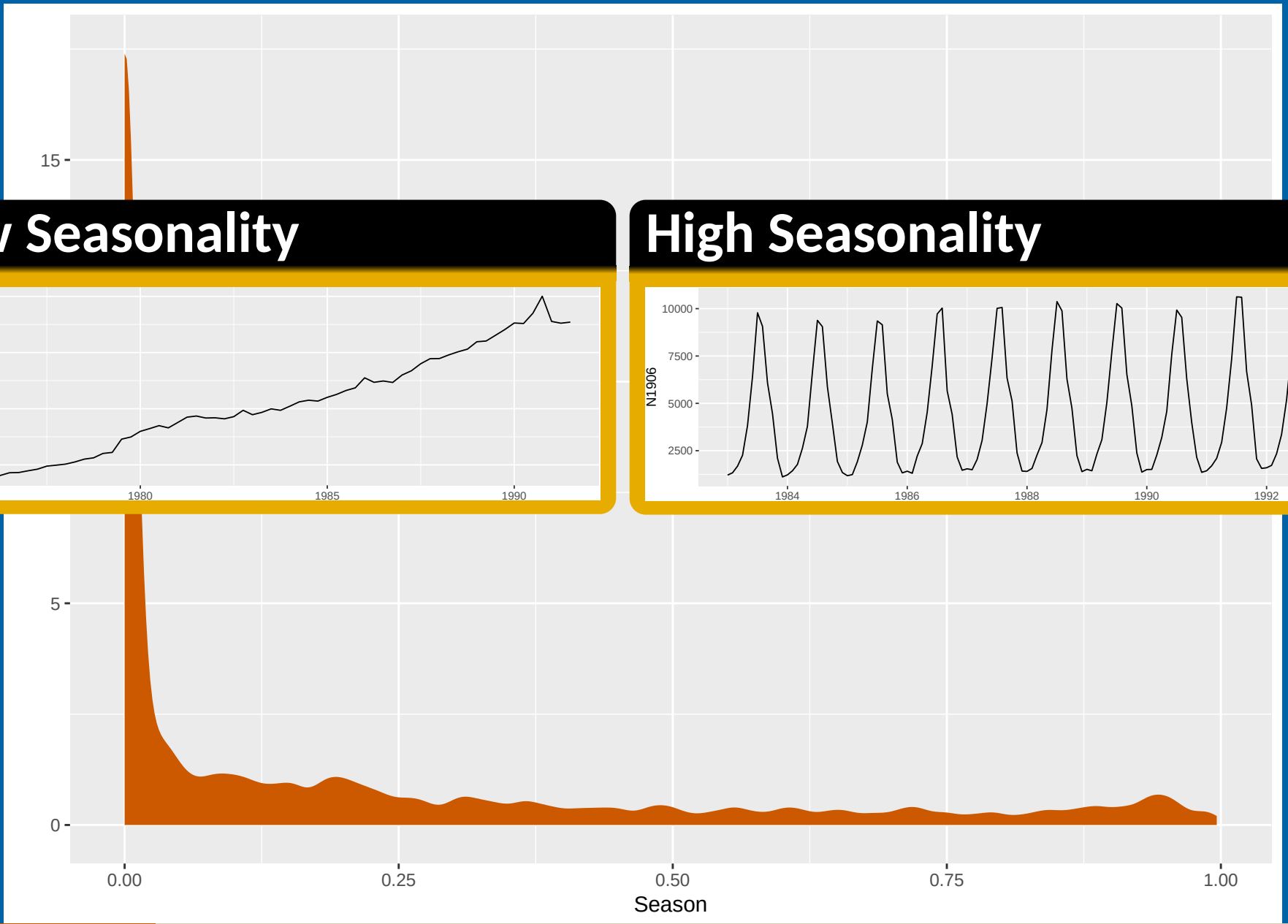
Distribution of Seasonality for M3



Distribution of Seasonality for M3



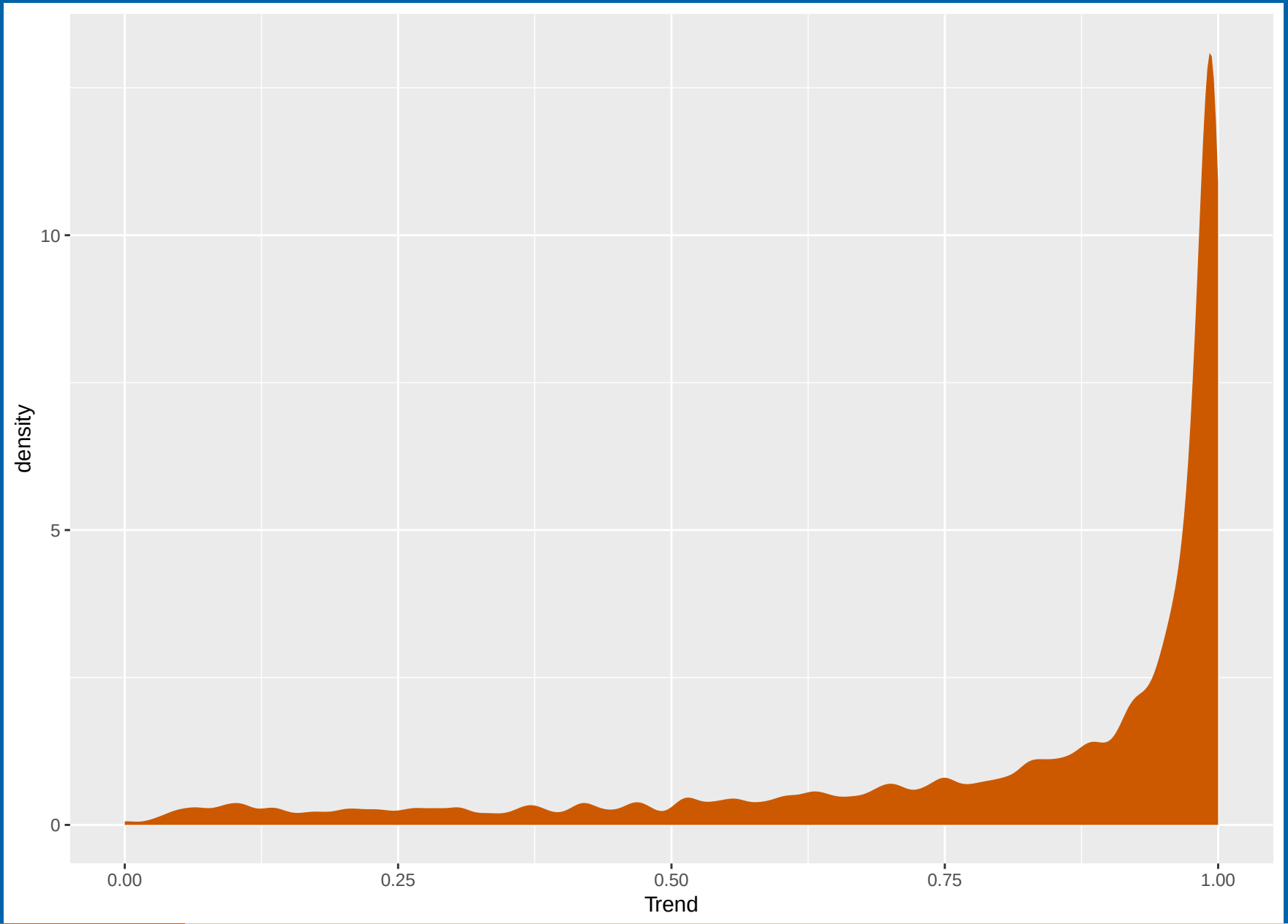
Distribution of Seasonality for M3



Low Seasonality

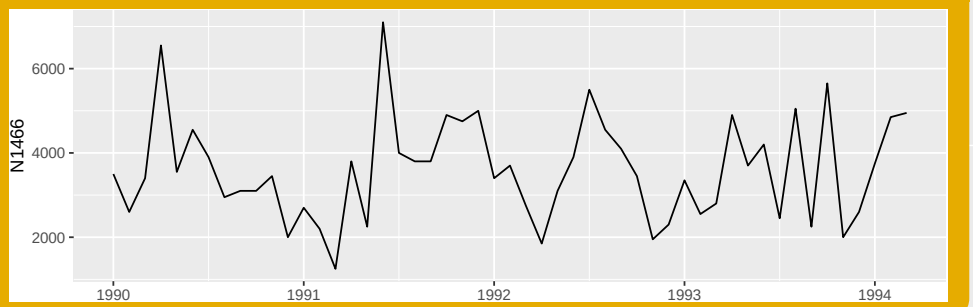
High Seasonality

Distribution of Trend for M3



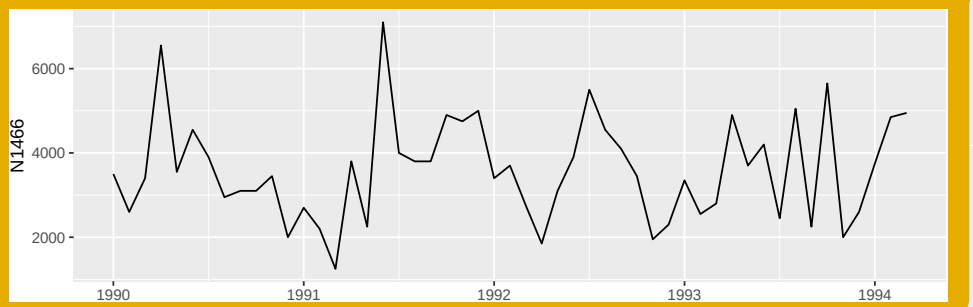
Distribution of Trend for M3

Low Trend

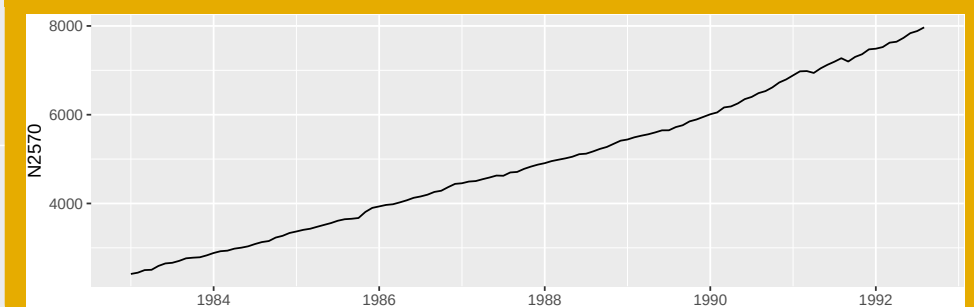


Distribution of Trend for M3

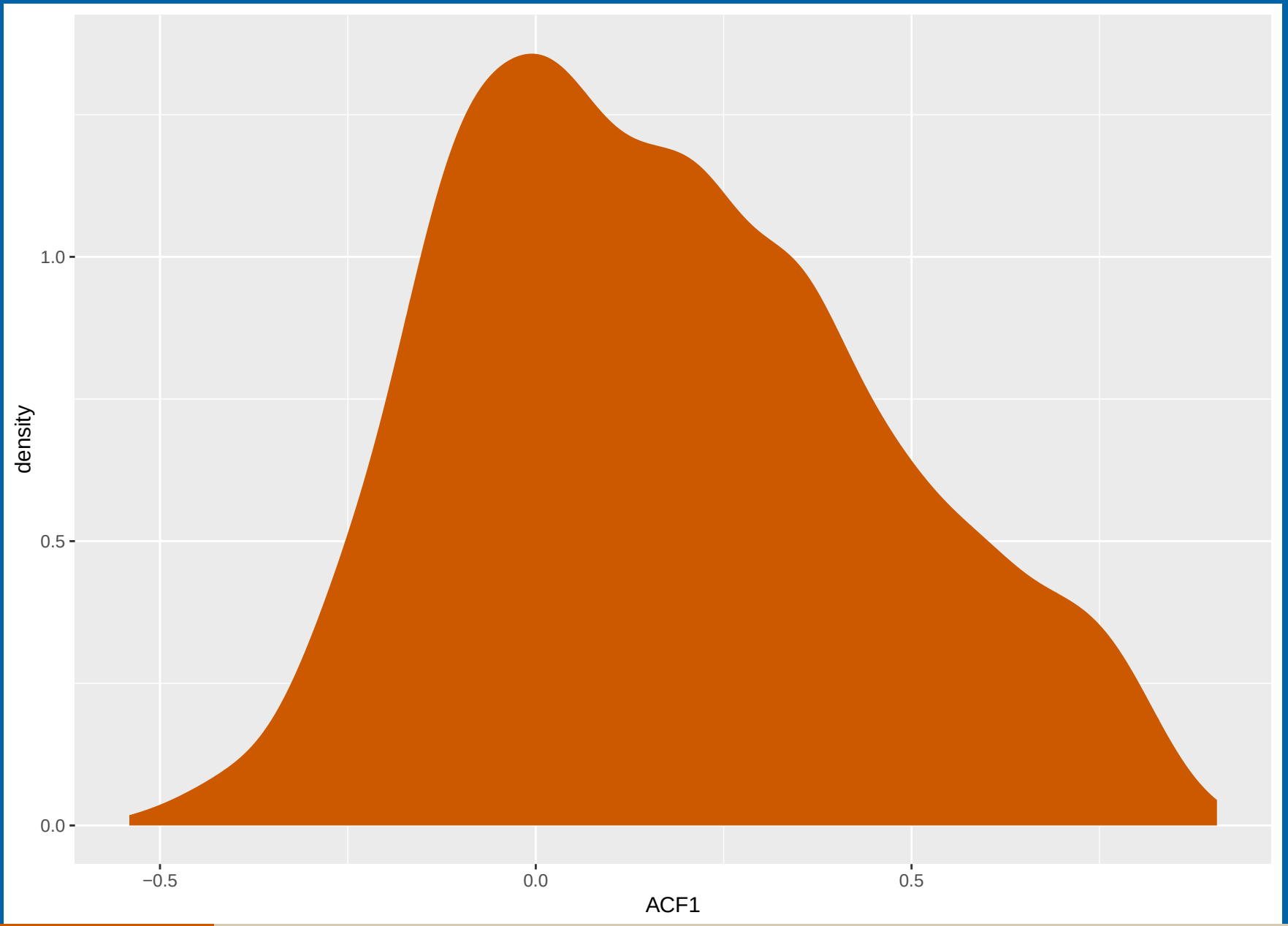
Low Trend



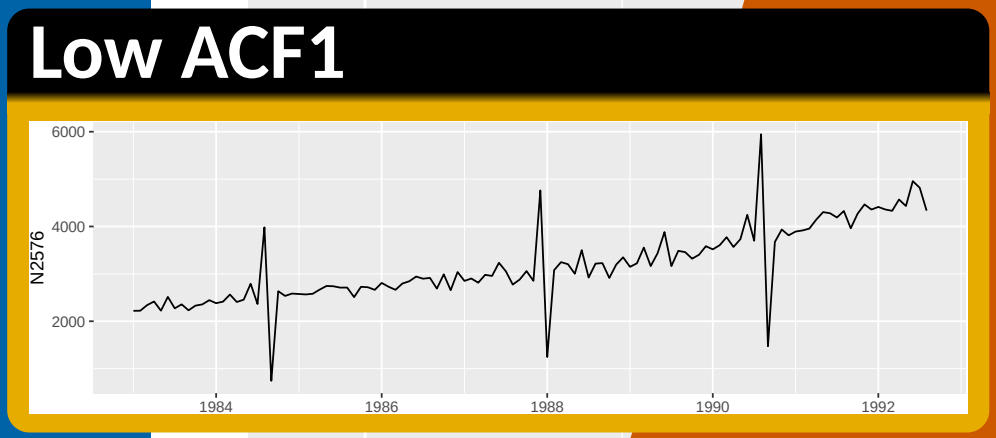
High Trend



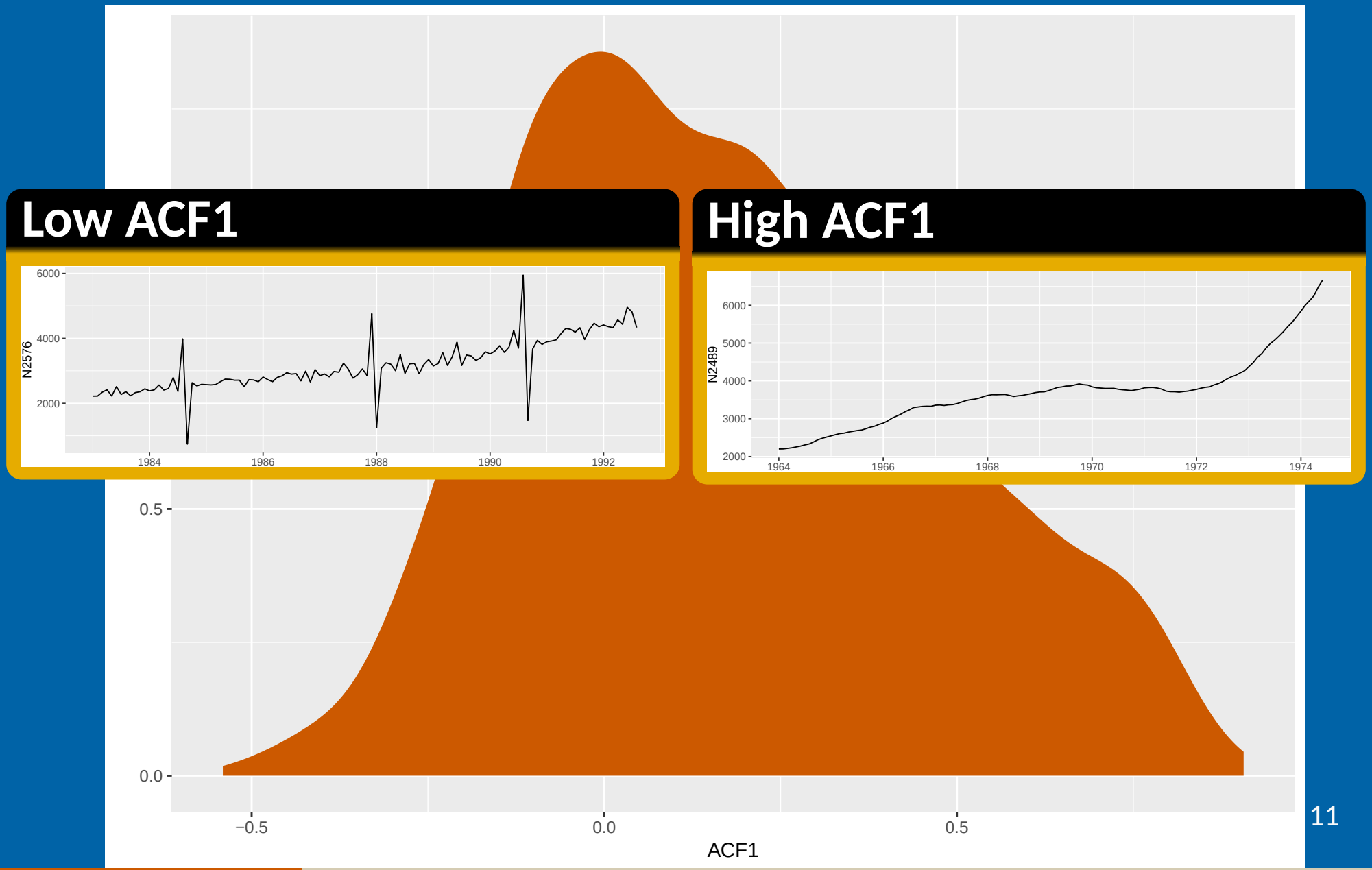
Distribution of residual ACF1 for M3



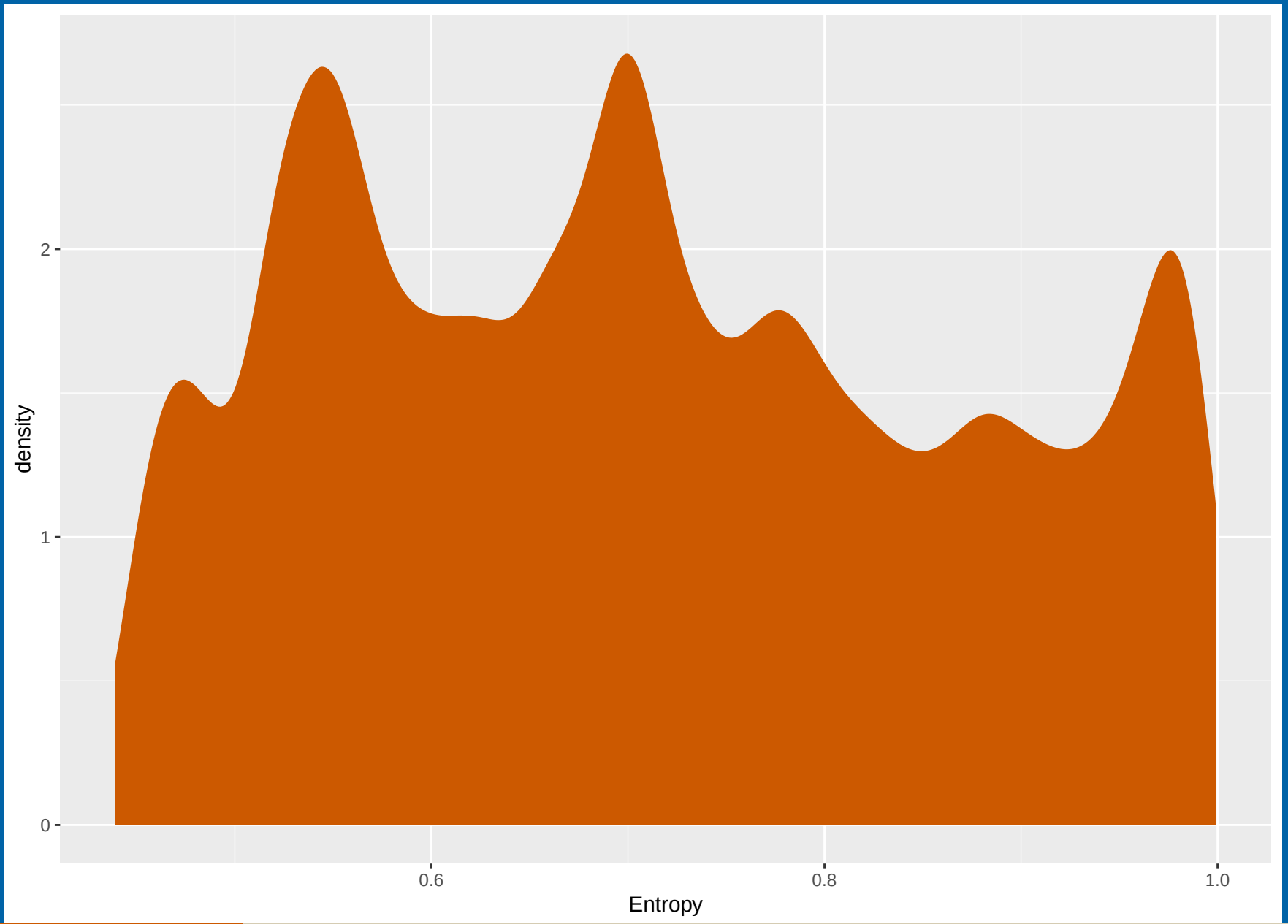
Distribution of residual ACF1 for M3



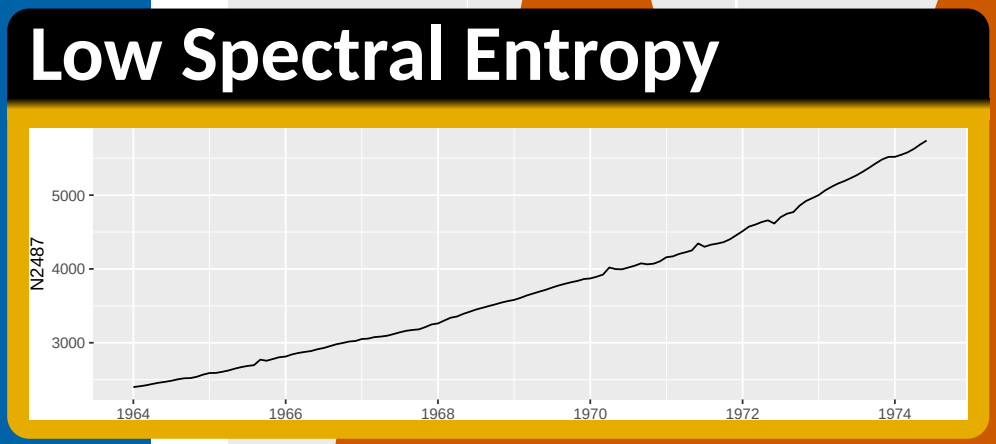
Distribution of residual ACF1 for M3



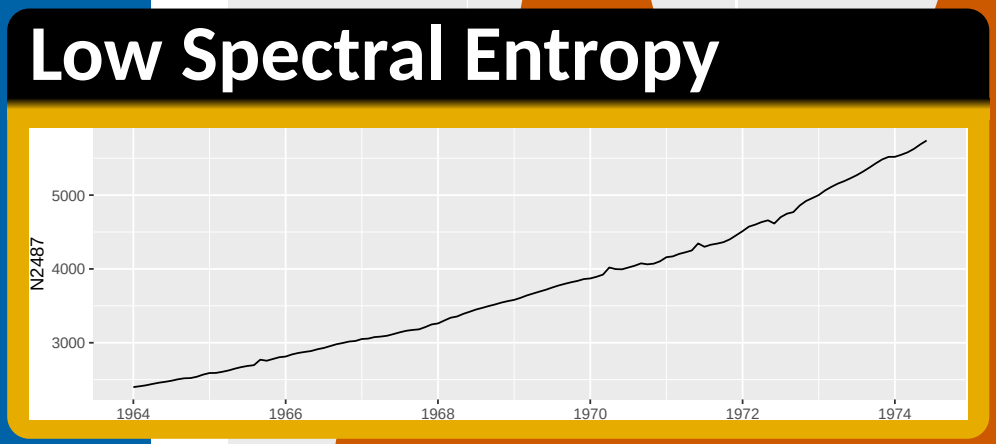
Distribution of Spectral Entropy for M3

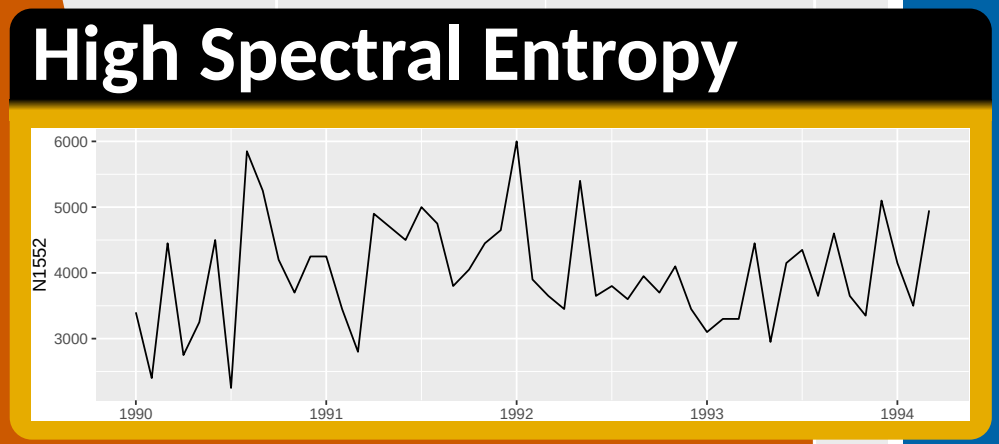


Distribution of Spectral Entropy for M3

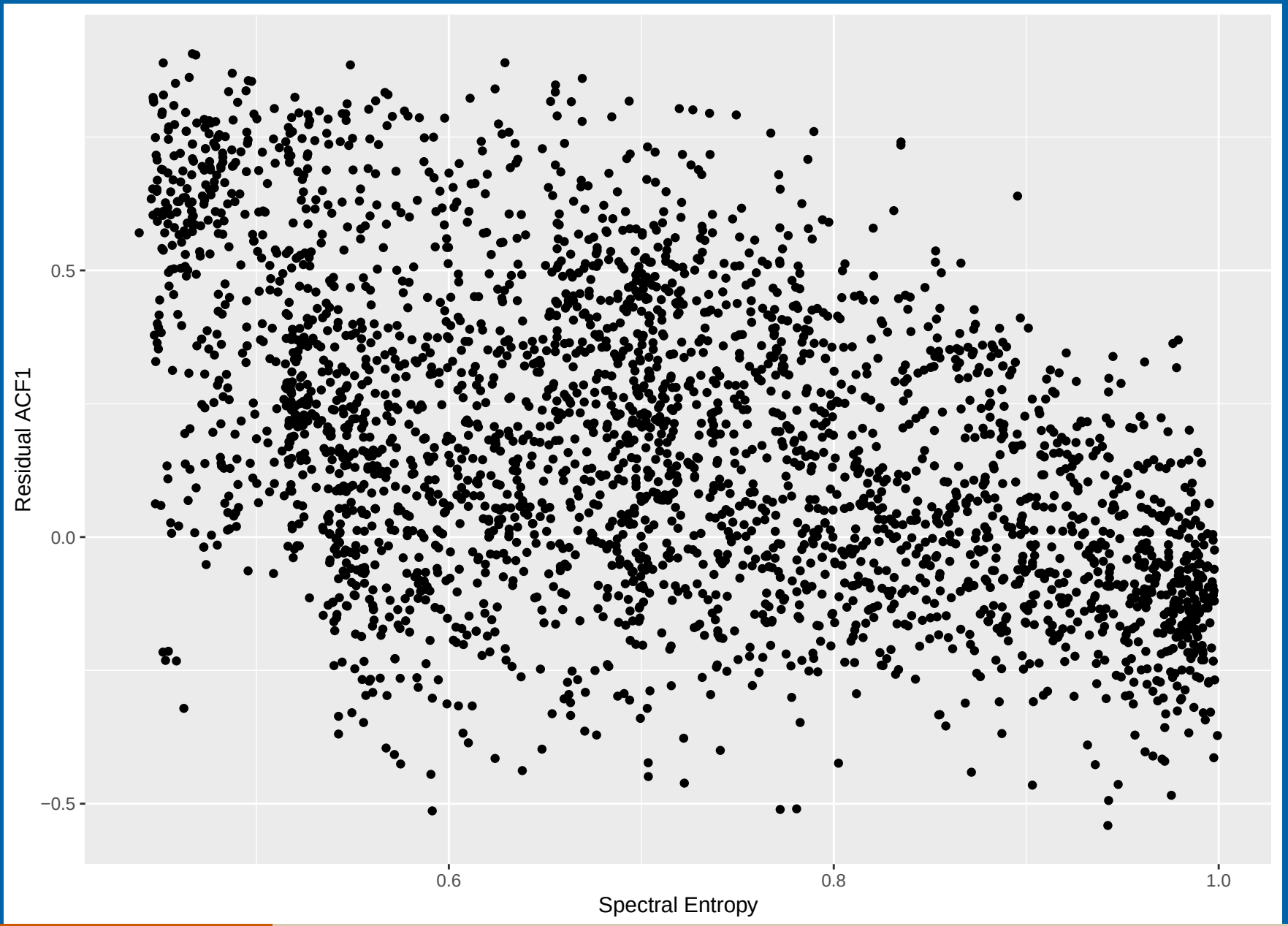


Distribution of Spectral Entropy for M3

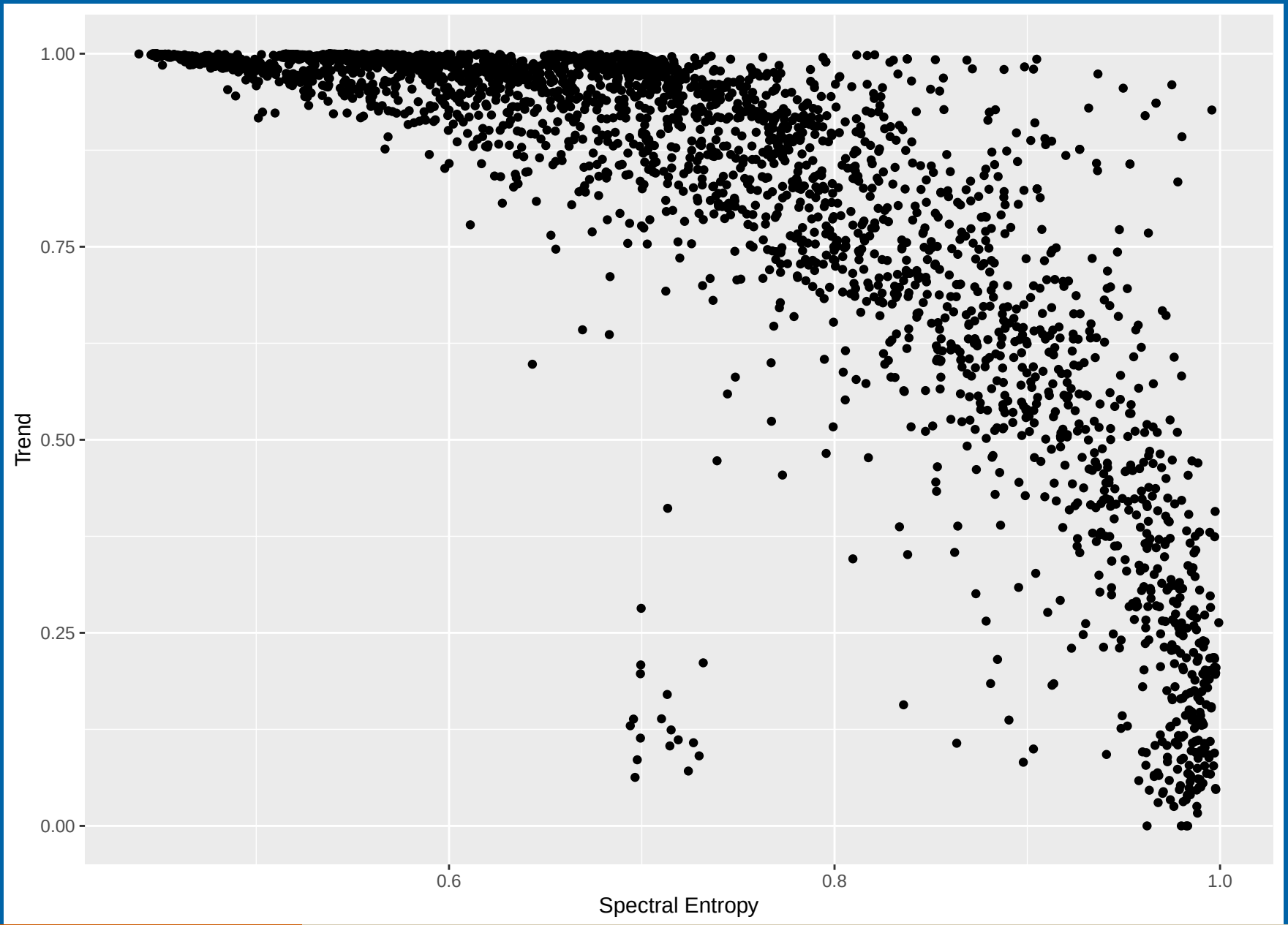




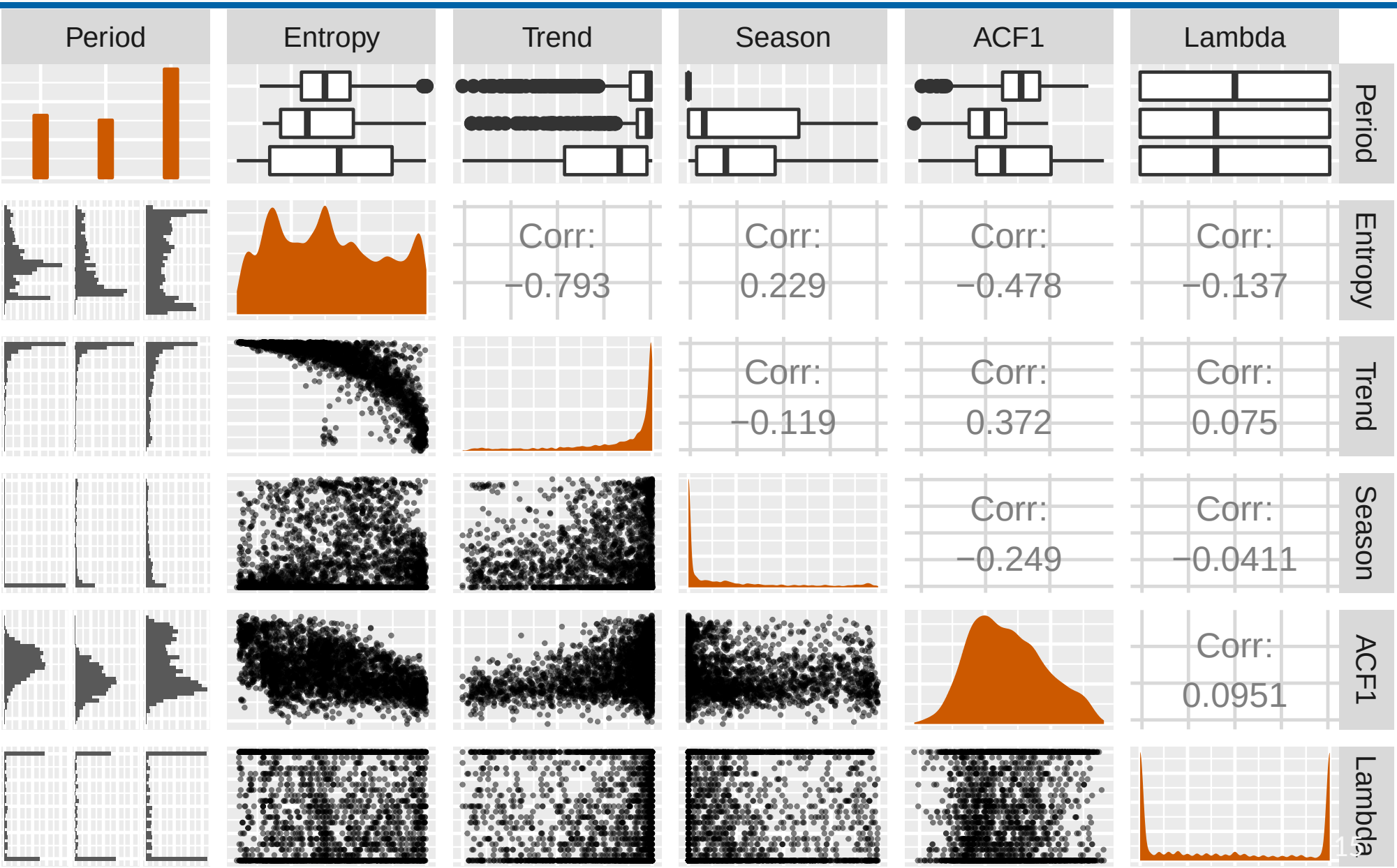
Feature distributions



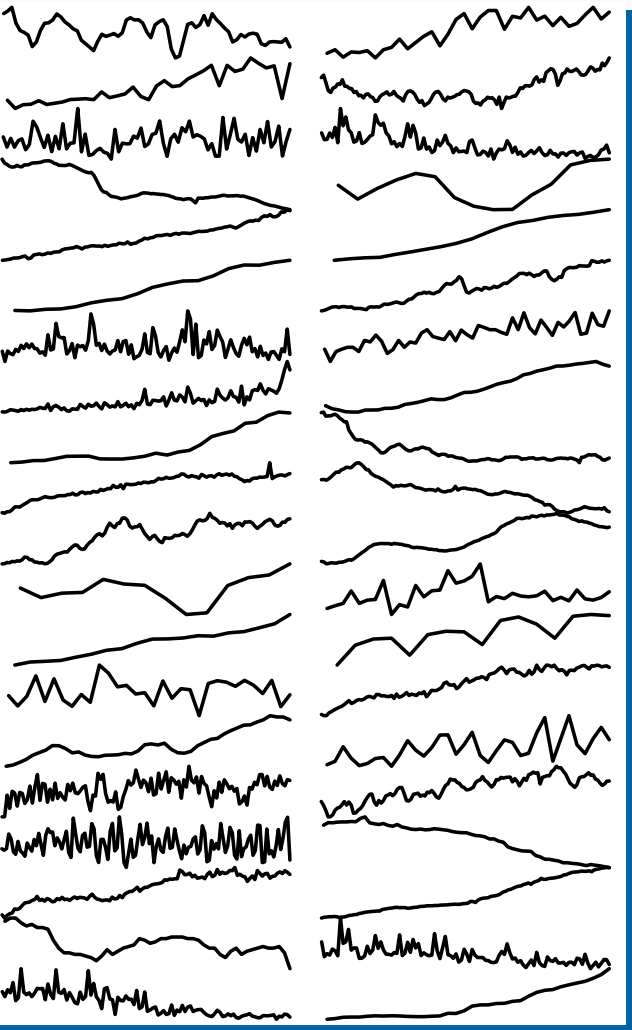
Feature distributions



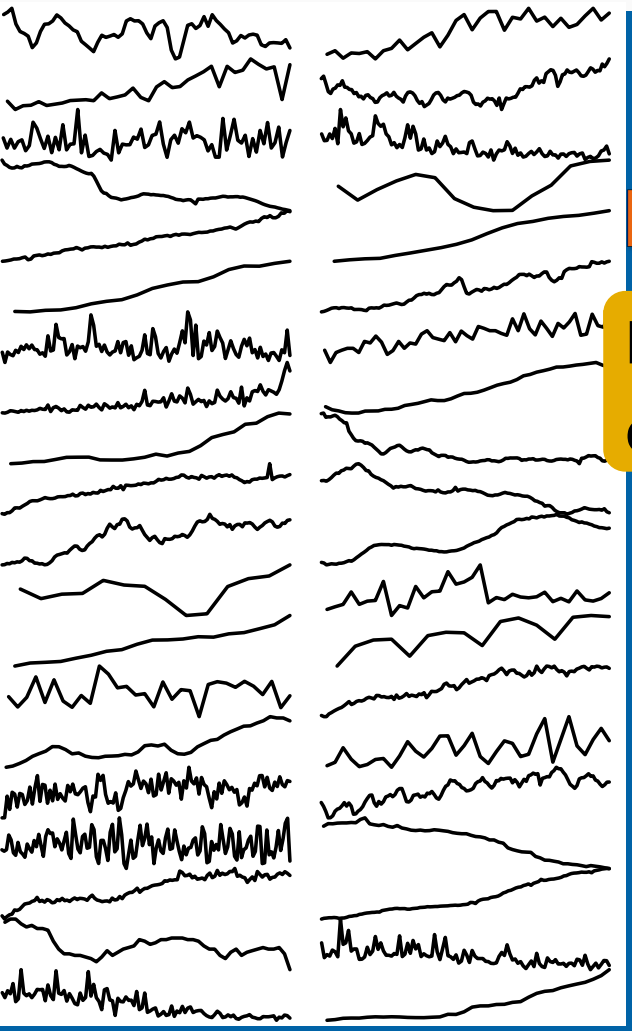
Feature distributions



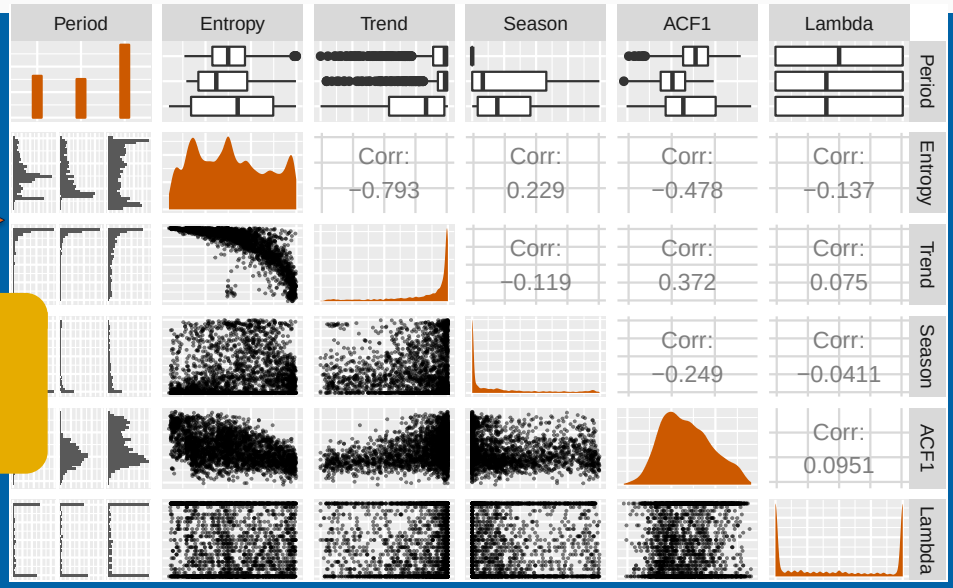
Dimension reduction for time series



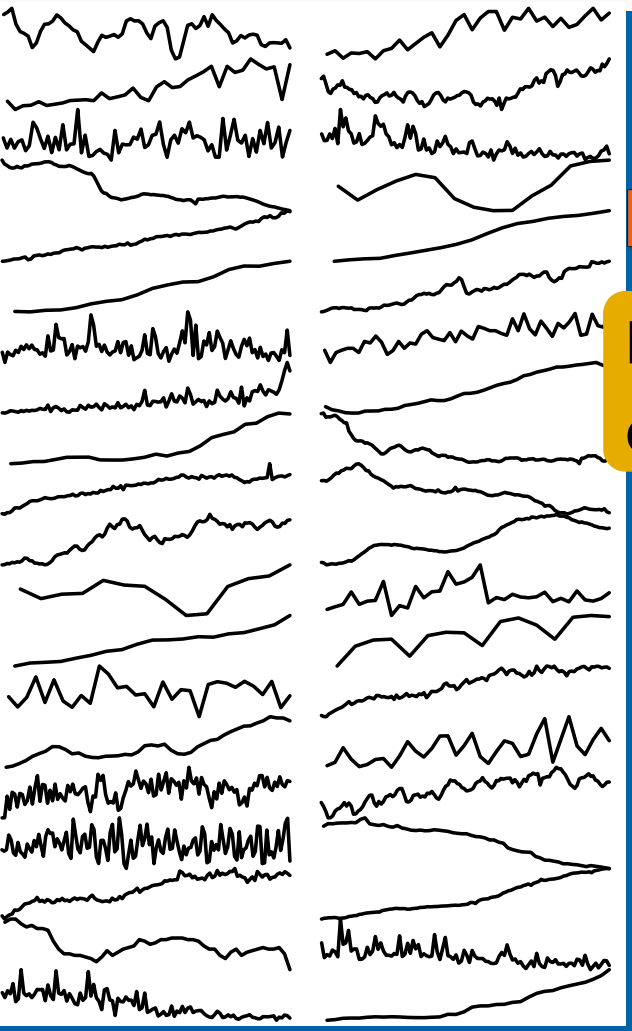
Dimension reduction for time series



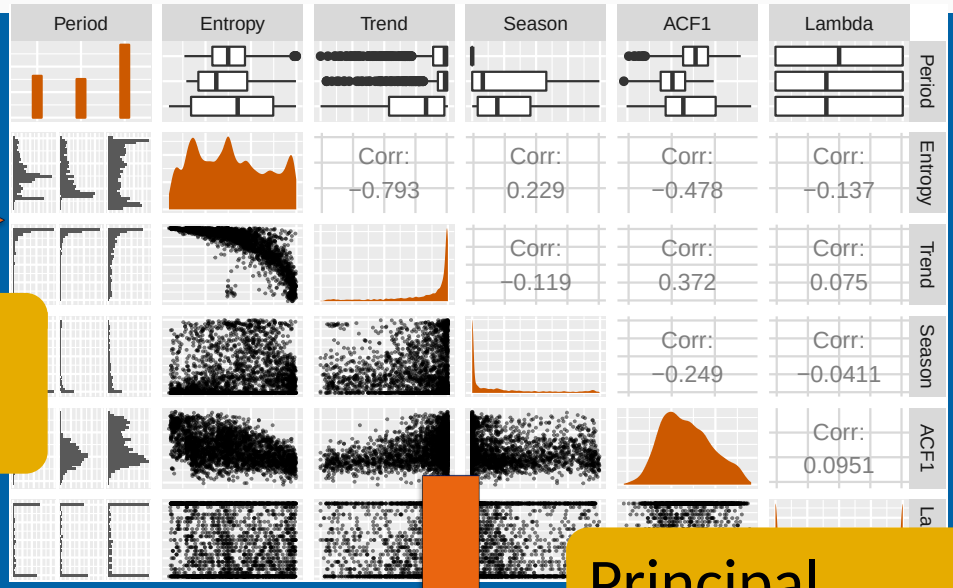
Feature calculation



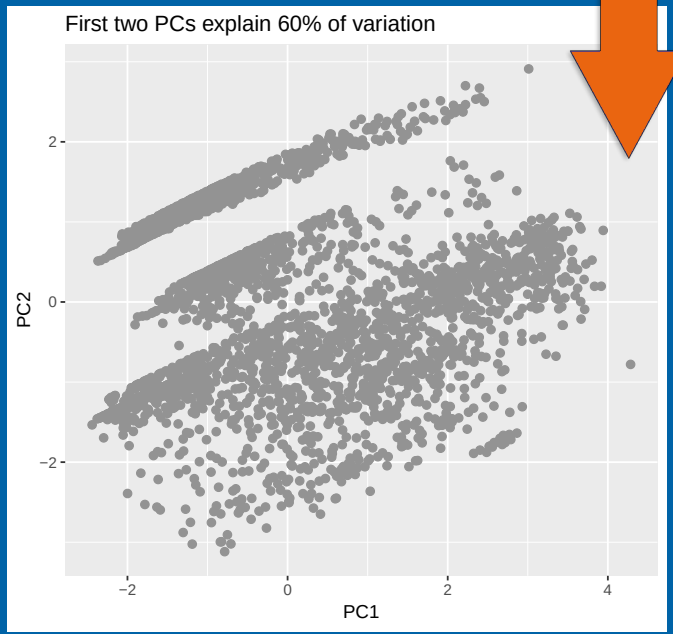
Dimension reduction for time series



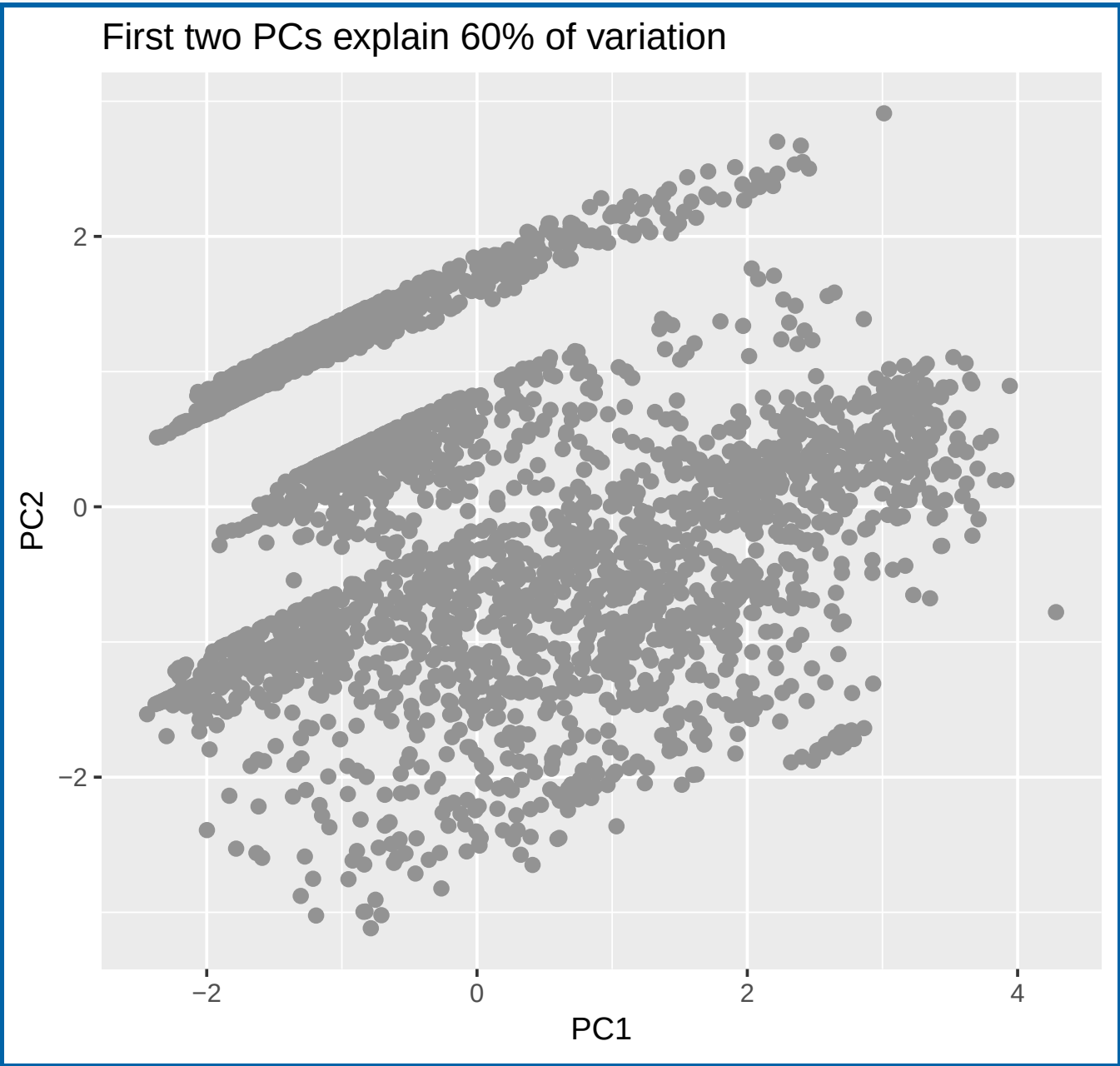
Feature calculation



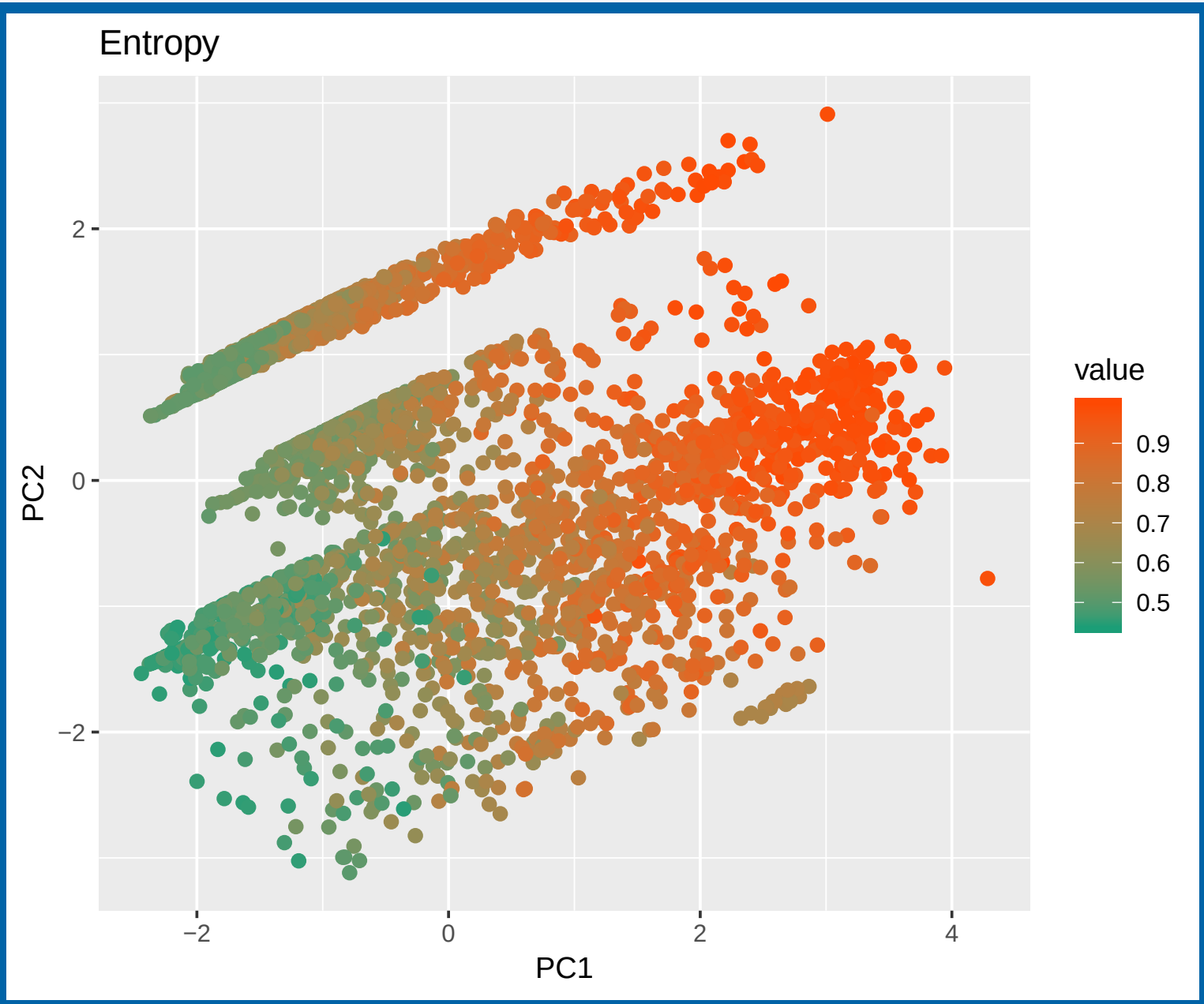
Principal component decomposition



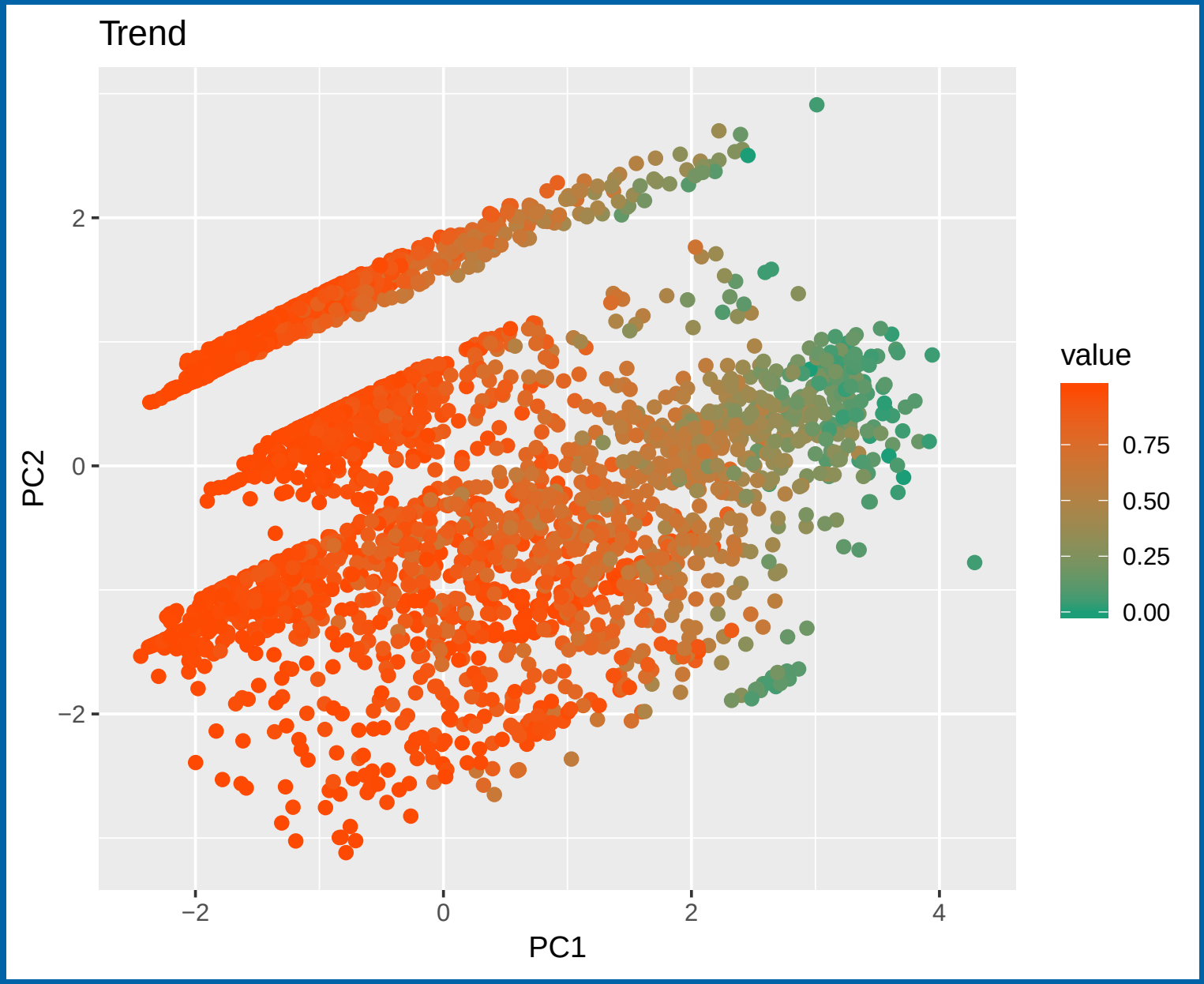
Feature space of M3 data



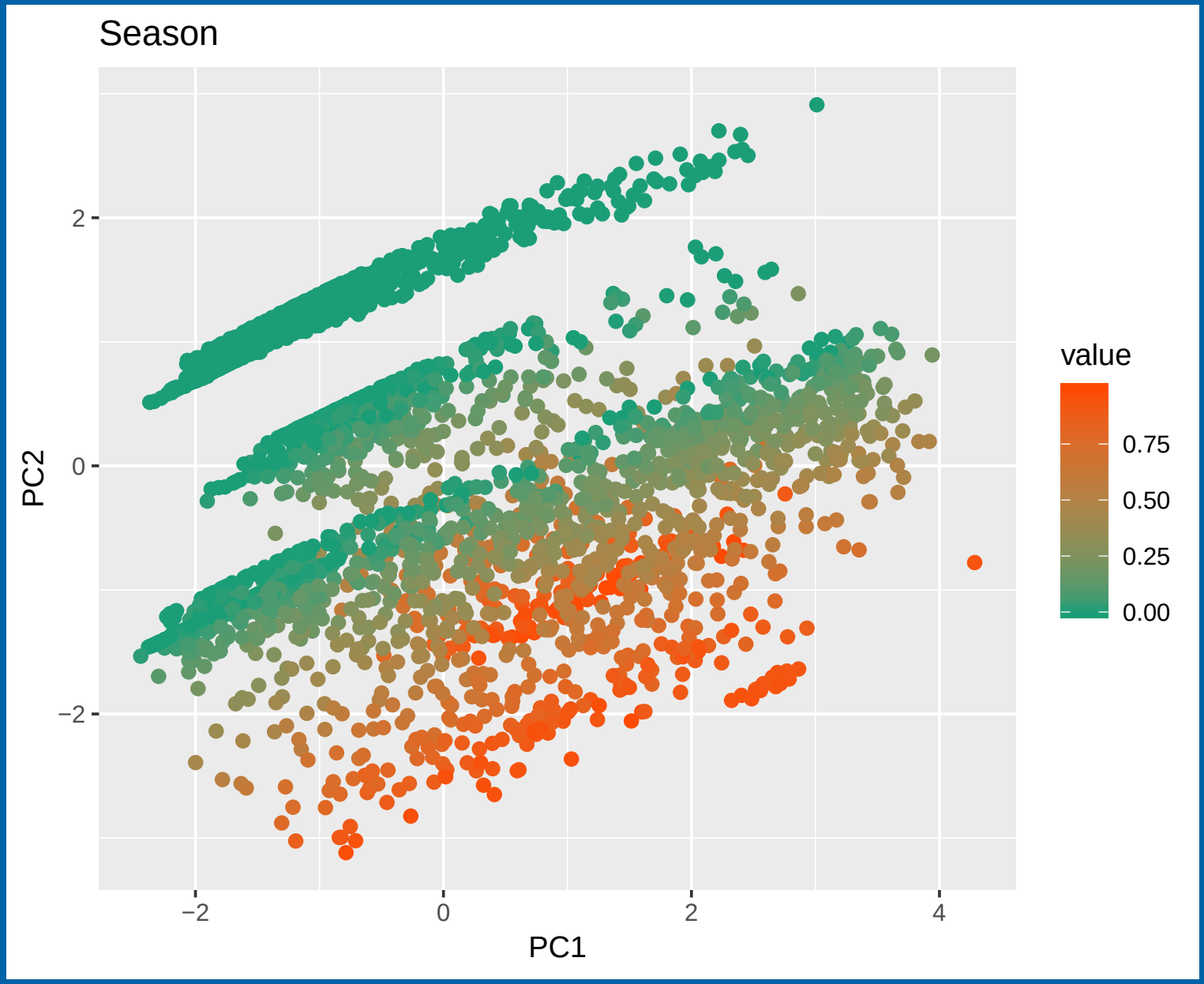
Feature space of M3 data



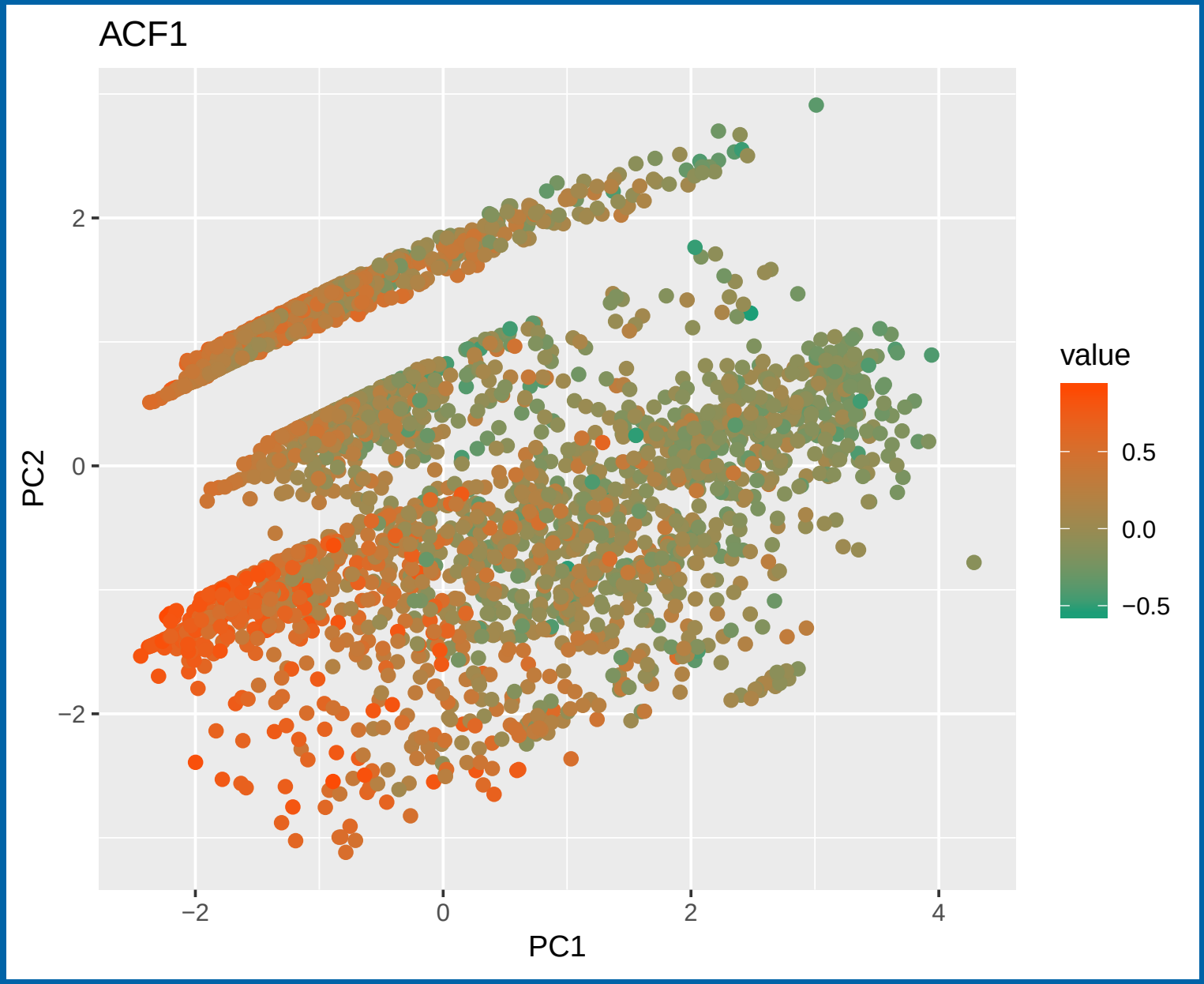
Feature space of M3 data



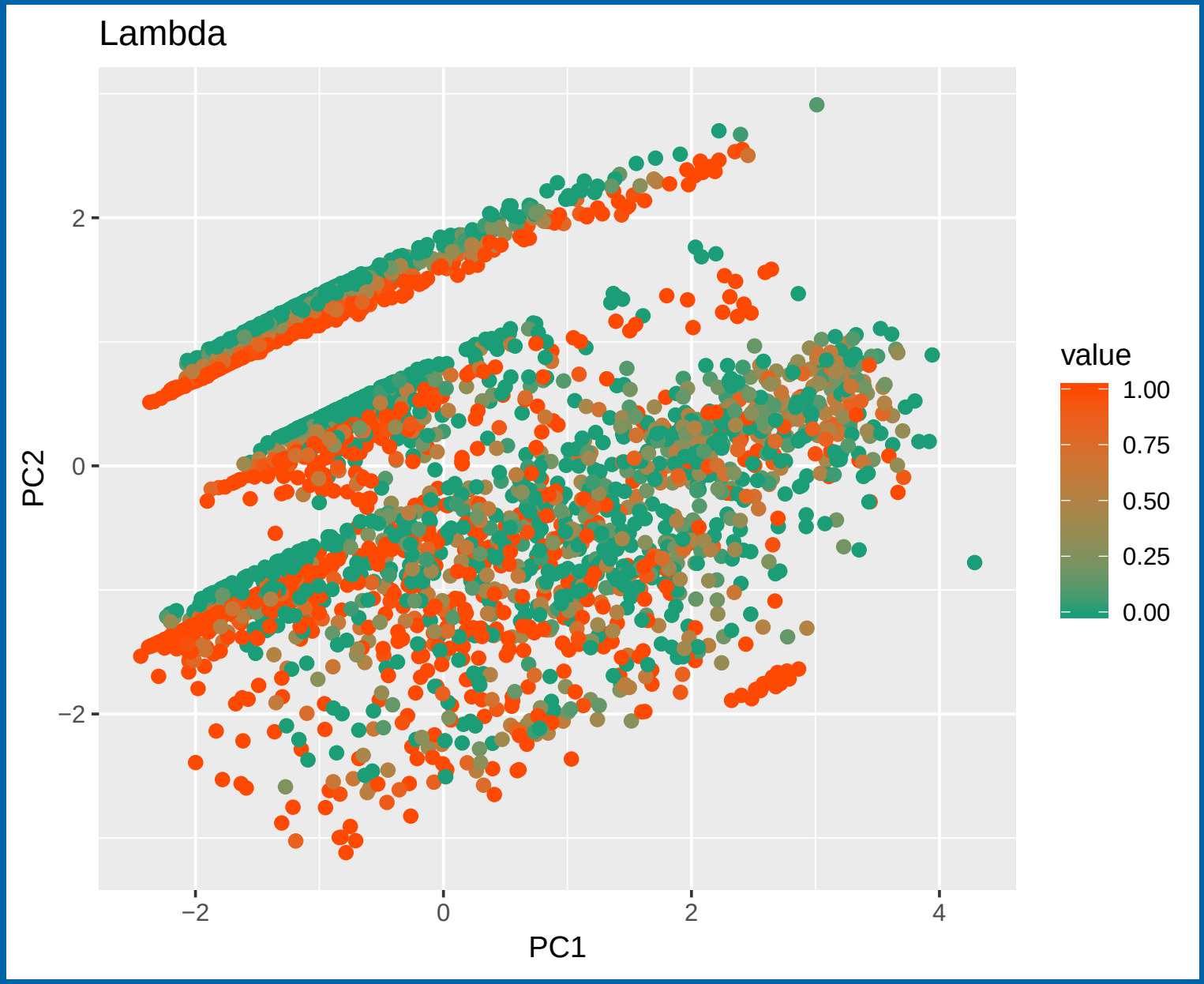
Feature space of M3 data



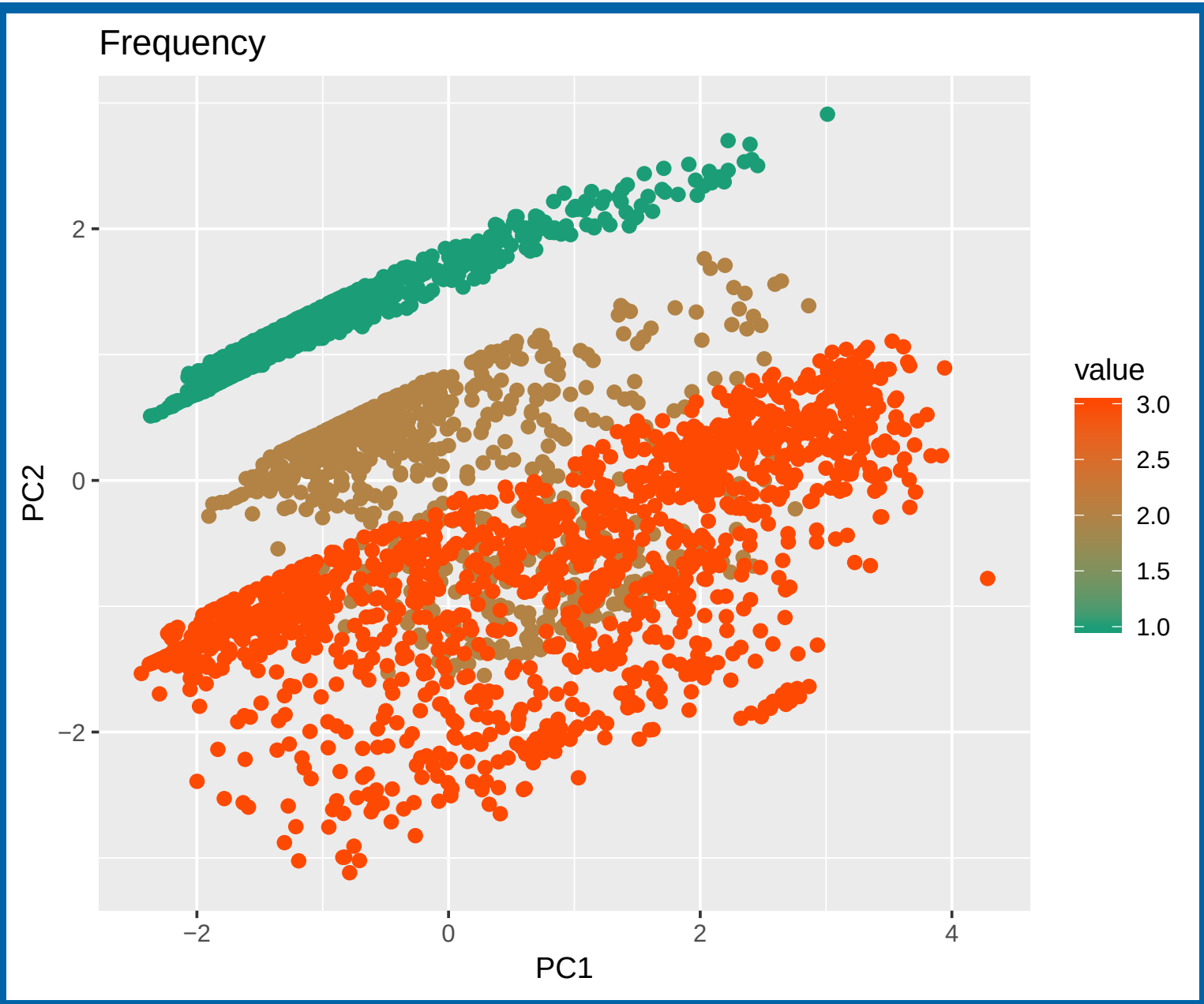
Feature space of M3 data



Feature space of M3 data



Feature space of M3 data



Papers and packages

Kang, Hyndman, & Smith-Miles, K. (2017) Visualising forecasting algorithm performance using time series instance spaces. *IJF*, **33**(2) 345–358.

Hyndman, Wang, Kang, Talagala & Montero-Manso (2018). **tsfeatures**: Time Series Feature Extraction. github.com/robjhyndman/tsfeatures/

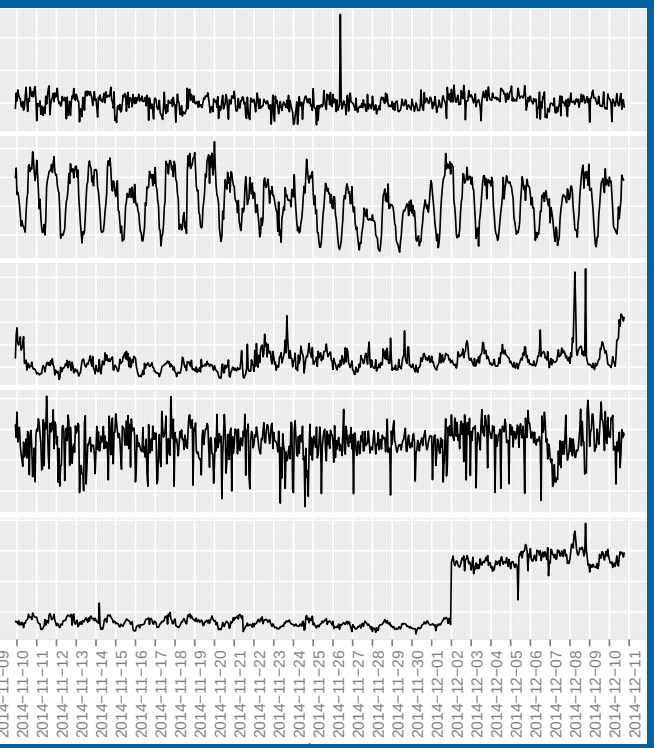
Outline

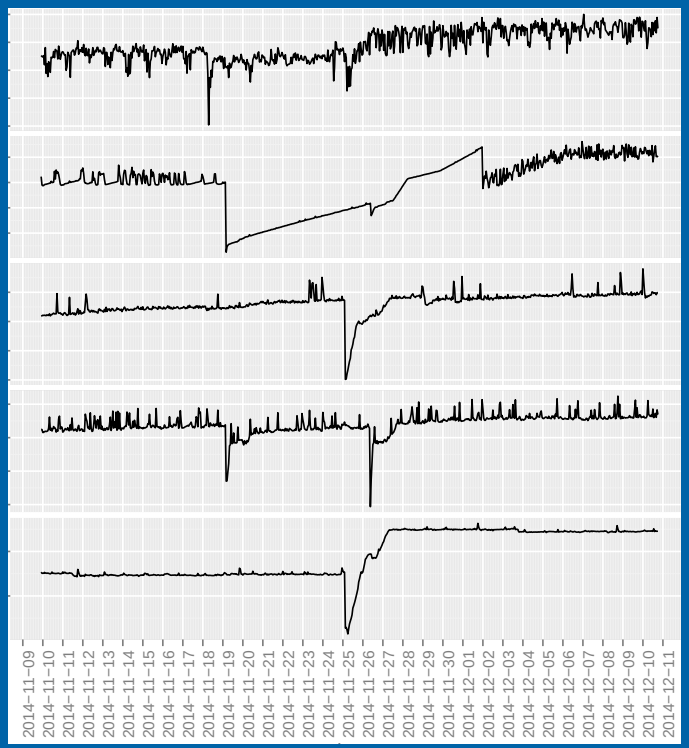
1. Time series features
2. Finding anomalies
3. Irish smart metre data
4. Finding typical and unusual households
5. Visualization via embedding

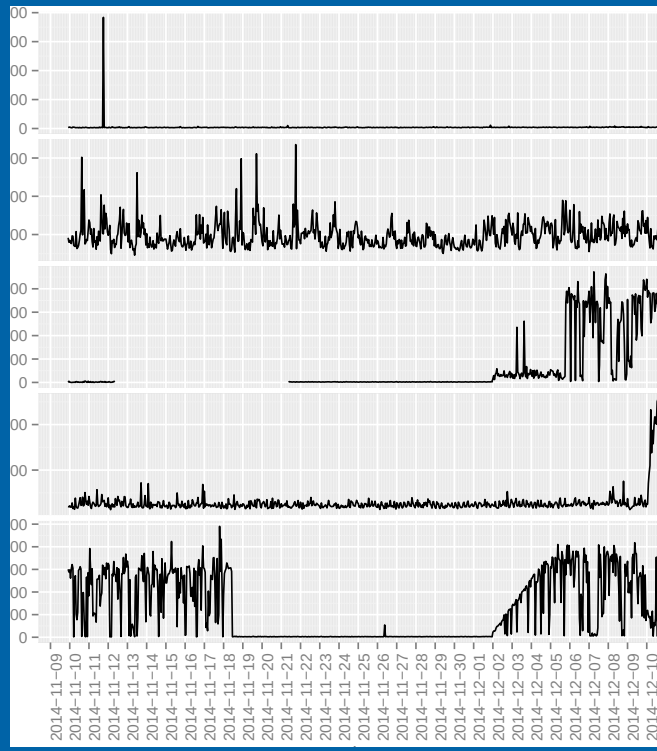
Yahoo web-traffic

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

Yahoo web-traffic



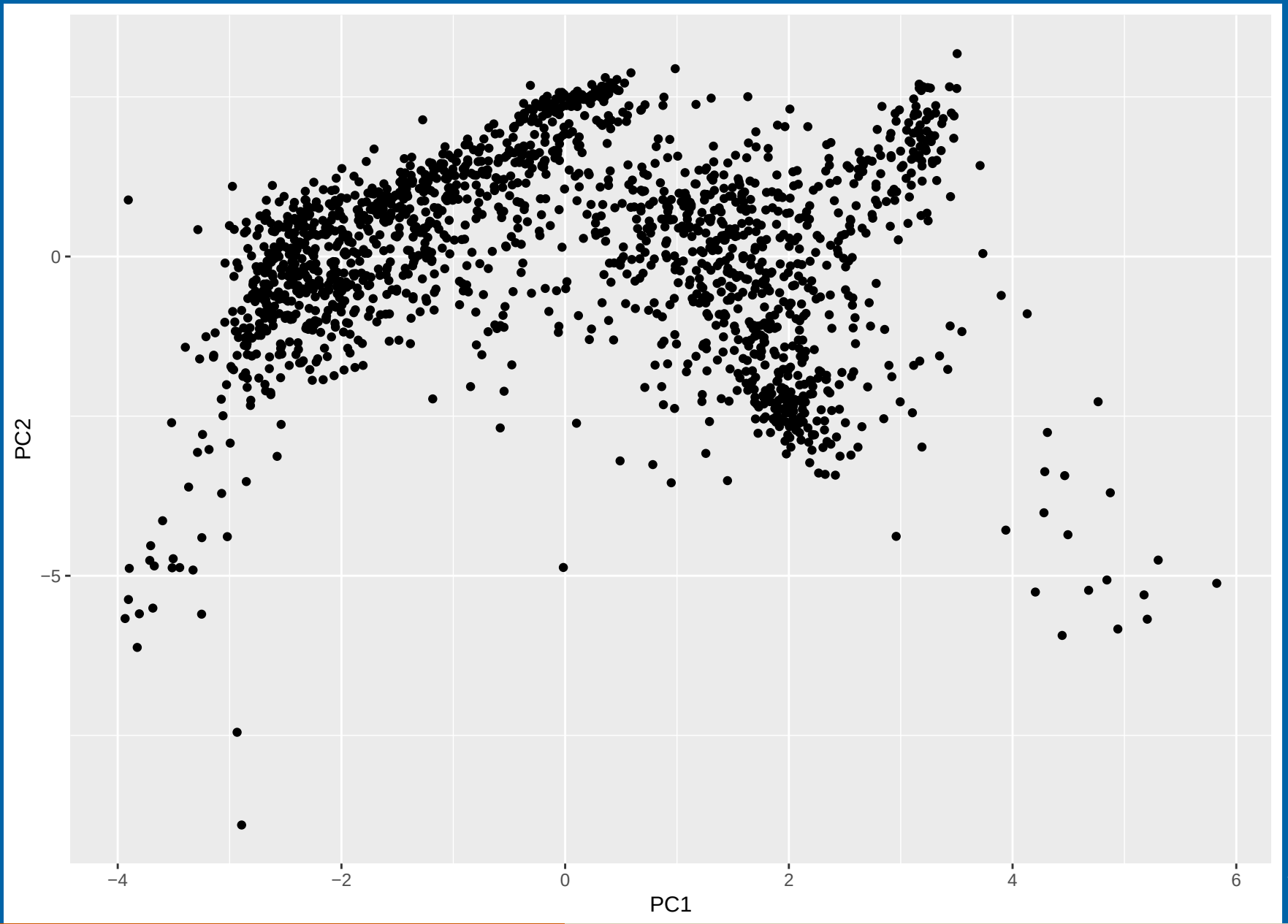




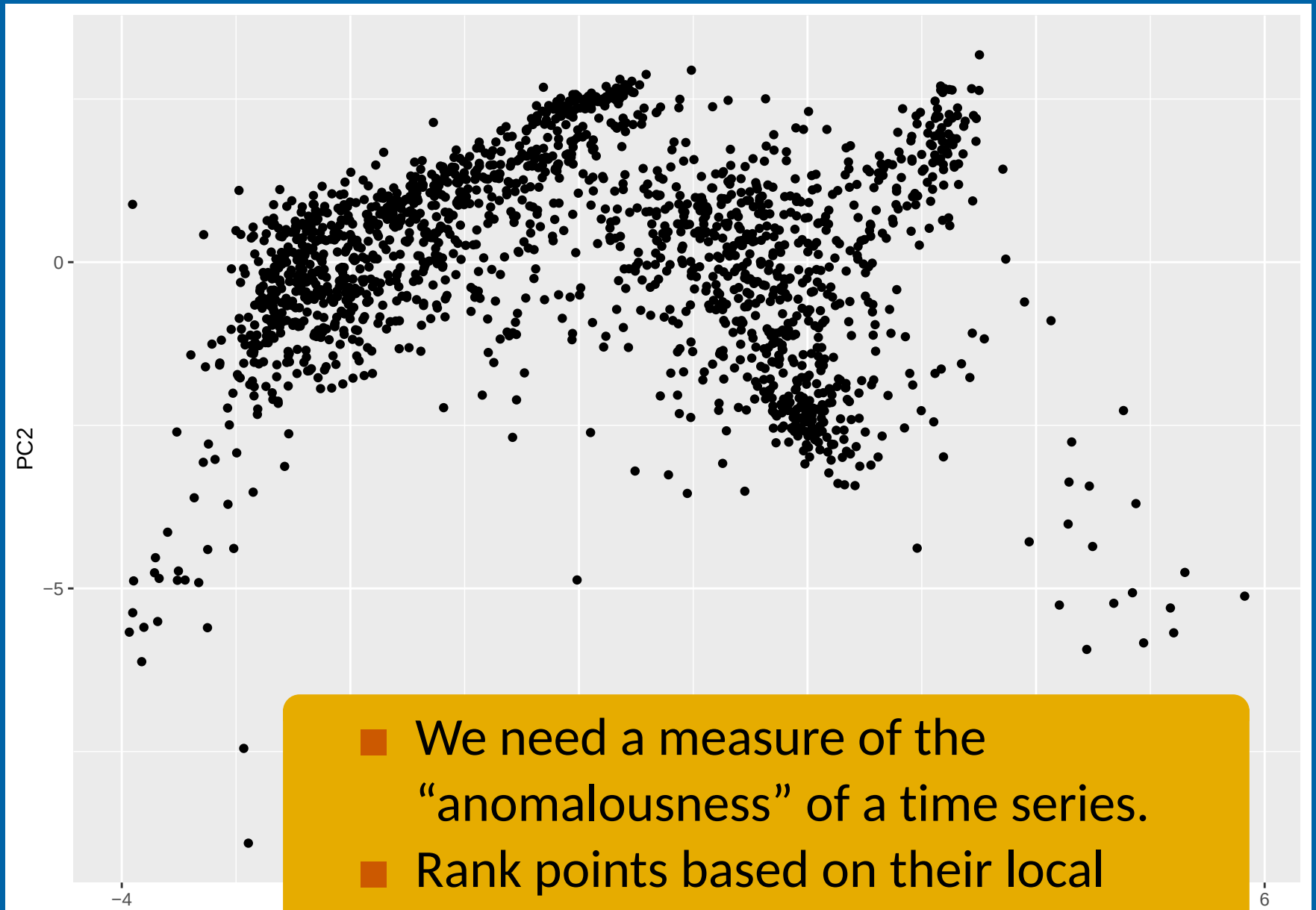
Feature space

- **ACF1**: first order autocorrelation = $\text{Corr}(Y_t, Y_{t-1})$
- Strength of **trend** and **seasonality** based on STL
- Trend **linearity** and **curvature**
- 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 of size 48.
- **Change index**: Time of maximum KL score

Principal component analysis



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

Bivariate kernel density

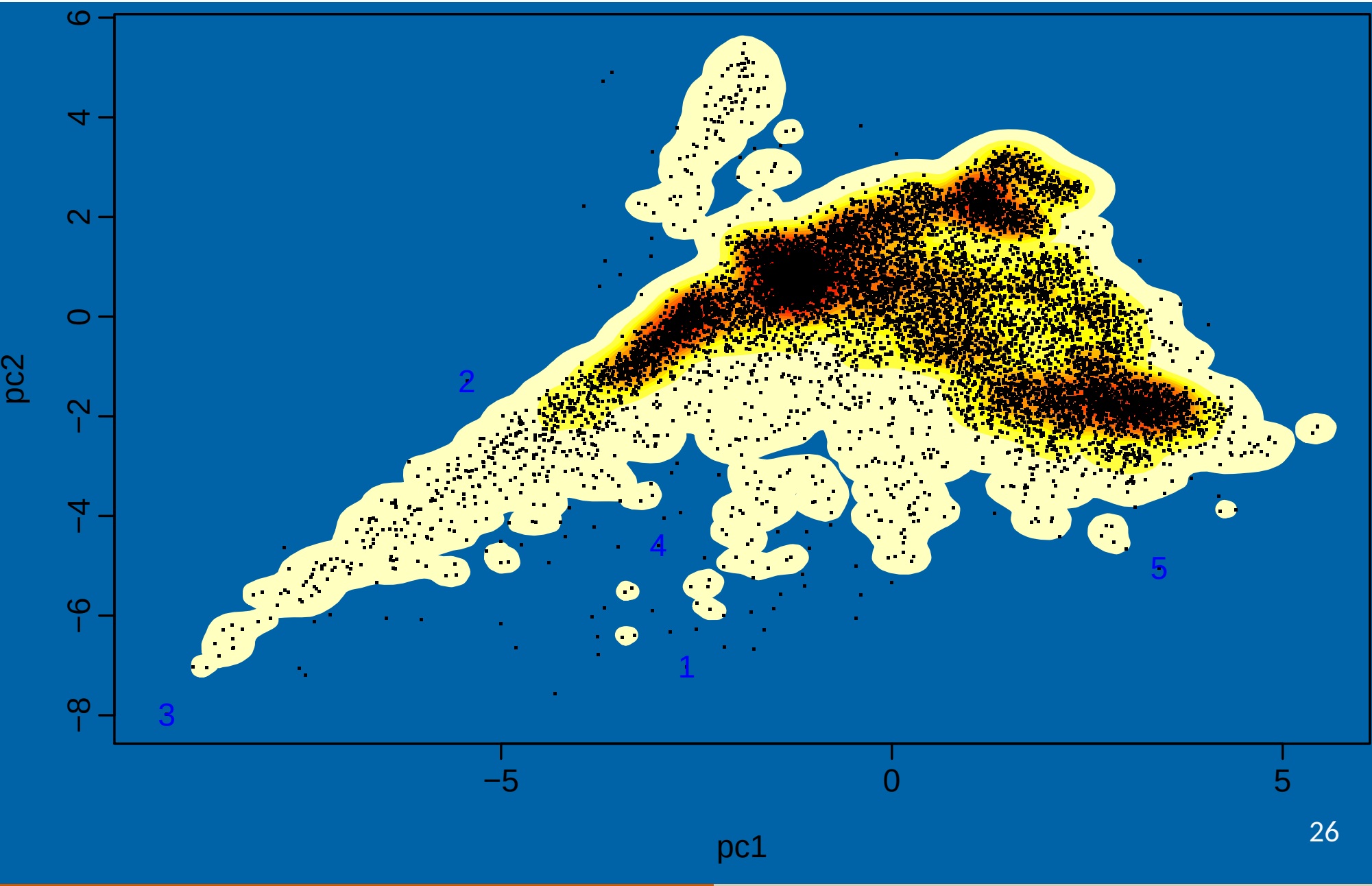
$$\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

Bivariate kernel density

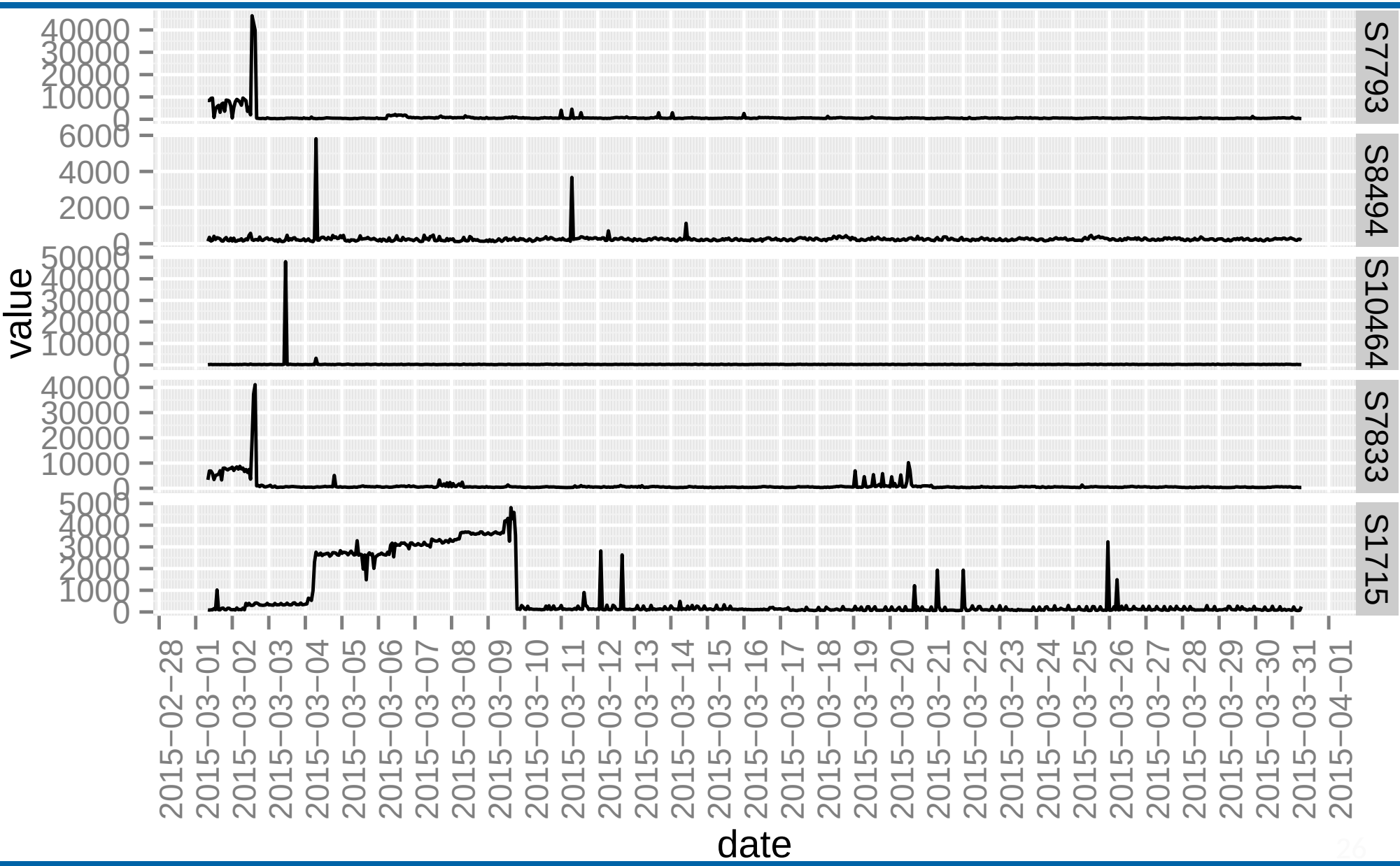
$$\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

- $\mathbf{X}_i \in$ a bivariate random sample $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$
- $K_{\mathbf{H}}(\mathbf{x})$ is the standard normal kernel function
- $\mathbf{H}$ estimated by minimizing the sum of AMISE
- Rank points based on $\hat{f}$ values in 2d PCA space.

Bivariate density ranking

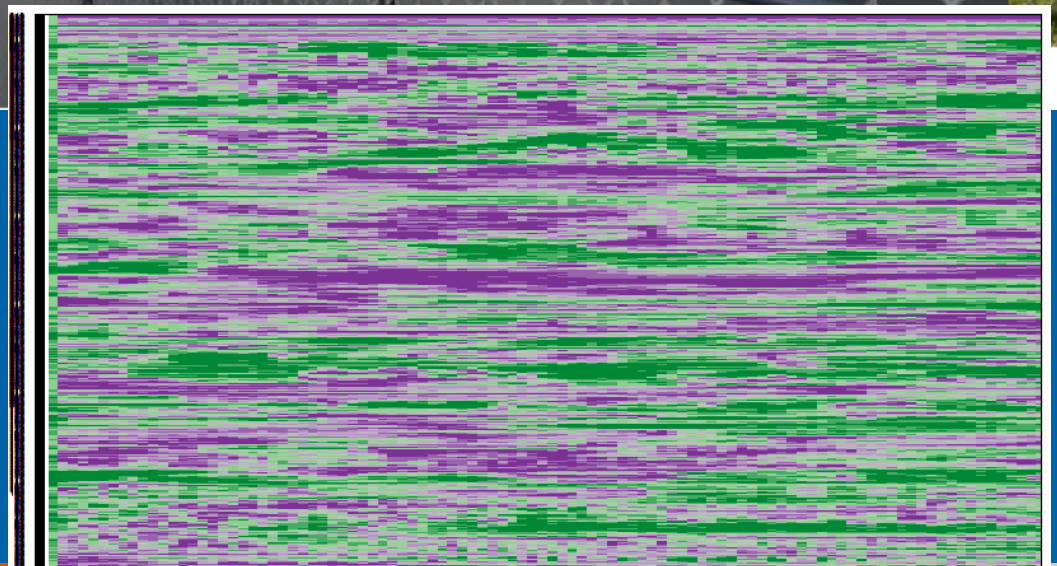


Bivariate density ranking



Security monitoring

Security monitoring



Time series anomaly detection

- Density-based outliers vs distance-based outliers.
- Fast feature calculation using windows needed for streaming data.
- **Oddstream**: Density-based anomalies, requiring a “typical” training period.
- **Stray**: Distance-based anomalies, requiring no “typical” training period.

Papers and packages

Hyndman, Wang & Laptev (2015). Large-scale unusual time series detection. *Proceedings of the IEEE International Conference on Data Mining*.

Talagala, Hyndman, Smith-Miles, Kandanaarachchi & Muñoz (2018) Anomaly detection in streaming nonstationary temporal data.

robjhyndman.com/publications/oddstream/

Hyndman, Wang, Kang, Talagala &

Outline

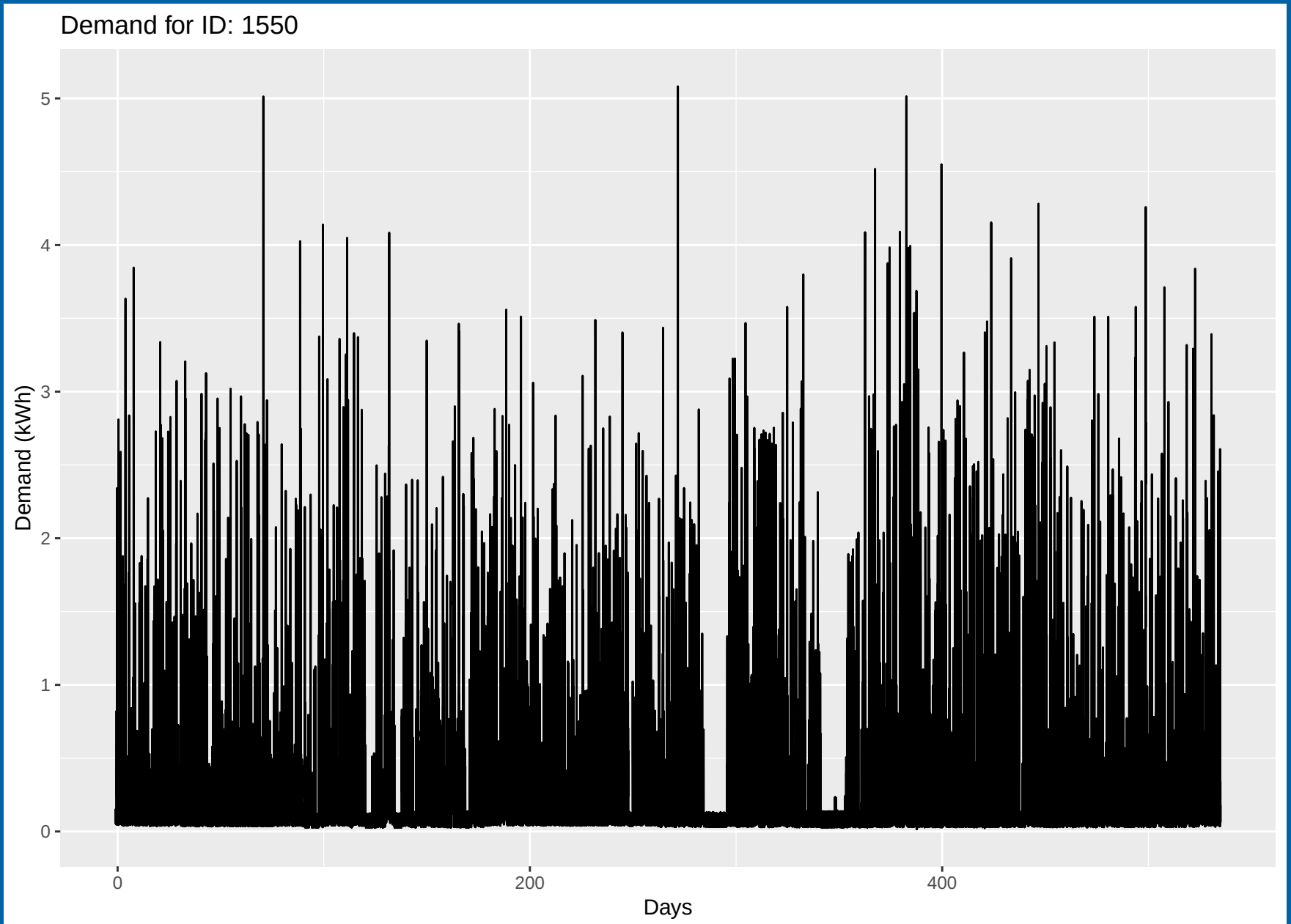
1. Time series features
2. Finding anomalies
3. Irish smart metre data
4. Finding typical and unusual households
5. Visualization via embedding

Irish smart metre data

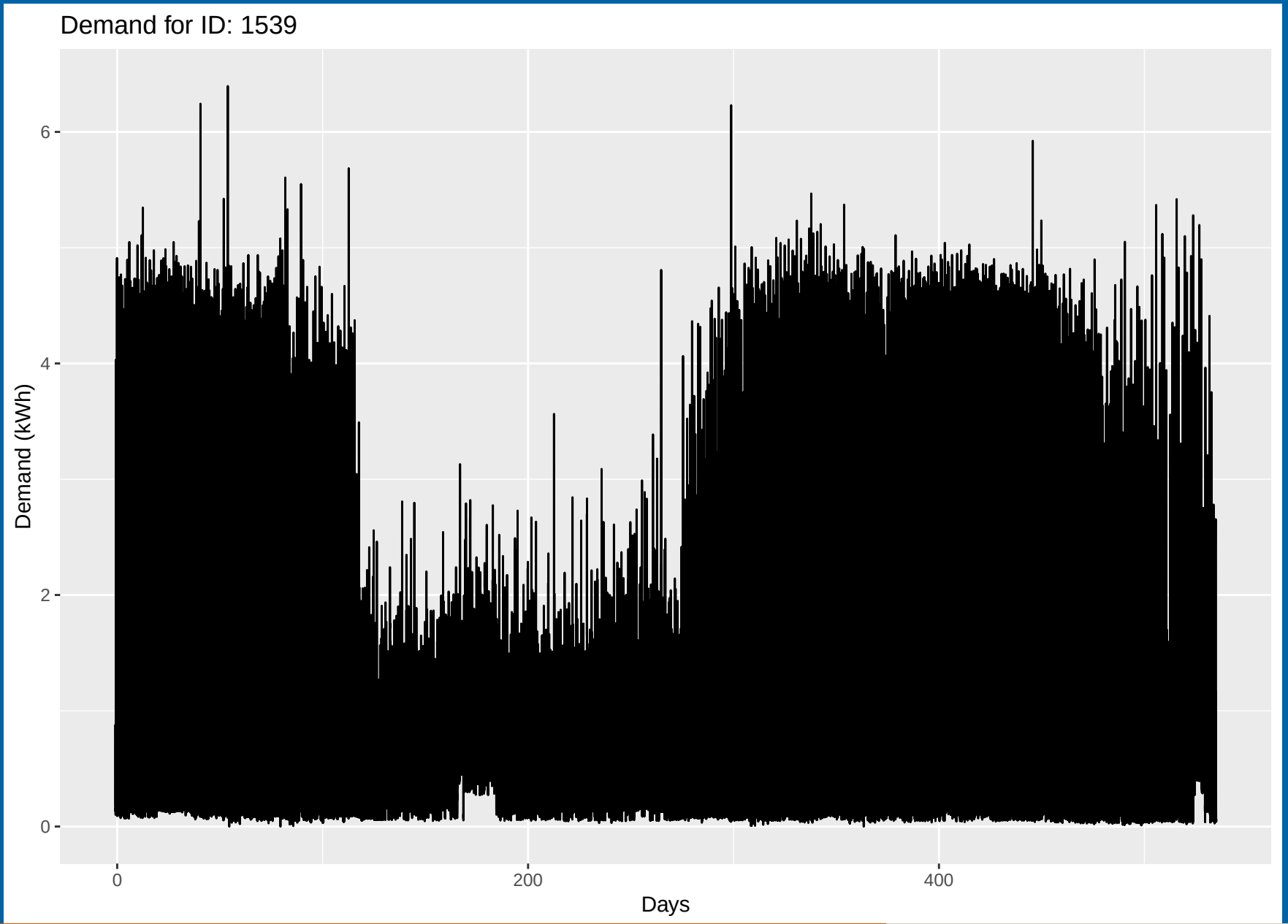
Figure: http://solutions.3m.com

- 500 households from smart metering trial
- Electricity consumption at 30-minute intervals between 14 July 2009 and 31 December 2010
- Heating/cooling energy usage excluded

Irish smart metre data

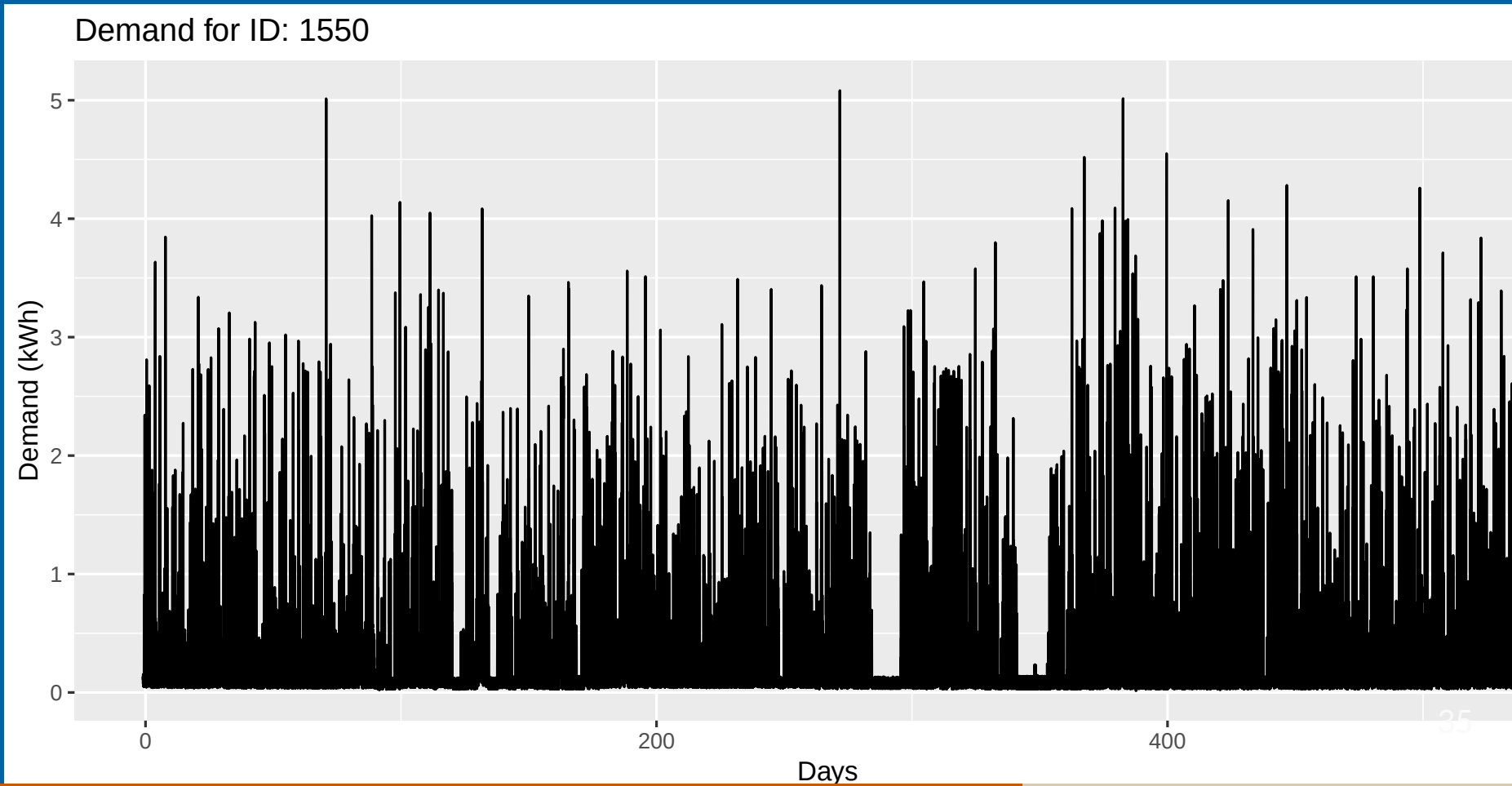


Irish smart metre data



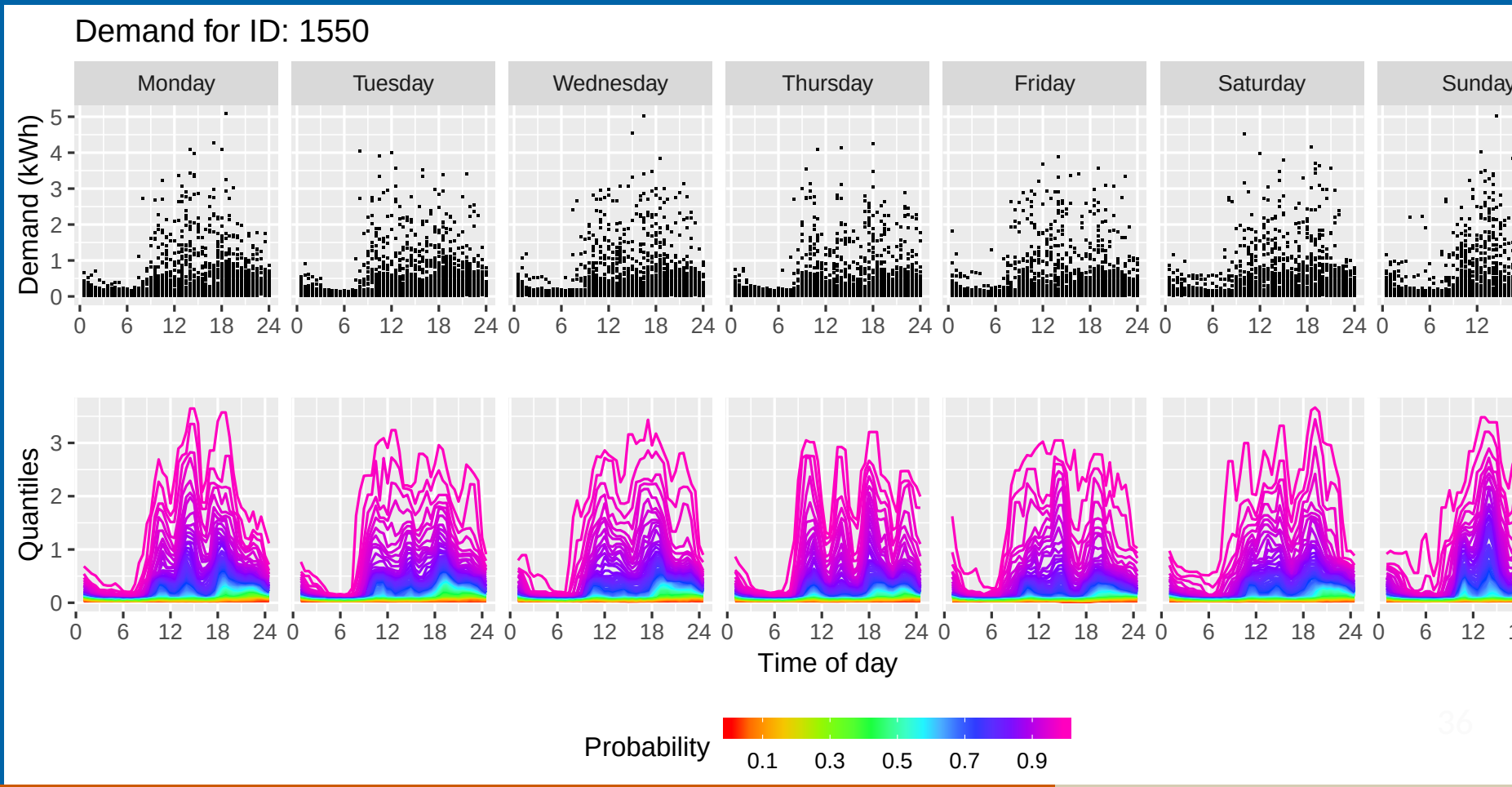
Quantiles conditional on time of week

- Compute sample quantiles at $p = 0.01, 0.02, \ldots, 0.99$ for each household and each half-hour of the week.
- 336 probability distributions per household.



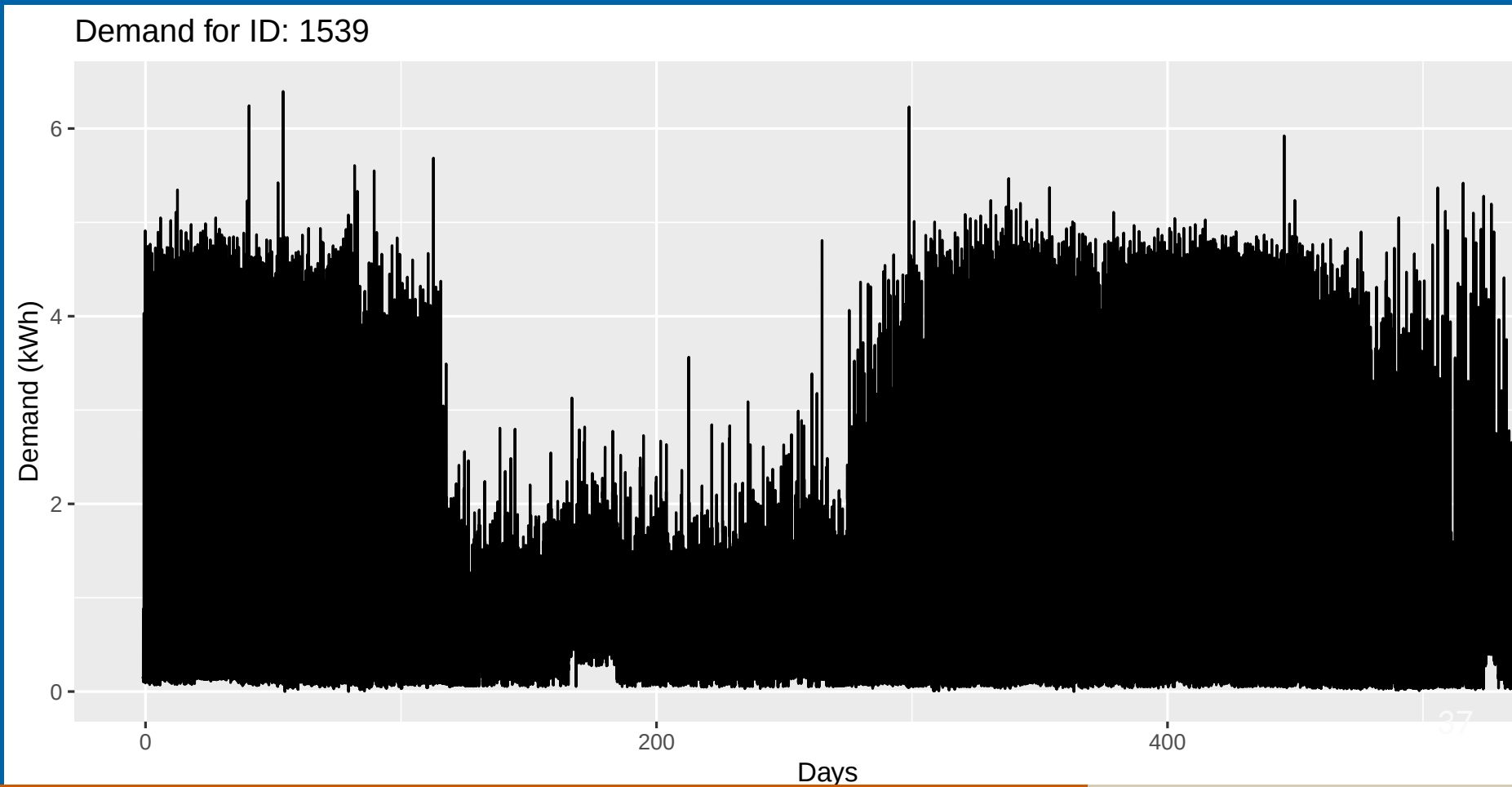
Quantiles conditional on time of week

- Compute sample quantiles at $p = 0.01, 0.02, \ldots, 0.99$ for each household and each half-hour of the week.
- 336 probability distributions per household.



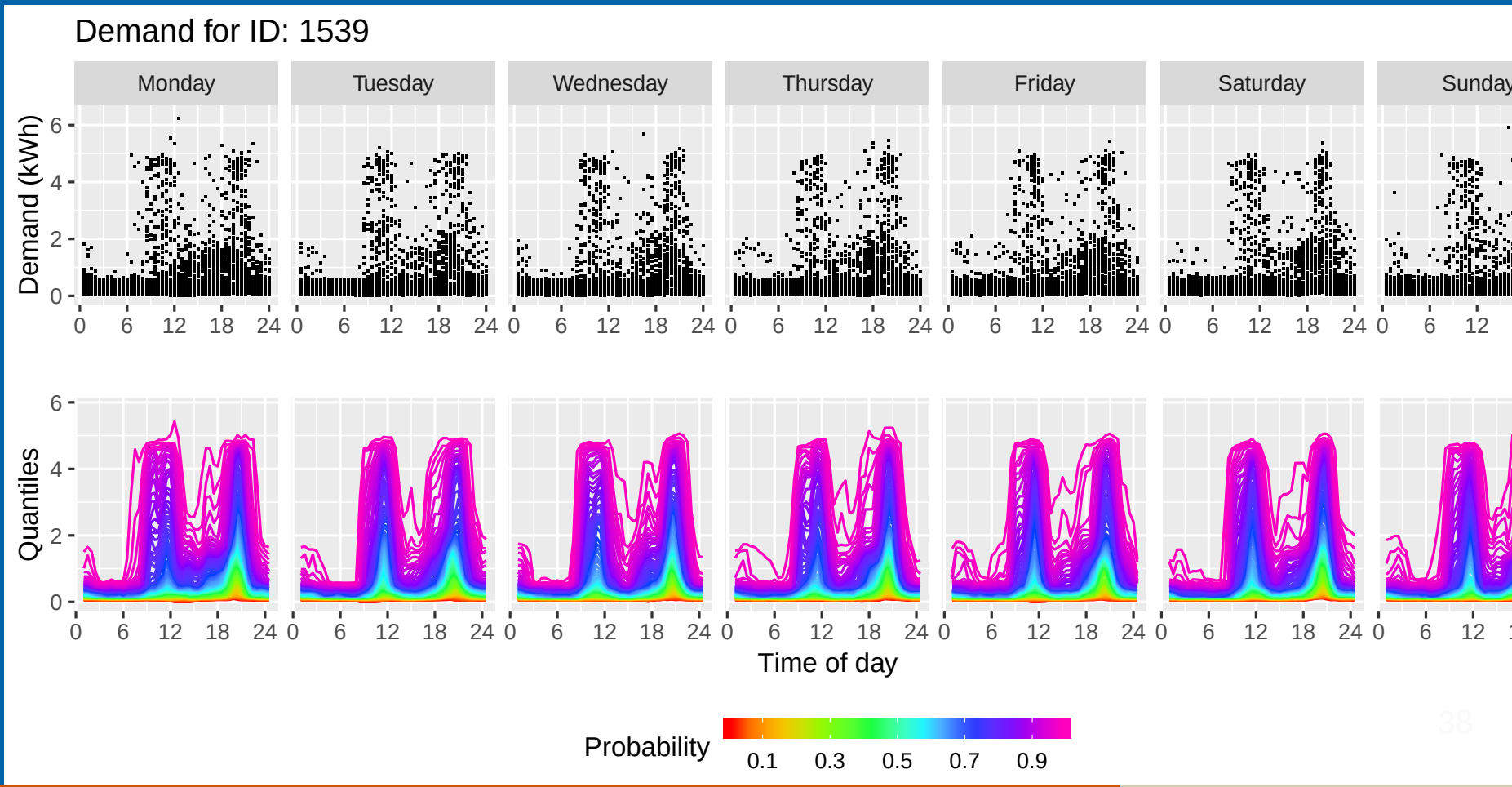
Quantiles conditional on time of week

- Compute sample quantiles at $p = 0.01, 0.02, \ldots, 0.99$ for each household and each half-hour of the week.
- 336 probability distributions per household.

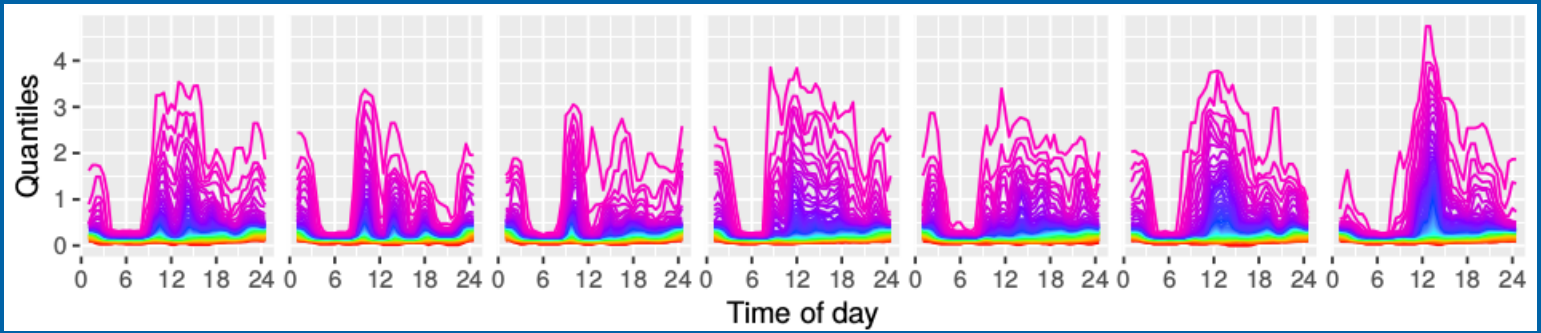


Quantiles conditional on time of week

- Compute sample quantiles at $p = 0.01, 0.02, \ldots, 0.99$ for each household and each half-hour of the week.
- 336 probability distributions per household.

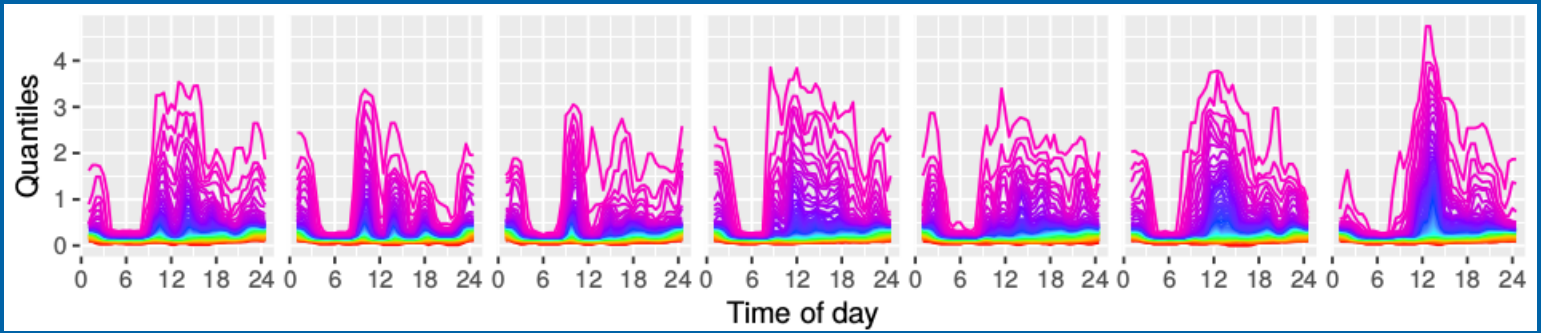


Quantiles conditional on time of week



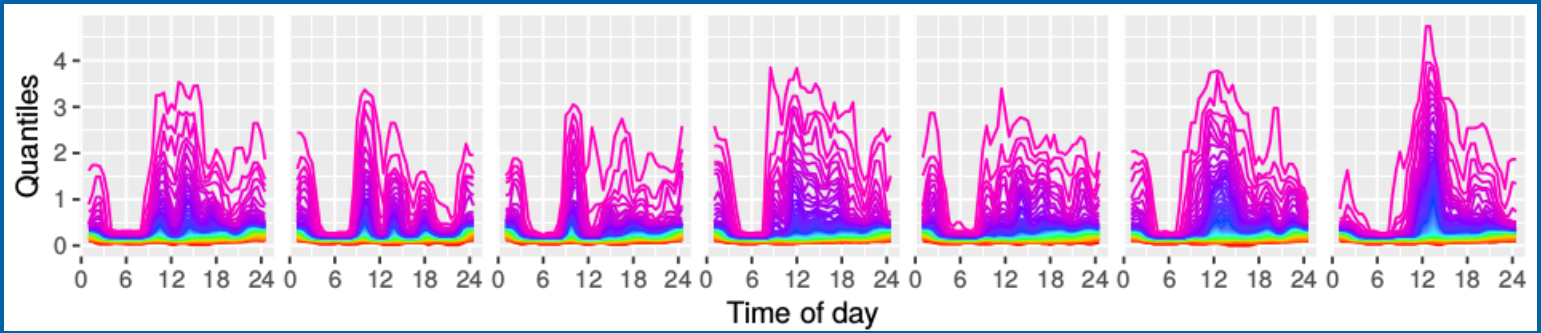
- Sample quantiles better than kernel density estimate:

Quantiles conditional on time of week



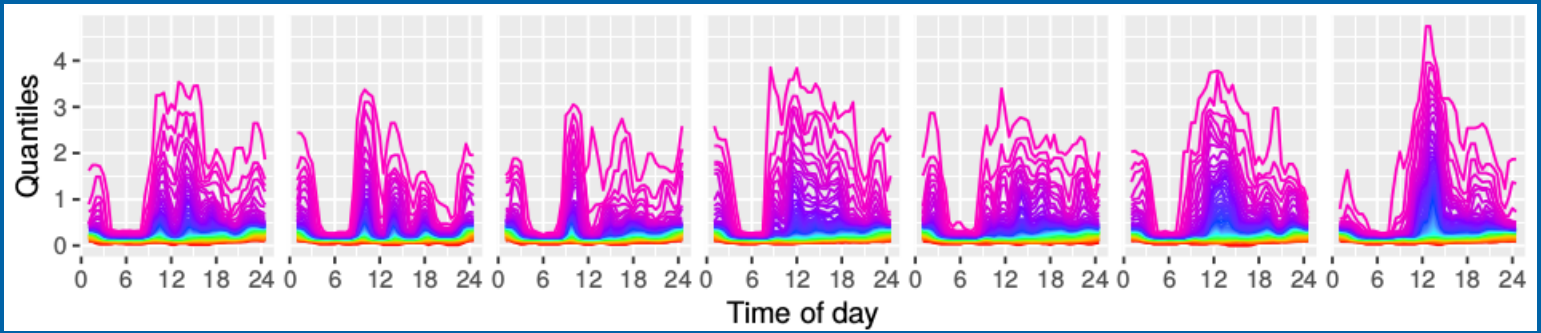
- Sample quantiles better than kernel density estimate:
- presence of zeros

Quantiles conditional on time of week



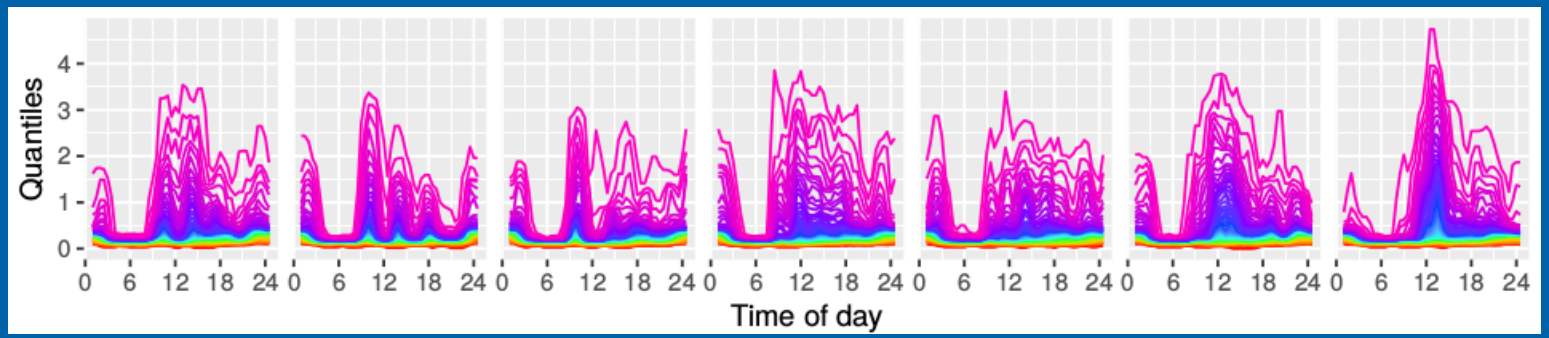
- Sample quantiles better than kernel density estimate:
- presence of zeros
- non-negative support

Quantiles conditional on time of week



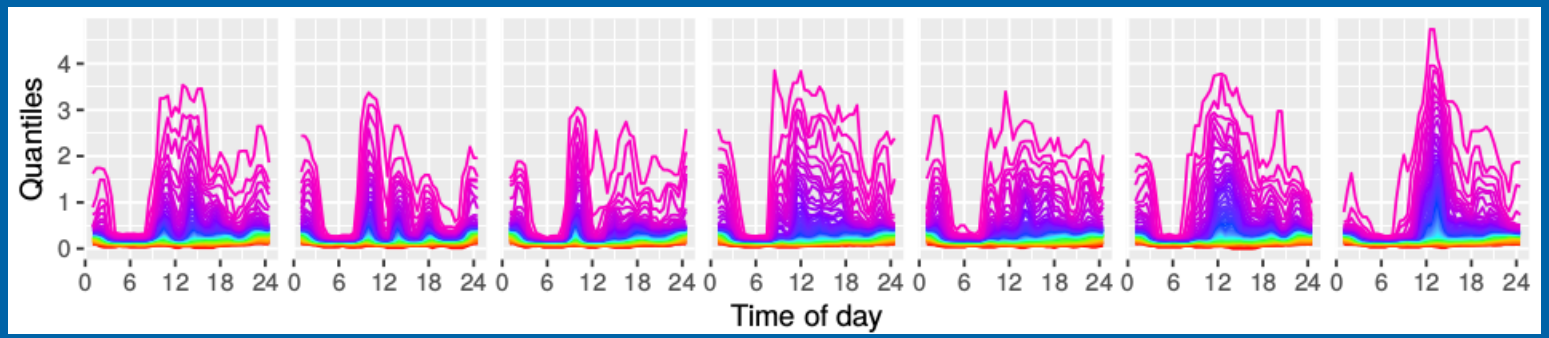
- Sample quantiles better than kernel density estimate:
- presence of zeros
- non-negative support
- high skewness

Quantiles conditional on time of week



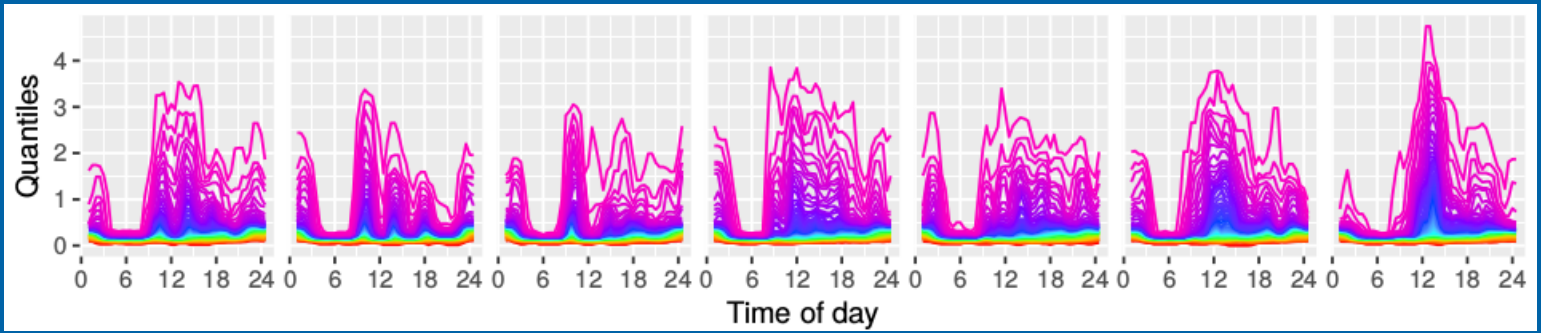
- Sample quantiles better than kernel density estimate:
- presence of zeros
- non-negative support
- high skewness
- Avoids missing data issues and variation in series length

Quantiles conditional on time of week



- Sample quantiles better than kernel density estimate:
- presence of zeros
- non-negative support
- high skewness
- Avoids missing data issues and variation in series length
- Avoids timing of household events, holidays, etc.

Quantiles conditional on time of week

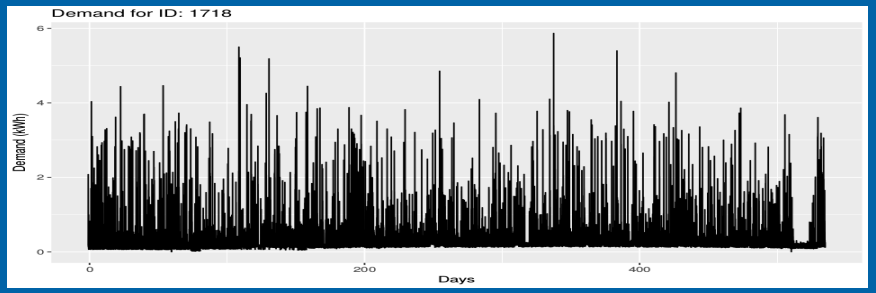


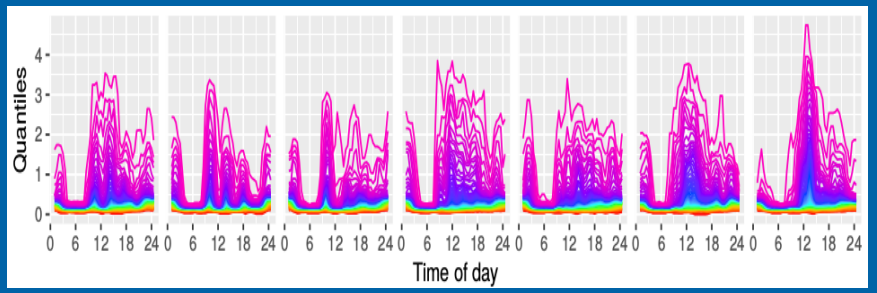
- Sample quantiles better than kernel density estimate:
- presence of zeros
- non-negative support
- high skewness
- Avoids missing data issues and variation in series length
- Avoids timing of household events, holidays, etc.
- Allows clustering of households based on probabilistic behaviour rather than coincident behaviour.

Outline

1. Time series features
2. Finding anomalies
3. Irish smart metre data
4. Finding typical and unusual households
5. Visualization via embedding

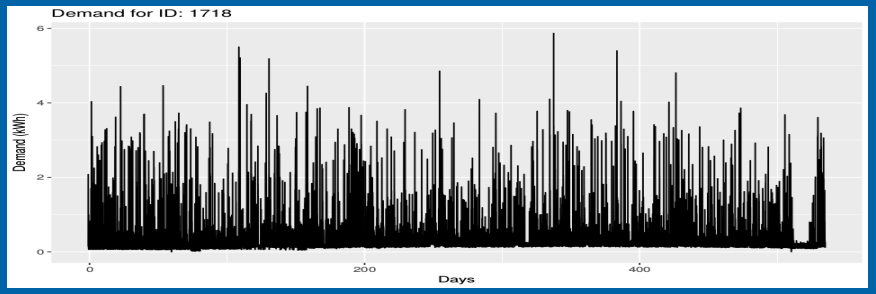
Pairwise distances

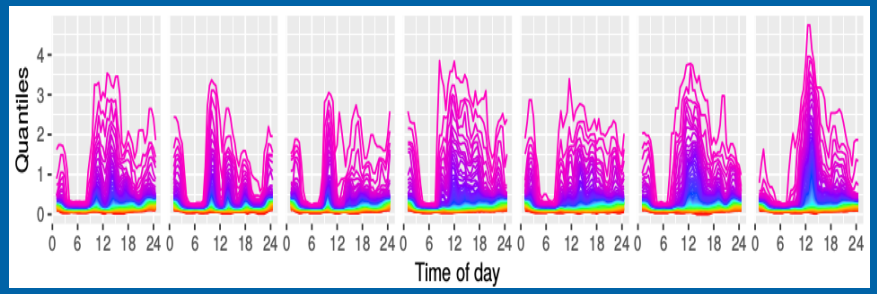




The time series of 535×48 observations per household is mapped to a set of $7 \times 48 \times 99$ quantiles giving a bivariate surface for each household.

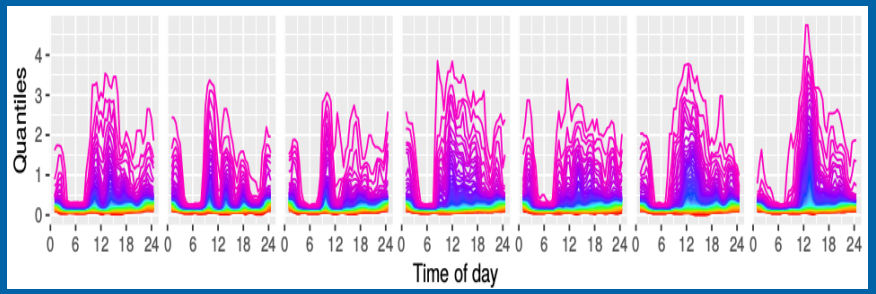
Pairwise distances



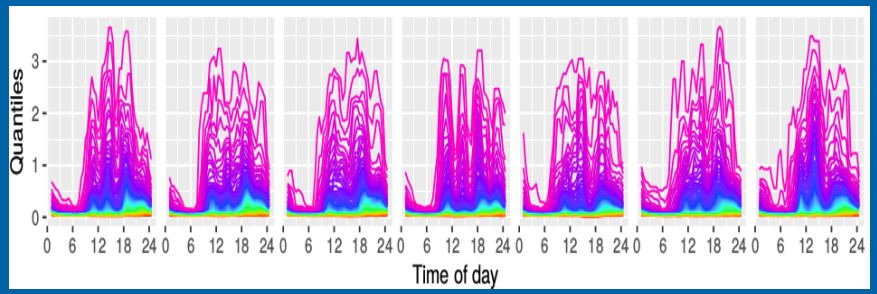


The time series of 535×48 observations per household is mapped to a set of $7 \times 48 \times 99$ quantiles giving a bivariate surface for each household.

Can we compute pairwise distances between all households?



$\leftarrow$? $\rightarrow$
Distance



Jensen-Shannon distances

Kullback-Leibler divergence between two densities

$$D(p,q) = \int_{\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Jensen-Shannon distances

Kullback-Leibler divergence between two densities

$$D(p,q) = \int_{\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Not symmetric: $D(p,q) \neq D(q,p)$

Jensen-Shannon distances

Kullback-Leibler divergence between two densities

$$D(p,q) = \int_{\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Not symmetric: $D(p,q) \neq D(q,p)$

Jensen-Shannon distance between two densities

$JS(p,q) = [D(p,r) + D(q,r)]/2$ where $r = (p+q)/2$

Jensen-Shannon distances

Kullback-Leibler divergence between two densities

$$D(p,q) = \int_{\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Not symmetric: $D(p,q) \neq D(q,p)$

Jensen-Shannon distance between two densities

$$\mathrm{JS}(p,q) = [D(p,r) + D(q,r)]/2 \qquad \text{where } r = (p+q)/2$$

Distance between two households

$$\Delta_{ij} = \sum_{t=1}^{7\times 48} \mathrm{JS}(p_t, q_t)$$

Kernel matrix and density ranking

Similarity between two households

$$w_{ij} = \exp(-\Delta_{ij}^2/h^2).$$

Kernel matrix and density ranking

Similarity between two households

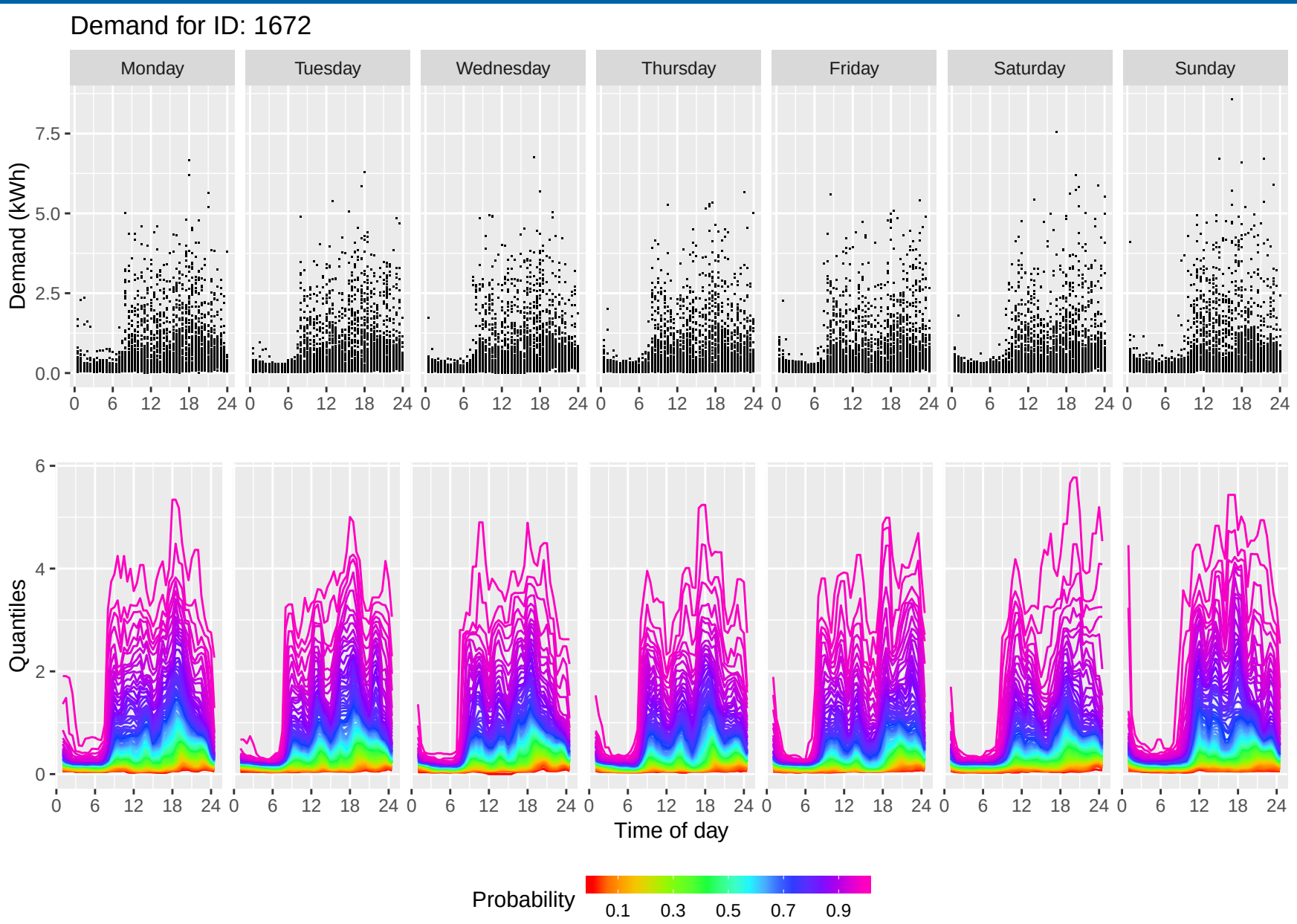
$$w_{ij} = \exp(-\Delta_{ij}^2/h^2).$$

Row sums of the kernel matrix gives a scaled kernel density estimate of households:

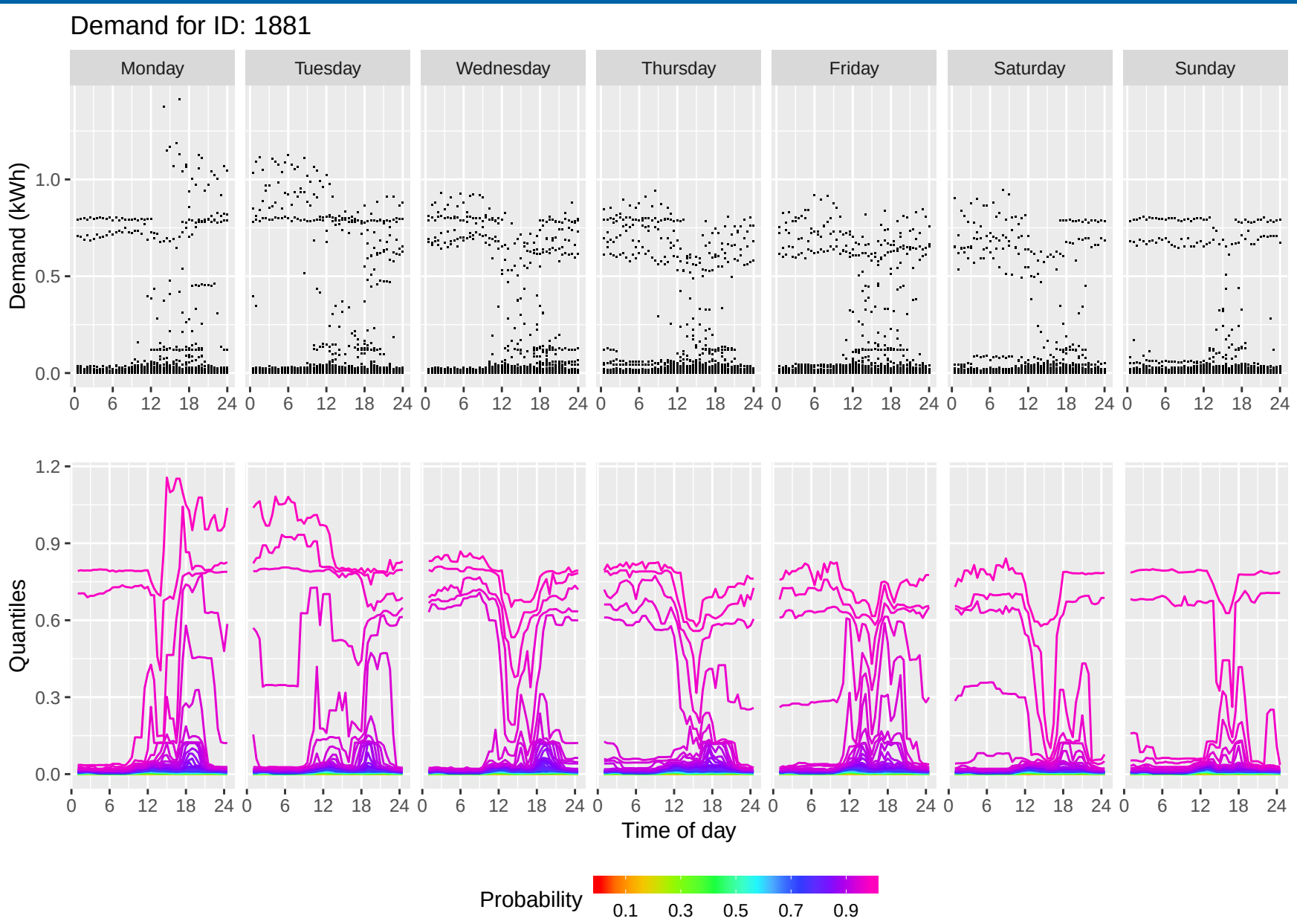
$$\hat{f}_i = \sum_{j=1}^{n} w_{ij}$$

- h is bandwidth in Gaussian kernel.
- Households can be ranked by density values.

Most typical household



Most anomalous household



Outline

1. Time series features
2. Finding anomalies
3. Irish smart metre data
4. Finding typical and unusual households
5. Visualization via embedding

Laplacian eigenmaps

- **Idea:** Embed conditional densities in a 2d space where the distances are preserved “as far as possible”.

Laplacian eigenmaps

- **Idea:** Embed conditional densities in a 2d space where the distances are preserved "as far as possible".
- Let $\boldsymbol{W} = [w_{ij}]$ where $w_{ij} = \exp(-\Delta_{ij}^2/h^2)$.

  $\boldsymbol{D} = \text{diag}(\hat{f}_i)$ where $\hat{f}_i = \sum_{j=1}^{n} w_{ij}$

  $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ (the Laplacian matrix).

Laplacian eigenmaps

- **Idea:** Embed conditional densities in a 2d space where the distances are preserved "as far as possible".
- Let $\boldsymbol{W} = [w_{ij}]$ where $w_{ij} = \exp(-\Delta_{ij}^2/h^2)$.

  $\boldsymbol{D} = \mathrm{diag}(\hat{f}_i)$ where $\hat{f}_i = \sum_{j=1}^{n} w_{ij}$

  $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ (the Laplacian matrix).
- Solve generalized eigenvector problem: $\boldsymbol{Le} = \lambda \boldsymbol{De}$.

Laplacian eigenmaps

- **Idea:** Embed conditional densities in a 2d space where the distances are preserved "as far as possible".
- Let $\boldsymbol{W} = [w_{ij}]$ where $w_{ij} = \exp(-\Delta_{ij}^2/h^2)$.

  $\boldsymbol{D} = \mathrm{diag}(\hat{f}_i)$ where $\hat{f}_i = \sum_{j=1}^{n} w_{ij}$

  $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ (the Laplacian matrix).
- Solve generalized eigenvector problem: $\boldsymbol{Le} = \lambda \boldsymbol{De}$.
- Let $\boldsymbol{e}_k$ be eigenvector corresponding to kth *smallest* eigenvalue.

Laplacian eigenmaps

- **Idea:** Embed conditional densities in a 2d space where the distances are preserved "as far as possible".
- Let $\boldsymbol{W} = [w_{ij}]$ where $w_{ij} = \exp(-\Delta_{ij}^2/h^2)$.

  $\boldsymbol{D} = \text{diag}(\hat{f}_i)$ where $\hat{f}_i = \sum_{j=1}^{n} w_{ij}$

  $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ (the Laplacian matrix).
- Solve generalized eigenvector problem: $\boldsymbol{Le} = \lambda \boldsymbol{De}$.
- Let $\boldsymbol{e}_k$ be eigenvector corresponding to kth *smallest* eigenvalue.
- Then $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$ create an embedding of households in 2d space.

Key property of Laplacian embedding

Let $y_i = (e_{2,i}, e_{3,i})$ be the embedded point corresponding to household i.

Then the Laplacian eigenmap minimizes

$$\sum_{ij} w_{ij}(y_i - y_j)^2 = \boldsymbol{y}'\boldsymbol{L}\boldsymbol{y} \qquad \text{such that} \quad \boldsymbol{y}'\boldsymbol{D}\boldsymbol{y} = 1.$$

Key property of Laplacian embedding

Let $y_i = (e_{2,i}, e_{3,i})$ be the embedded point corresponding to household i.

Then the Laplacian eigenmap minimizes

$$\sum_{ij} w_{ij}(y_i - y_j)^2 = \boldsymbol{y}'\boldsymbol{L}\boldsymbol{y} \qquad \text{such that} \quad \boldsymbol{y}'\boldsymbol{D}\boldsymbol{y} = 1.$$

- the most similar points are as close as possible.

Key property of Laplacian embedding

Let $y_i = (e_{2,i}, e_{3,i})$ be the embedded point corresponding to household i.

Then the Laplacian eigenmap minimizes

$$\sum_{ij} w_{ij}(y_i - y_j)^2 = \boldsymbol{y}'\boldsymbol{L}\boldsymbol{y} \qquad \text{such that} \quad \boldsymbol{y}'\boldsymbol{D}\boldsymbol{y} = 1.$$

- the most similar points are as close as possible.
- First eigenvalue is 0 due to translation invariance.

Key property of Laplacian embedding

Let $y_i = (e_{2,i}, e_{3,i})$ be the embedded point corresponding to household i.

Then the Laplacian eigenmap minimizes

$$\sum_{ij} w_{ij}(y_i - y_j)^2 = \boldsymbol{y}'\boldsymbol{L}\boldsymbol{y} \qquad \text{such that} \quad \boldsymbol{y}'\boldsymbol{D}\boldsymbol{y} = 1.$$

- the most similar points are as close as possible.
- First eigenvalue is 0 due to translation invariance.
- Equivalent to optimal embedding using Laplace-Beltrami operator on manifolds.

Outliers shown in embedded space

Features and limitations

Features of approach

- Converting time series to quantile surfaces conditional on time of week.
- Using pairwise distances between households.
- Using kernel matrices for density ranking and embedding.

Features and limitations

Features of approach

- Converting time series to quantile surfaces conditional on time of week.
- Using pairwise distances between households.
- Using kernel matrices for density ranking and embedding.

Unresolved issues

- Need to select the bandwidth *h* in constructing the similarity matrix.
- Two different uses of bandwidth: density-ranking, embedding. Different bandwidth in each case?
- The use of pairwise distances makes it hard to scale this algorithm.

Features and limitations

Features of approach

- Converting time series to quantile surfaces conditional on time of week.
- Using pairwise distances between households.
- Using kernel matrices for density ranking and embedding.

Unresolved issues

- Need to select the bandwidth h in constructing the similarity matrix.
- Two different uses of bandwidth: density-ranking, embedding. Different bandwidth in each case?
- The use of pairwise distances makes it hard to scale this algorithm.