Model generation and selection for internet of things

Vadim Strijov

Moscow Institute of Physics and Technology

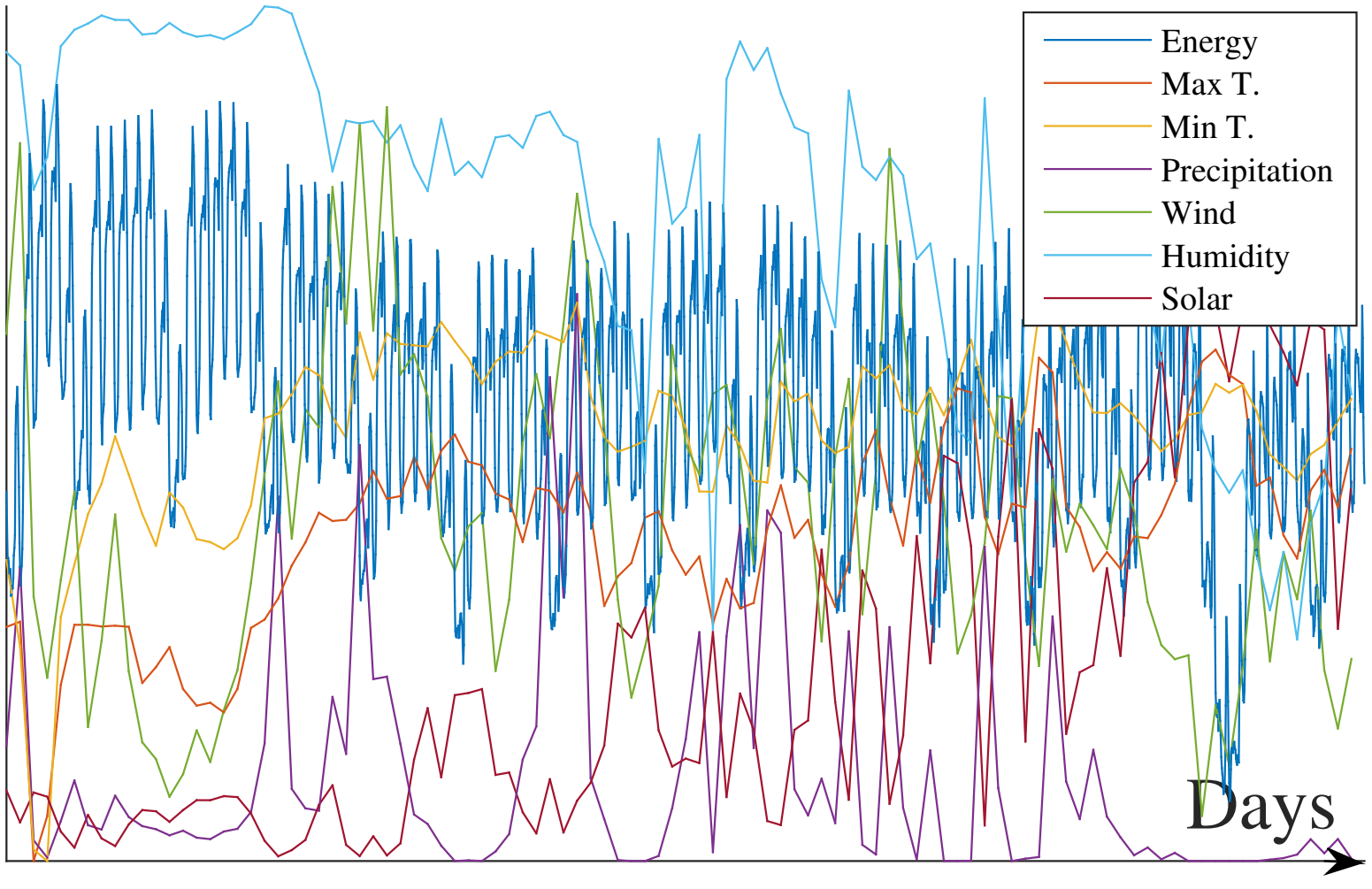
Data Fest, September 10–11, 2016

Internet of things

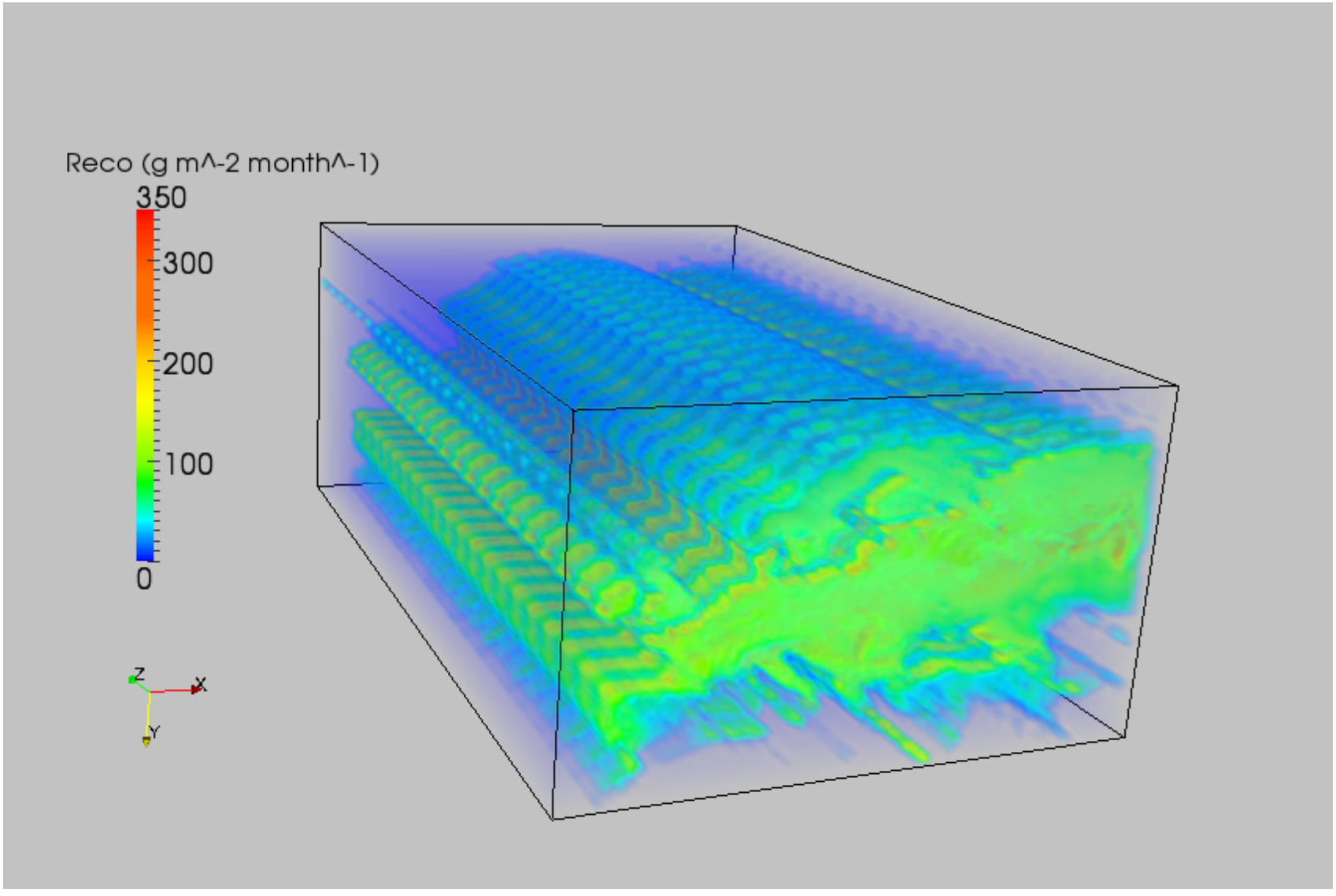
IoT is the networking of devices (portables, vehicles, buildings) embedded with sensors and software.

- Environment and energy monitoring
- Medical and health monitoring
- Consumer support, sales monitoring
- Urban management and manufactiring

Example of a multiscale dataset

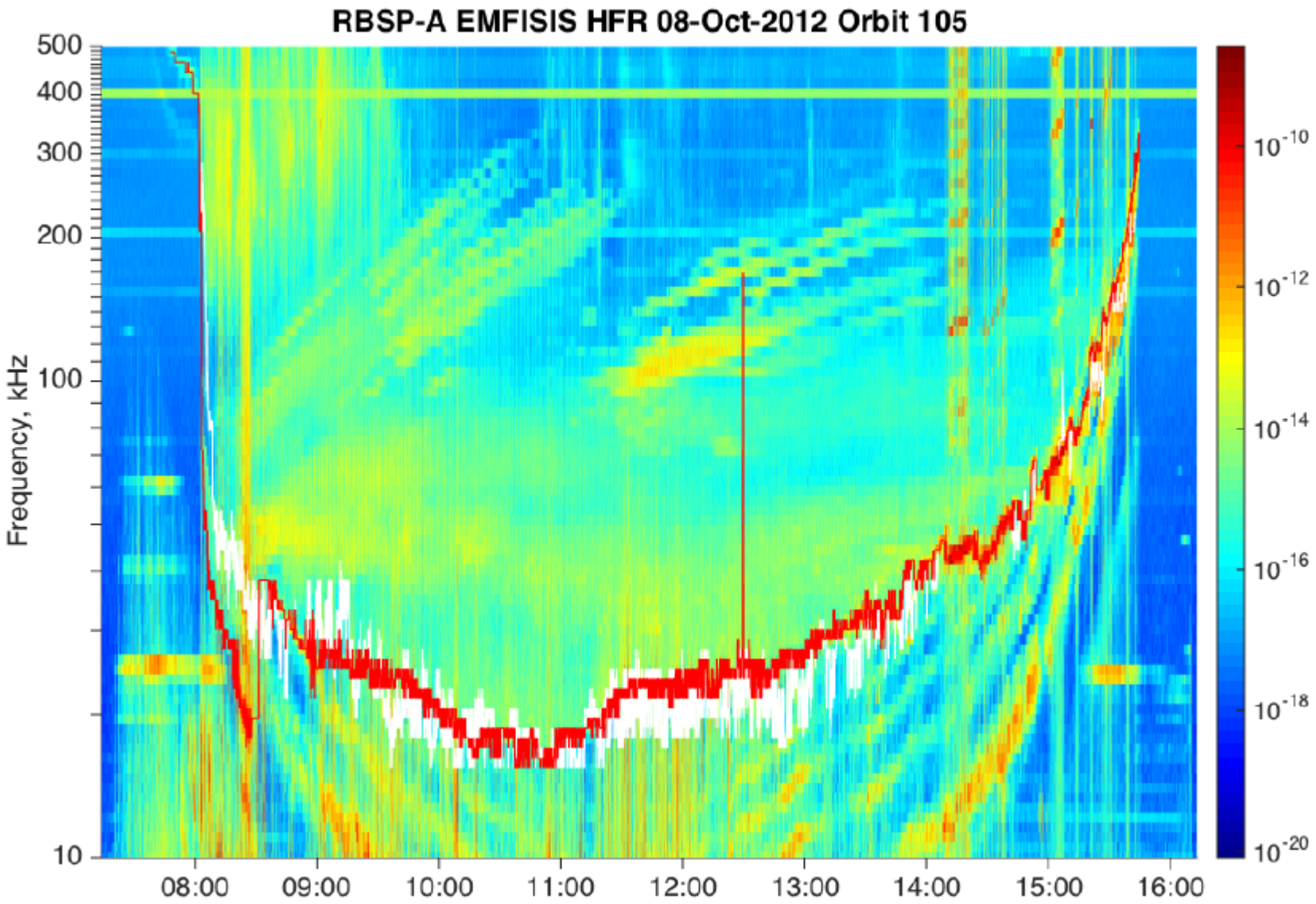


Example of a spatio temporal dataset



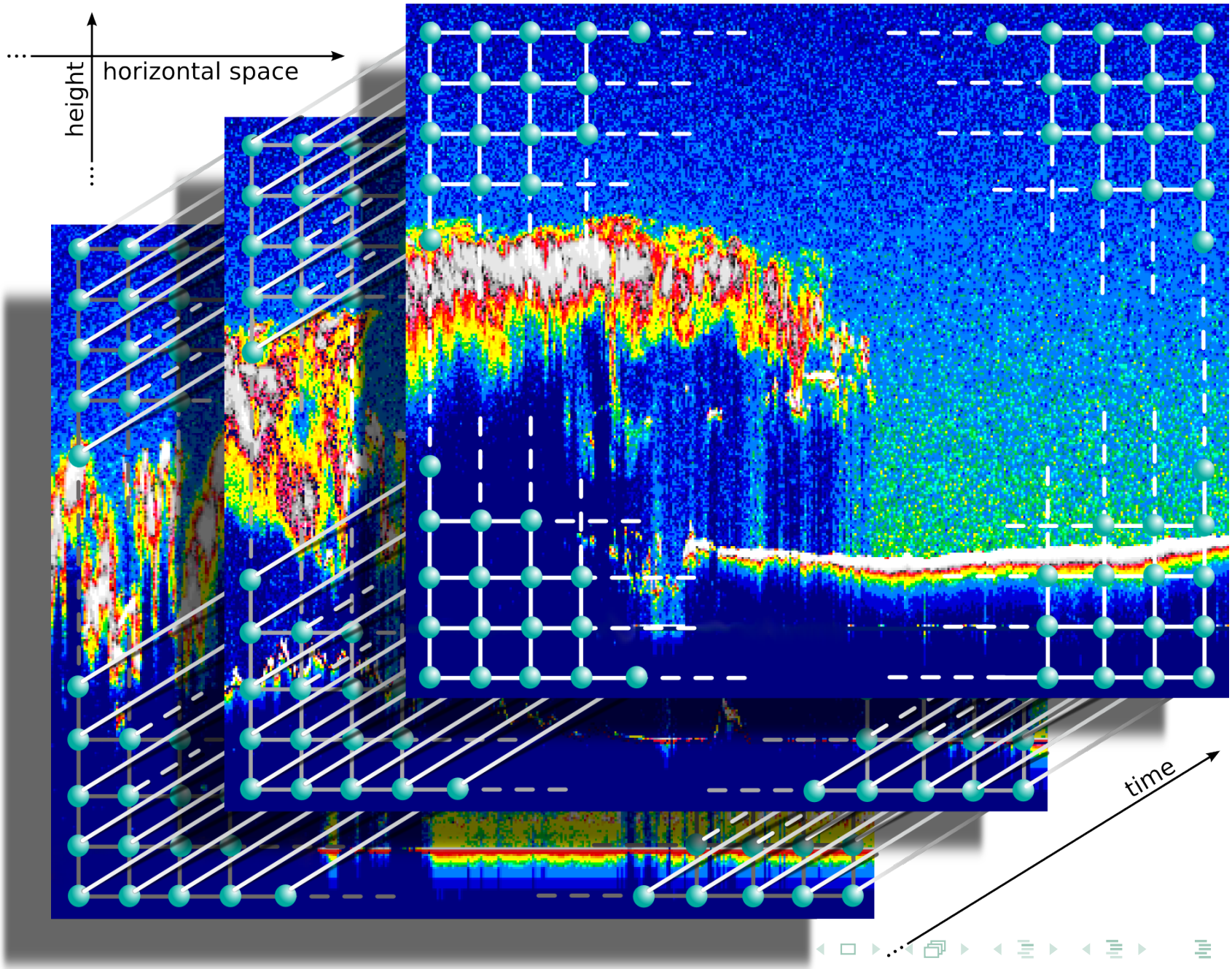
Monthly Ecosystem Respiration by M. Reichstein, GHG-Europe.

Example of a spatio temporal dataset



Electric field measurement, the Van Allen probes by I. Zhelavskaya, Skoltech.

Example of a spatio temporal dataset



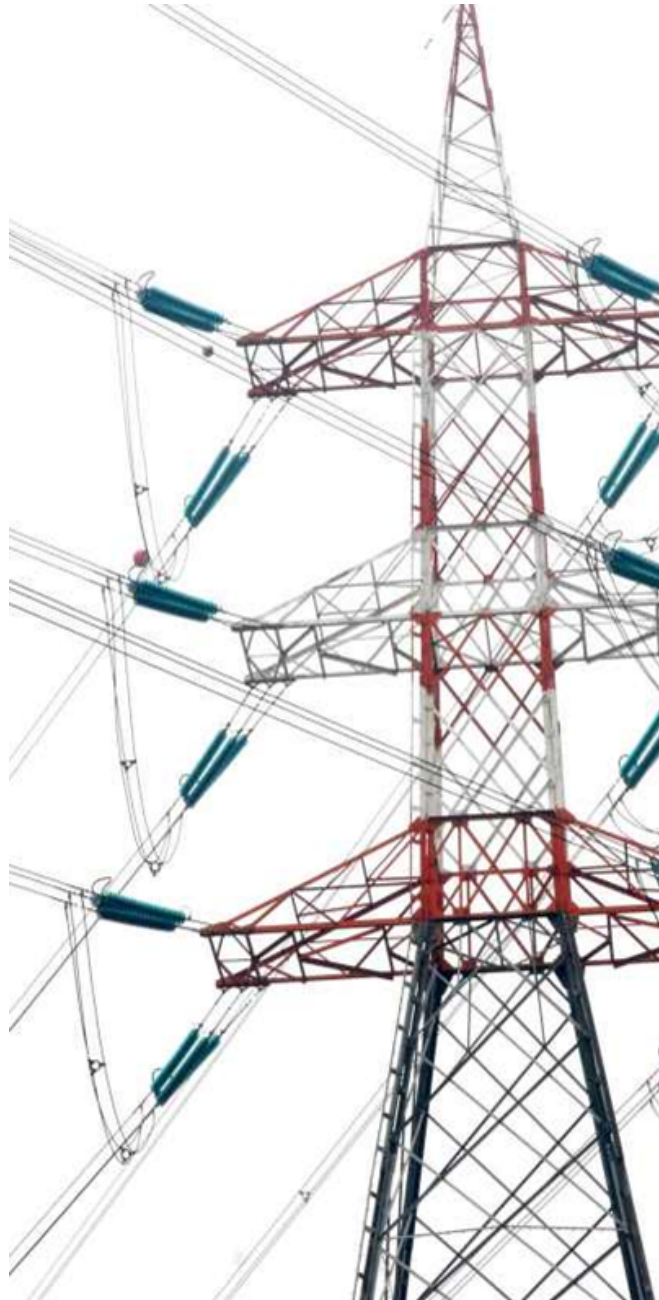
The periodic components of the multivariate time series

The time series:

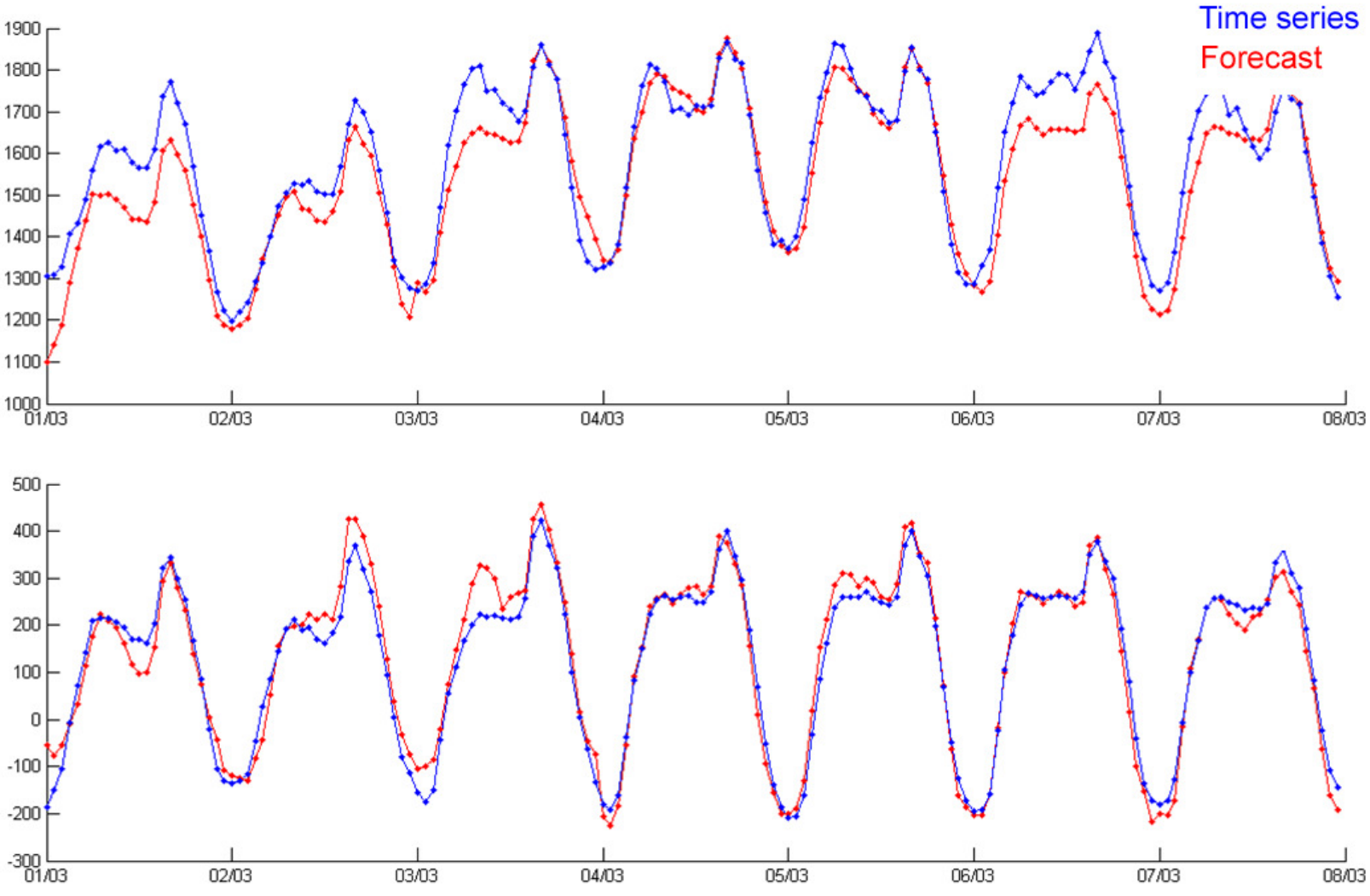
- energy price,
- consumption,
- daytime,
- temperature,
- humidity,
- wind force,
- holiday schedule.

Periods:

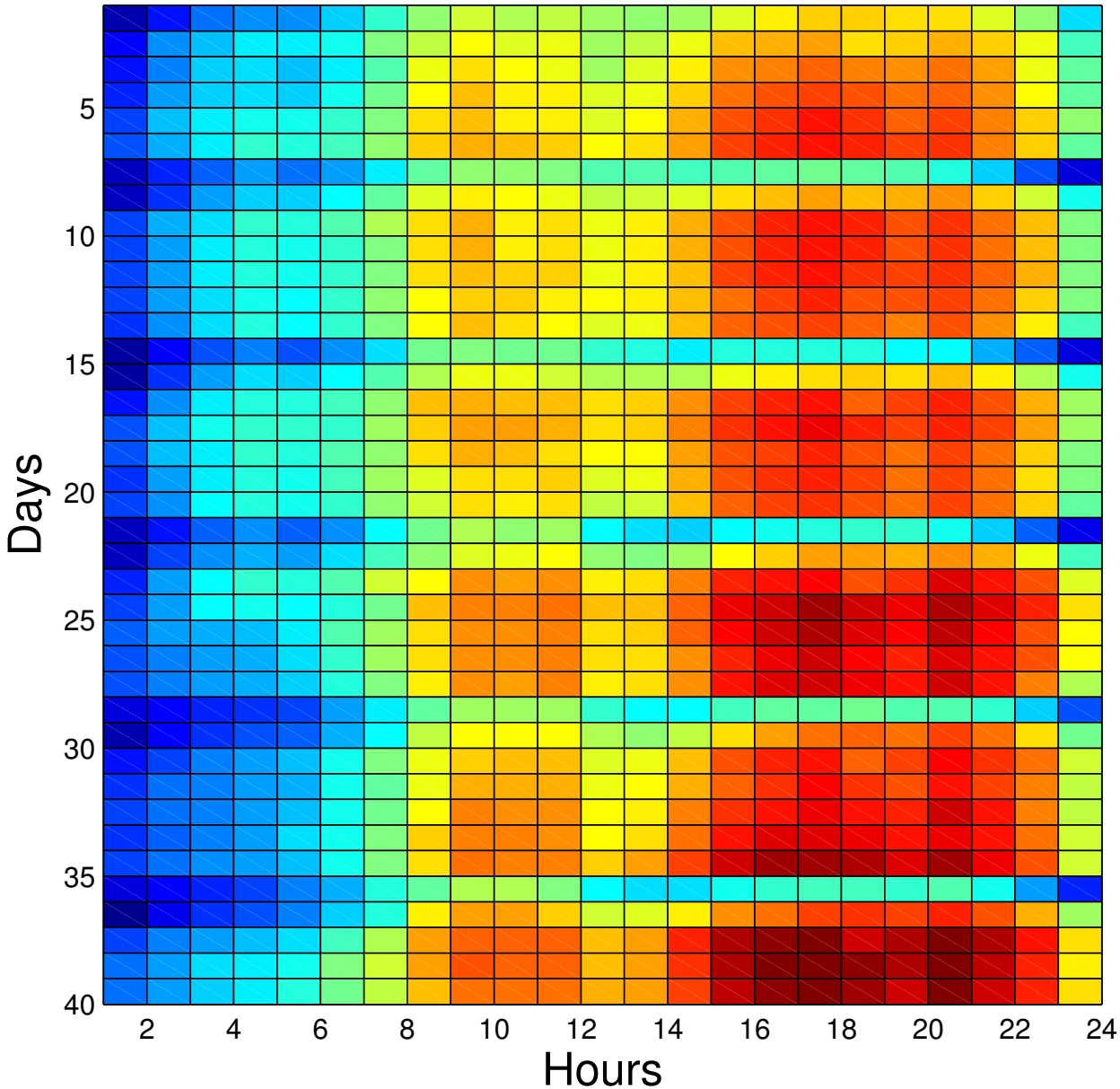
- one year seasons (temperature, daytime),
- one week,
- one day (working day, week-end),
- a holiday,
- aperiodic events.



Energy consumption one-week forecast for each hour



The autoregressive matrix, five week-ends



The autoregressive matrix and the linear model

$$\underset{(m+1)\times(n+1)}{\mathbf{X}^*} = \left(\begin{array}{c|ccc} s_T & s_{T-1} & \cdots & s_{T-\kappa+1} \\ \hline s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \cdots & s_{(m-2)\kappa+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n\kappa} & s_{n\kappa-1} & \cdots & s_{n(\kappa-1)+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{\kappa} & s_{\kappa-1} & \cdots & s_1 \end{array}\right).$$

In a nutshell,

$$\mathbf{X}^* = \left[\begin{array}{c|c} \underset{1\times 1}{s_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

In terms of linear regression:

$$\mathbf{y} = \mathbf{X}\mathbf{w},$$

$$y_{m+1} = s_T = \mathbf{w}^\intercal \mathbf{x}_{m+1}^\intercal.$$

Model generation

Introduce a set of the primitive functions $\mathfrak{G} = \{g_1, \ldots, g_r\}$, for example $g_1 = 1$, $g_2 = \sqrt{x}$, $g_3 = x$, $g_4 = x\sqrt{x}$, etc.

The generated set of features $\mathbf{X} =$

$$\left(\begin{array}{cccc|c|ccc}
g_1 \circ s_{T-1} & \ldots & g_r \circ s_{T-1} & \ldots & g_1 \circ s_{T-\kappa+1} & \ldots & g_r \circ s_{T-\kappa+1} \\
\hline
g_1 \circ s_{(m-1)\kappa-1} & \ldots & g_r \circ s_{(m-1)\kappa-1} & \ldots & g_1 \circ s_{(m-2)\kappa+1} & \ldots & g_r \circ s_{(m-2)\kappa+1} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
g_1 \circ s_{n\kappa-1} & \ldots & g_r \circ s_{n\kappa-1} & \ldots & g_1 \circ s_{n(\kappa-1)+1} & \ldots & g_r \circ s_{n(\kappa-1)+1} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
g_1 \circ s_{\kappa-1} & \ldots & g_r \circ s_{\kappa-1} & \ldots & g_1 \circ s_1 & \ldots & g_r \circ s_1
\end{array}\right).$$

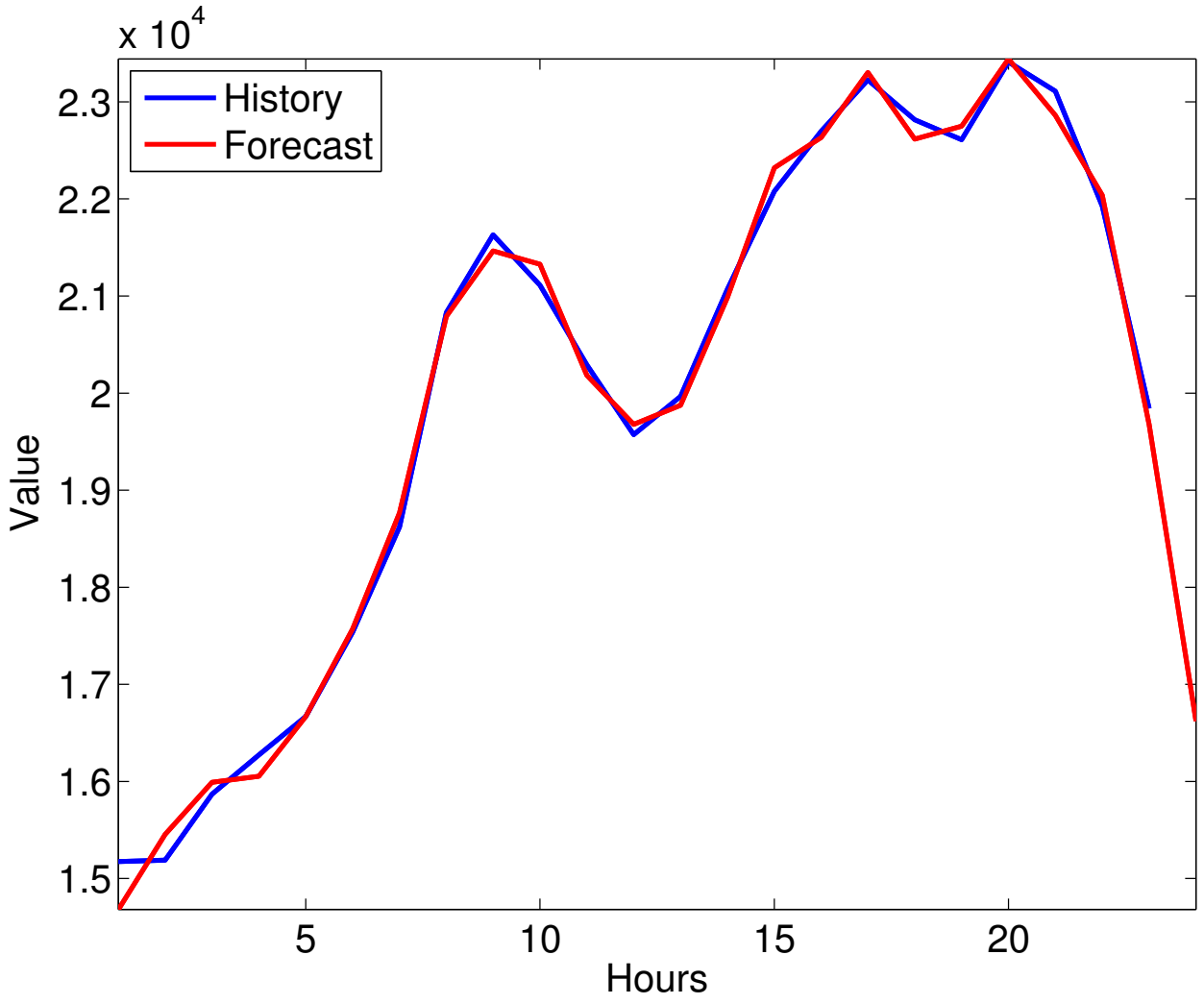
Kolmogorov-Gabor polynomial as a variant for model generation

$$y = w_0 + \sum_{i=1}^{UV} w_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_i x_j + \cdots + \sum_{i=1}^{n} \ldots \sum_{z=1}^{n} w_{i\ldots z} x_i \ldots x_z,$$

where the coefficients

$$\mathbf{w} = \left(w_0, w_i, w_{ij}, \ldots, w_{i\ldots z}\right)_{i,j,\ldots,z=1,\ldots,n}.$$

The one-day forecast (an example)



The function $y = f(\mathbf{x}, \mathbf{w})$ could be a linear model, neural network, deep NN, SVN, ...

Ill-conditioned matrix, or curse of dimensionality

Assume we have hourly data on price/consumption for three years. Then the matrix $\underset{(m+1)\times(n+1)}{\mathbf{X}^*}$ is

$$156 \times 168, \text{ in details: } 52\text{w} \cdot 3\text{y} \times 24\text{h} \cdot 7\text{d};$$

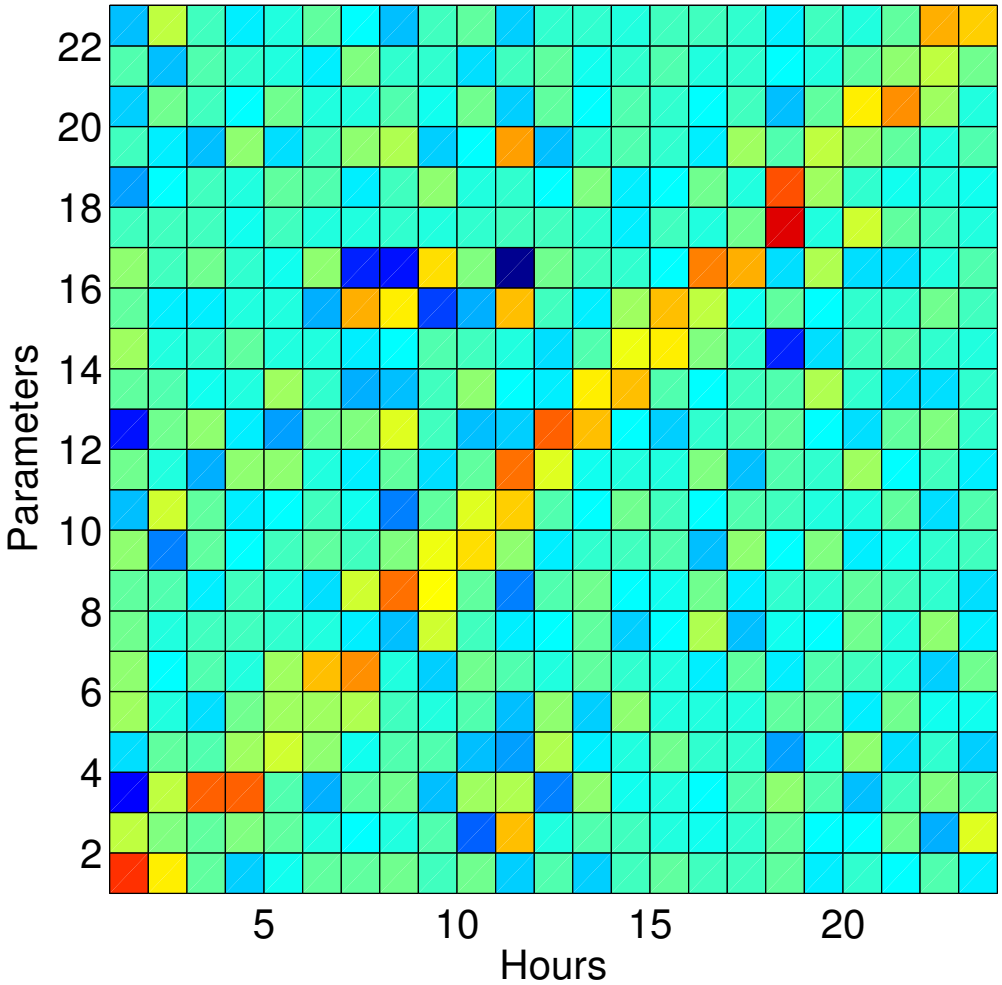
- for 6 time series the matrix $\mathbf{X}$ is 156×1008,
- for 4 primitive functions it is 156×4032,

$$m << n.$$

The autoregressive matrix could be considered as *ill-conditioned* and *multi-correlated*. The model selection procedure is required.

How many parameters must be used to forecast?

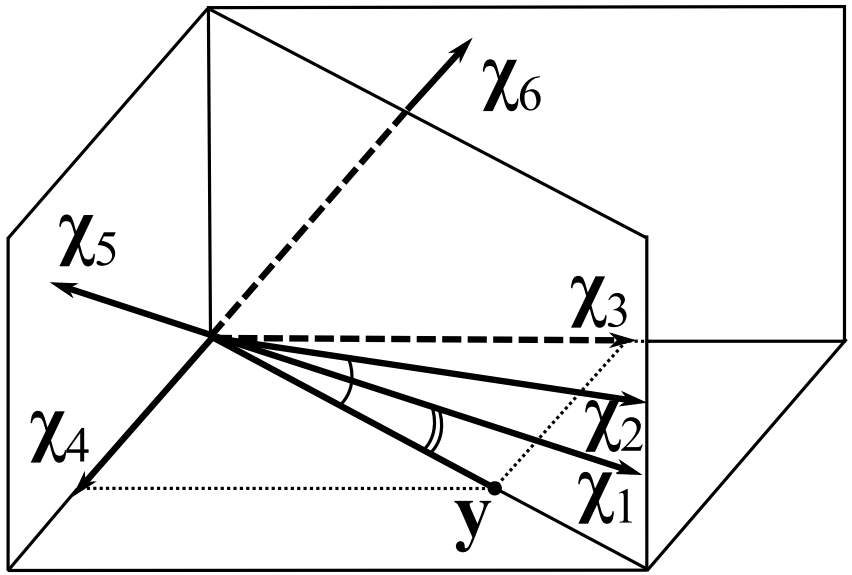
The color shows the value of a parameter for each hour.



Estimate parameters $\mathbf{w}(\tau) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$, then calculate the sample $s(\tau) = \mathbf{w}^\mathsf{T}(\tau)\mathbf{x}_{m+1}$ for each τ of the next $(m+1\text{-th})$ period.

Selection of a stable set of features of restricted size

The sample contains multicollinear χ_1, χ_2 and noisy χ_5, χ_6 features, columns of the design matrix $\mathbf{X}$. We want to select two features from six.



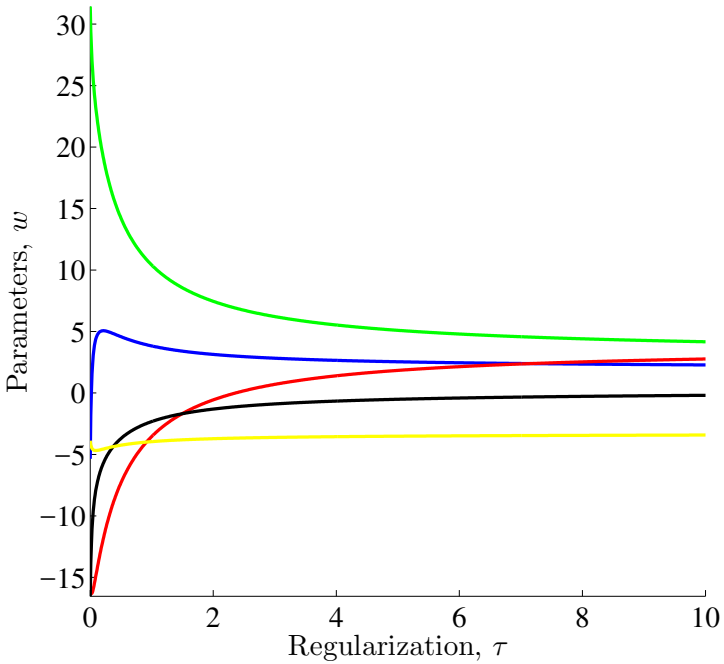
Stability and accuracy for a fixed complexity

The solution: $\chi_3, \chi_4$is an orthogonal set of features minimizing the error function.

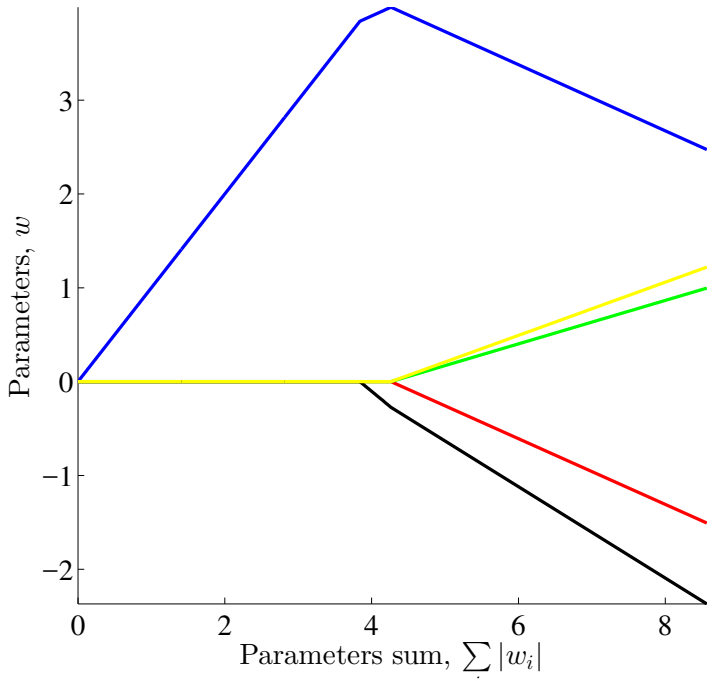
Algorithms: Add/Del, GMDH, Stepwise, Ridge, Lasso, Stagewise, FOS, LARS, Genetics, ...

Model parameter values with regularization

Vector-function $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w}, \mathbf{x}_m)]^\mathsf{T} \in \mathbb{Y}^m$.



$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2 + \gamma^2 \|\mathbf{w}\|^2$$



$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2,$$
$$T(\mathbf{w}) \leqslant \tau$$

Empirical distribution of model parameters

There given a sample $\{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$ of realizations of the m.r.v. $\mathbf{w}$ and an error function $S(\mathbf{w}|\mathfrak{D}, \mathbf{f})$. Consider the set of points $\{s_k = \exp\big(-S(\mathbf{w}_k|\mathfrak{D}, \mathbf{f})\big) | k = 1, \ldots, K\}$.

x- and y-axis: parameters $\mathbf{w}$, z-axis: $\exp\big(-S(\mathbf{w})\big)$.

Generating extra features

To augment feature description, consider the following types of features:

1) the local history of all time series themselves,
2) transformations (non-parametric and parametric) of local history,
3) parameters of the local models,
4) distances to the centroids of local clusters.

Multiscale data

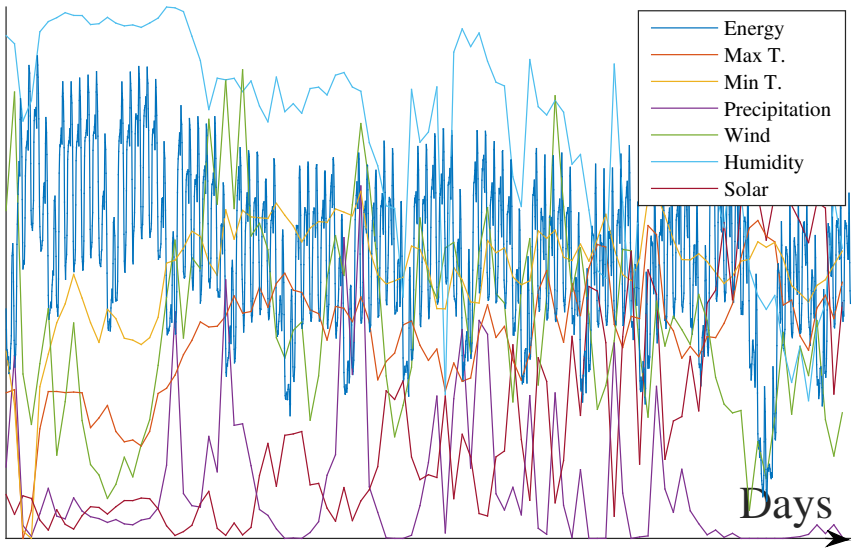
Consider a large set of time series $\mathfrak{D} = \{\mathbf{s}^{(q)} | \; q = 1 \ldots, Q\}$.

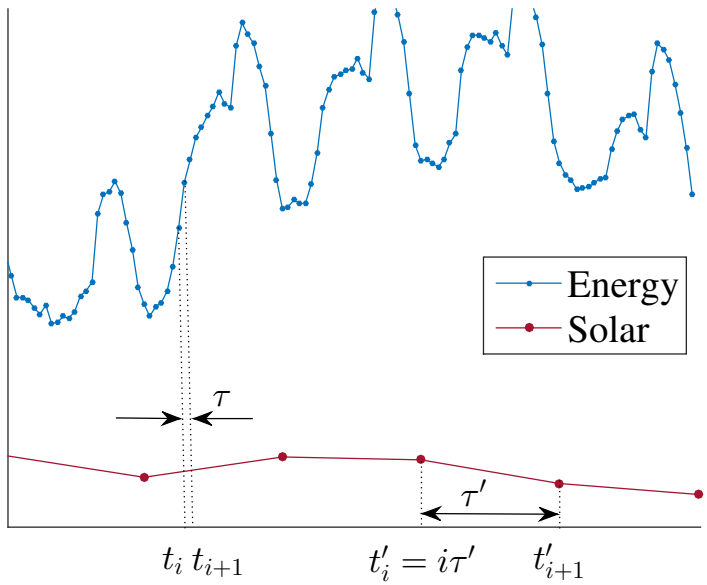
Each real-valued time series $\mathbf{s}$

$$\mathbf{s} = [s_1, \ldots, s_i, \ldots, s_T], \quad s_i = s(t_i), \quad 0 \leq t_i \leq t_{\max}$$

is a sequence of observations of some real-valued signal $s(t)$.

Each time series $\mathbf{s}^{(q)}$ has its own sampling rate $\tau^{(q)}$.





Examples of nonparametric transformation functions

- Univariate

Formula	Output dimension
$\sqrt{x}$	1
$x\sqrt{x}$	1
$\arctan x$	1
$\ln x$	1
$x \ln x$	1

- Bivariate

Plus	$x_1 + x_2$
Minus	$x_1 - x_2$
Product	$x_1 \cdot x_2$
Division	$\frac{x_1}{x_2}$
	$x_1\sqrt{x_2}$
	$x_1 \ln x_2$

Nonparametric aggregation: sample statistics

Nonparametric transformations include basic data statistics:

- Sum or average value of each row $\mathbf{x}_i$, $i = 1, \ldots, m$:

$$\phi_i = \sum_{j=1}^{n} x_{ij}, \text{ or } \phi_i' = \frac{1}{n} \sum_{j=1}^{n} x_{ij}.$$

- Min and max values: $\phi_i = \min_j x_{ij}$, $\phi_i' = \max_j x_{ij}$.
- Standard deviation:

$$\phi_i = \frac{1}{n-1} \sqrt{\sum_{j=1}^{n} (x_{ij} - \text{mean}(\mathbf{x}_i))^2}.$$

- Data quantiles: $\phi_i = [X_1, \ldots, X_K]$, where

$$\sum_{j=1}^{n} [X_{k-1} < x_{ij} \leq X_k] = \frac{1}{K}, \text{ for } k = 1, \ldots, K.$$

Nonparametric transformations: Haar's transform

Applying Haar's transform produces multiscale representations of the same data.

Assume that $n = 2^K$ and init $\phi_{i,j}^{(0)} = \phi_{i,j}'^{(0)} = x_{ij}$ for $j = 1, \ldots, n$.

To obtain coarse-graining and fine-graining of the input feature vector $\mathbf{x}_i$, for $k = 1, \ldots, K$ repeat:

- data averaging step

$$\phi_{i,j}^{(k)} = \frac{\phi_{i,2j-1}^{(k-1)} + \phi_{i,2j}^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k},$$

- and data differencing step

$$\phi_{i,j}'^{(k)} = \frac{\phi_{i,2j}'^{(k-1)} - \phi_{i,2j-1}'^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k}.$$

The resulting multiscale feature vectors are $\phi_i = [\phi_i^{(1)}, \ldots, \phi_i^{(K)}]$ and $\phi_i' = [\phi_i'^{(1)}, \ldots, \phi_i'^{(K)}]$.

Examples of parametric transformation functions

Function name	Formula	Output dim.	Num. of args	Num. of pars
Add constant	$x + w$	1	1	1
Quadratic	$w_2 x^2 + w_1 x + w_0$	1	1	3
Cubic	$w_3 x^3 + w_2 x^2 + w_1 x + w_0$	1	1	4
Logarithmic sigmoid	$1/(w_0 + \exp(-w_1 x))$	1	1	2
Exponent	$\exp x$	1	1	0
Normal	$\frac{1}{w_1\sqrt{2\pi}} \exp\left(\frac{(x-w_2)^2}{2w_1^2}\right)$	1	1	2
Multiply by constant	$x \cdot w$	1	1	1
Monomial	$w_1 x^{w_2}$	1	1	2
Weibull-2	$w_1 w_2 x^{w_2-1} \exp -w_1 x^{w_2}$	1	1	2
Weibull-3	$w_1 w_2 x^{w_2-1} \exp -w_1 (x - w_3)^{w_2}$	1	1	3
...	...	...	...	...

Monotone functions

- **By grow rate**

Function name	Formula	Constraints
Linear	$w_1 x + w_0$	
Exponential rate	$\exp(w_1 x + w_0)$	$w_1 > 0$
Polynomial rate	$\exp(w_1 \ln x + w_0)$	$w_1 > 1$
Sublinear polynomial rate	$\exp(w_1 \ln x + w_0)$	$0 < w_1 < 1$
Logarithmic rate	$w_1 \ln x + w_0$	$w_1 > 0$
Slow convergence	$w_0 + w_1/x$	$w_1 \neq 0$
Fast convergence	$w_0 + w_1 \cdot \exp(-x)$	$w_1 \neq 0$

- **Other**

Soft ReLu	$\ln(1 + e^x)$	
Sigmoid	$1/(w_0 + \exp(-w_1 x))$	$w_1 > 0$
Softmax	$1/(1 + \exp(-x))$	
Hiberbolic tangent	$\tanh(x)$	
softsign	$\frac{\lvert x \rvert}{1+\lvert x \rvert}$	

Parametric transformations

Optimization of the transformation function parameters $\mathbf{b}$ is iterative:

1. Fix the vector $\hat{\mathbf{b}}$, collected over all the primitive functions $\{g\}$, which generate features ϕ:

$$\hat{\mathbf{w}} = \arg\min S\big(\mathbf{w}|\mathbf{f}(\mathbf{w},\mathbf{x}),\mathbf{y}\big), \quad \text{where} \quad \phi(\hat{\mathbf{b}},\mathbf{s}) \subseteq \mathbf{x}.$$

2. Optimize transformation parameters $\hat{\mathbf{b}}$ given model parameters $\hat{\mathbf{w}}$

$$\hat{\mathbf{b}} = \arg\min S\big(\mathbf{b}|\mathbf{f}(\hat{\mathbf{w}},\mathbf{x}),\mathbf{y}\big).$$

Repeat these steps until vectors $\hat{\mathbf{w}}, \hat{\mathbf{b}}$ converge.

Parameters of the local models

More feature generation options:

- Parameters of SSA approximation of the time series $\mathbf{x}^{(q)}$.
- Parameters of the FFT of each $\mathbf{x}^{(q)}$.
- Parameters of polynomial/spline approximation of each $\mathbf{x}^{(q)}$.

Parameters of the local models: SSA

For the time series $\mathbf{s}$ construct the Hankel matrix with a period k and shift p, so that for $\mathbf{s} = [s_1, \ldots, s_T]$ the matrix

$$\mathbf{H}^* = \left[\begin{array}{c|cc} s_T & \ldots & s_{T-k+1} \\ \vdots & \ddots & \vdots \\ s_{k+p} & \ldots & s_{1+p} \\ s_k & \ldots & s_1 \end{array}\right], \text{ where } 1 \geqslant p \geqslant k.$$

Reconstruct the regression to the first column of the matrix $\mathbf{H}^* = [\mathbf{h}, \mathbf{H}]$ and denote its least square parameters as the feature vector

$$\phi(\mathbf{s}) = \arg\min \|\mathbf{h} - \mathbf{H}\phi\|_2^2.$$

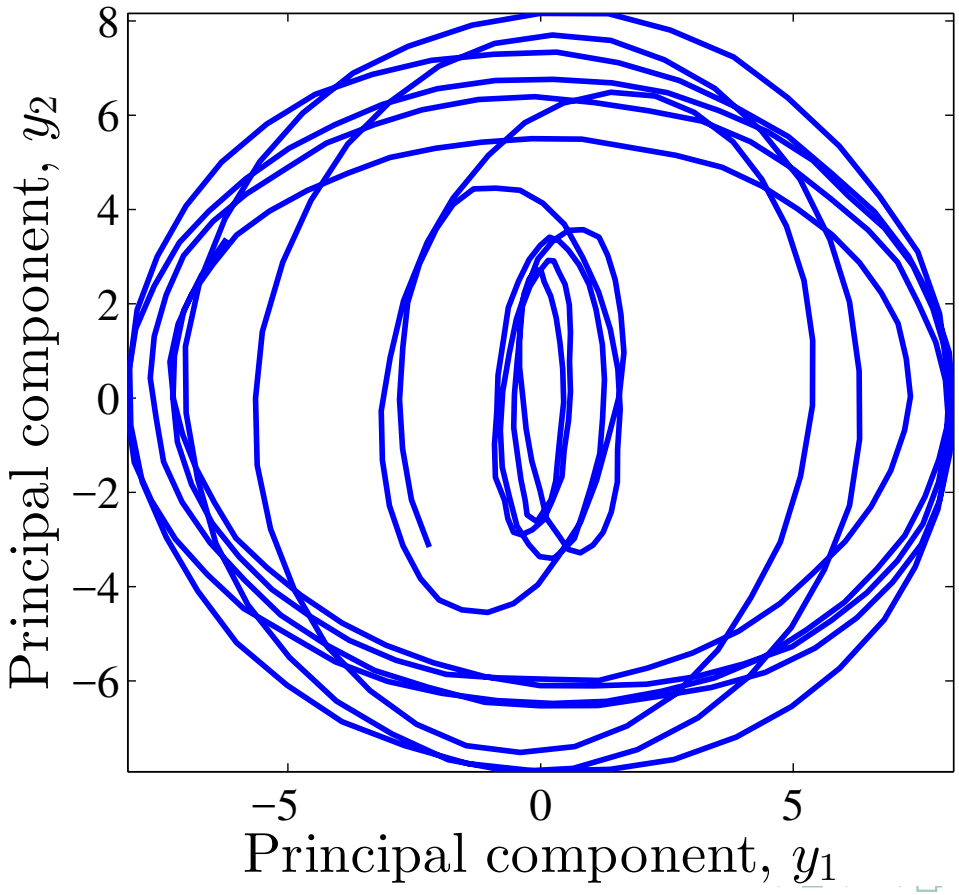
For the orignal feature vector $\mathbf{x} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(Q)}]$ use the parameters $\phi(\mathbf{x}^{(q)})$, $q = 1, \ldots, Q$ as the features.

SSA: principal components

Compute the SVD of covariance matrix of $\mathbf{H}$

$$\frac{1}{N}\mathbf{H}^{\mathsf{T}}\mathbf{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$$

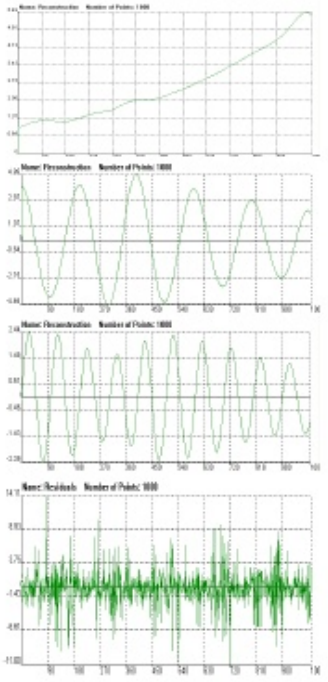
and find the principal components $\mathbf{y}_j = \mathbf{H}\mathbf{v}_j$.



SSA: decomposition

$$\mathbf{H}^* = \sum_j \mathbf{H}_j, \quad \mathbf{H}_j = \mathbf{v}_j \mathbf{y}_j^{\mathsf{T}}.$$

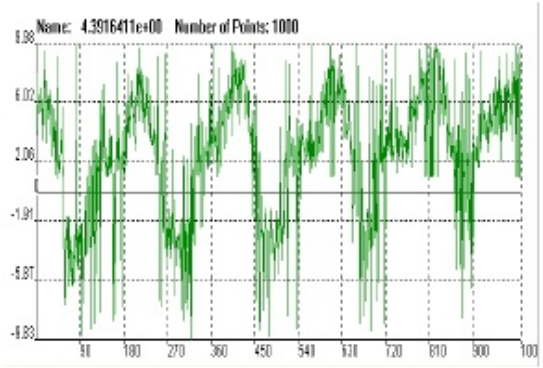
Singular Spectrum Analysis



Works directly in the time domain

addition

SSA



Component <=> % total information

Ref: Broomhead & King (1986)

Metric features: distances to the centroids of local clusters

Apply kernel trick to the time series.

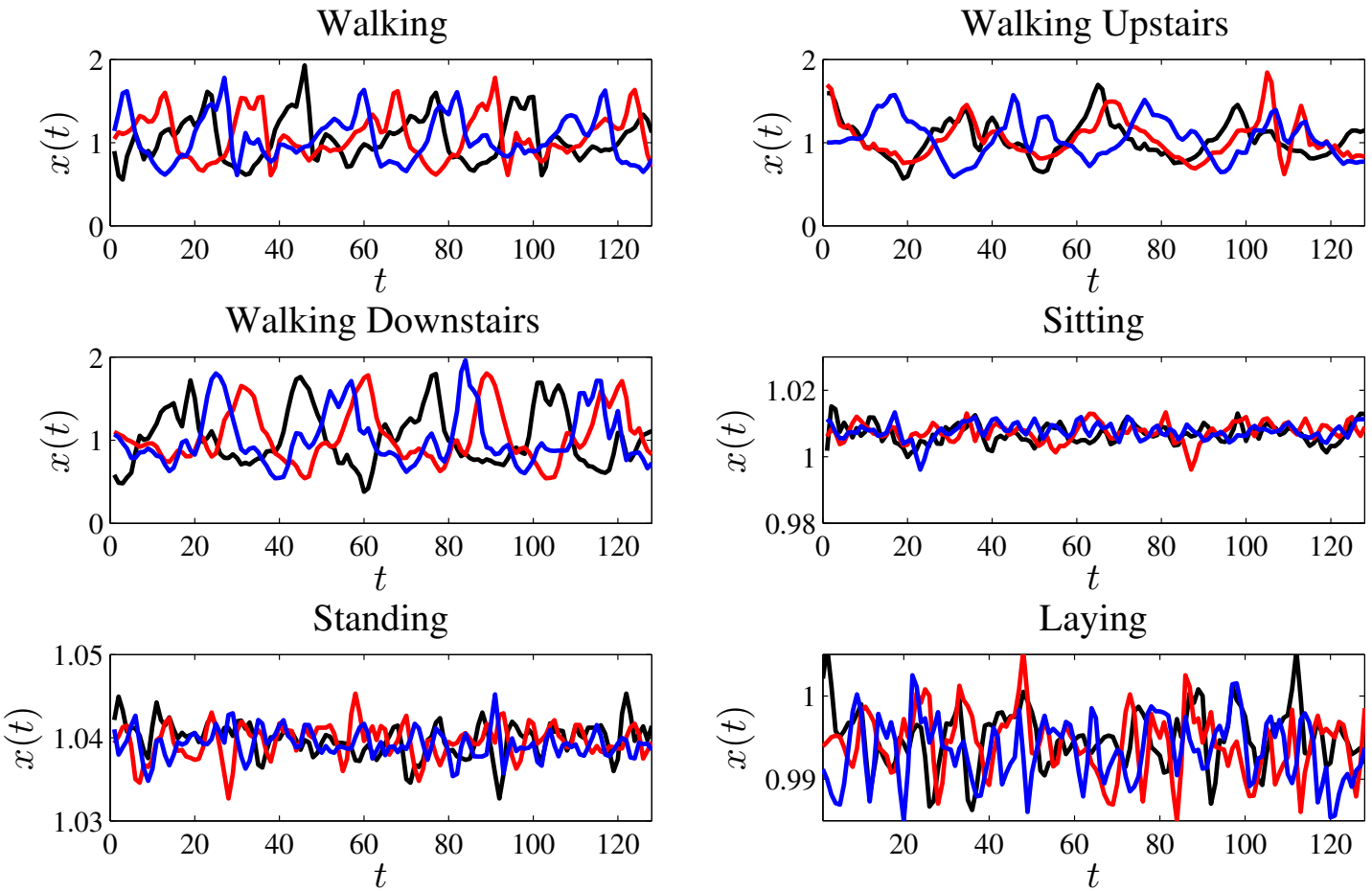
1. For given local feature vector $\mathbf{x}_i^{(q)}$, $q = 1, \ldots, Q$ compute k-means centroids $\mathbf{c}_p^{(m)}$, $p = 1, \ldots, P$.
2. With the selected k-means distance function ρ construct the feature vector

$$\phi_i^{(q)} = [\rho(\mathbf{c}_1^{(q)}, \mathbf{x}_i^{(q)}), \ldots, \rho(\mathbf{c}_P^{(q)}, \mathbf{x}_i^{(q)})] \in \mathbb{R}_+^P.$$

The procedure may be applied to each $\mathbf{x}^{(q)}$ or directly to the $\mathbf{x} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(Q)}]$, resulting in only P additional features instead of $Q \cdot P$

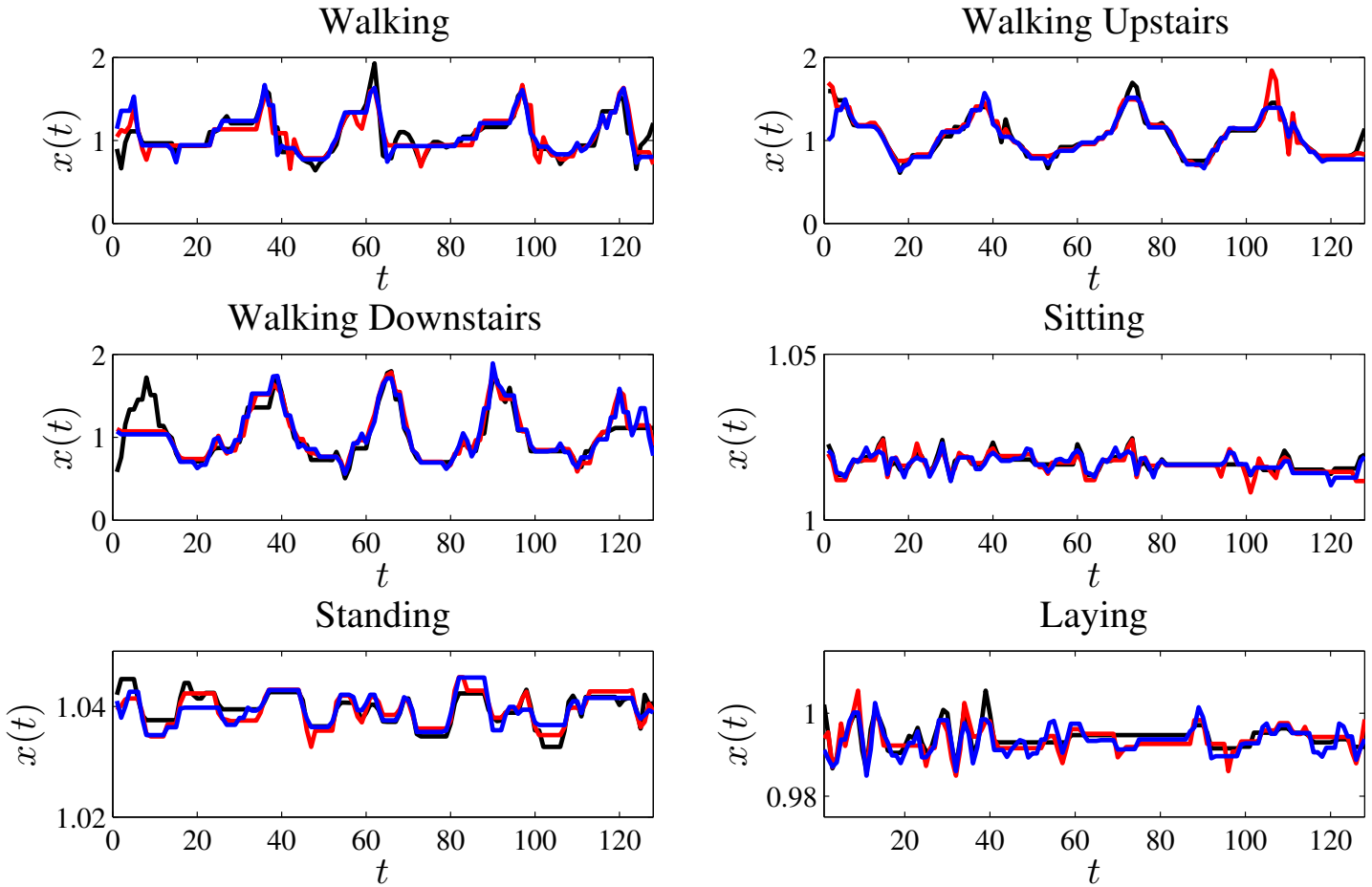
Computing distances to centroids of the time series

Sets of time series.



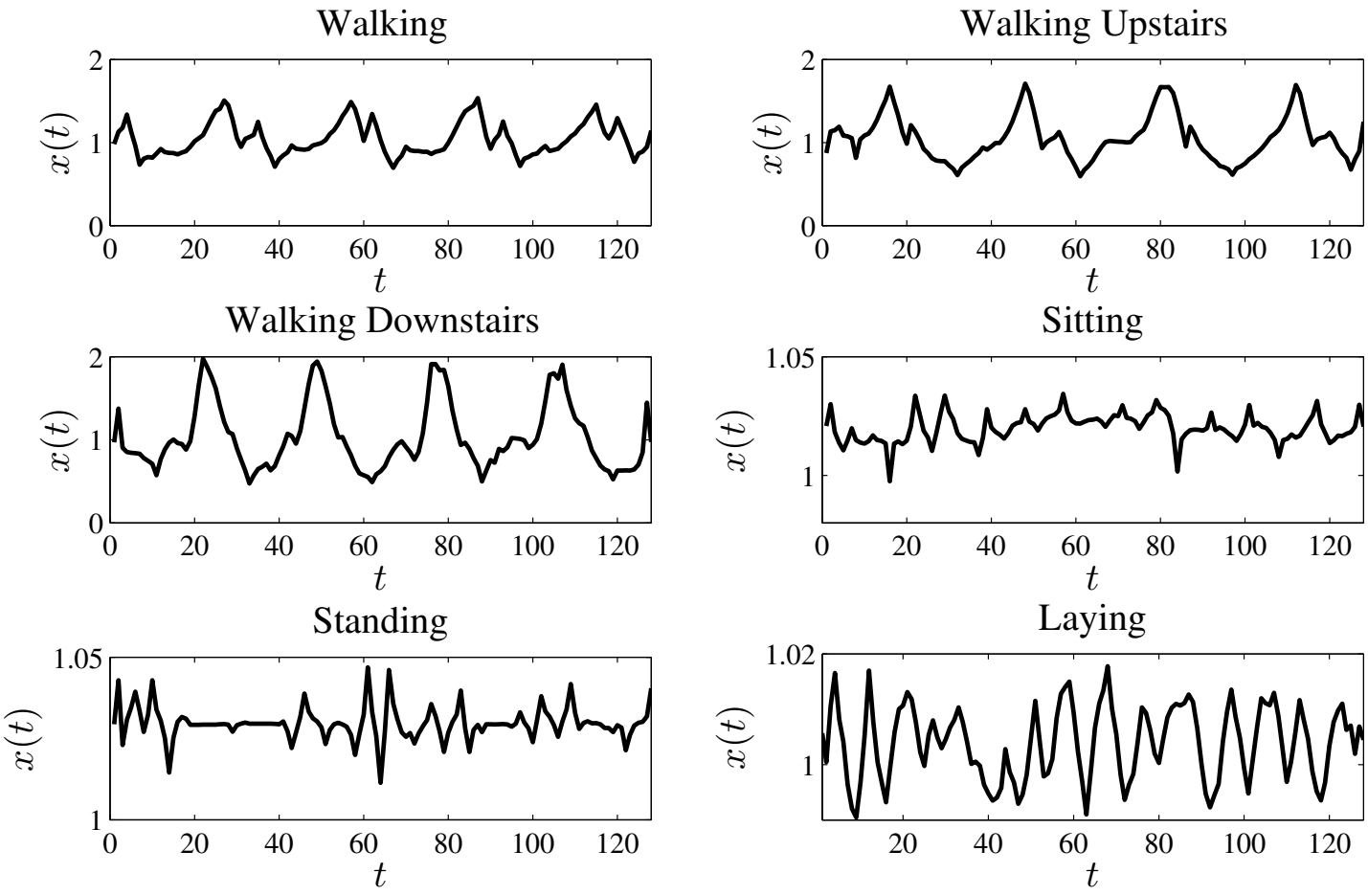
Computing distances to centroids of the time series

Aligned time series.



Computing distances to centroids of the time series

Centroids of the time series.



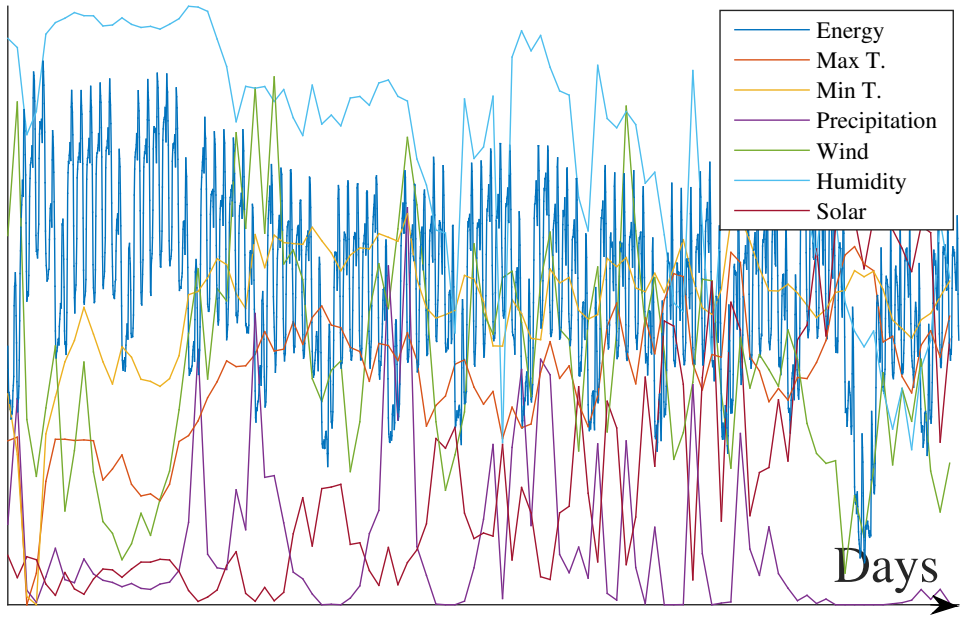
Data

1. Original Polish electricity load time series, 1999–2004, including:

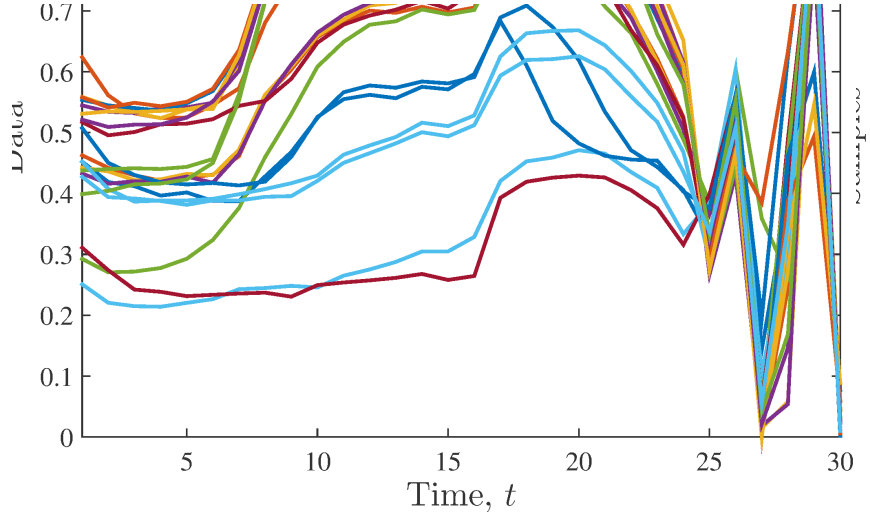
- hourly energy time series (total of 52512 observations),
- six daily weather time series from Warsaw (2188 observations): Max Temperature, Min Temperature, Precipitation, Wind, Relative Humidity, Solar.

2–5. Data sets with artificial inserted missing values, 1, 3, 5 and 10% missing.

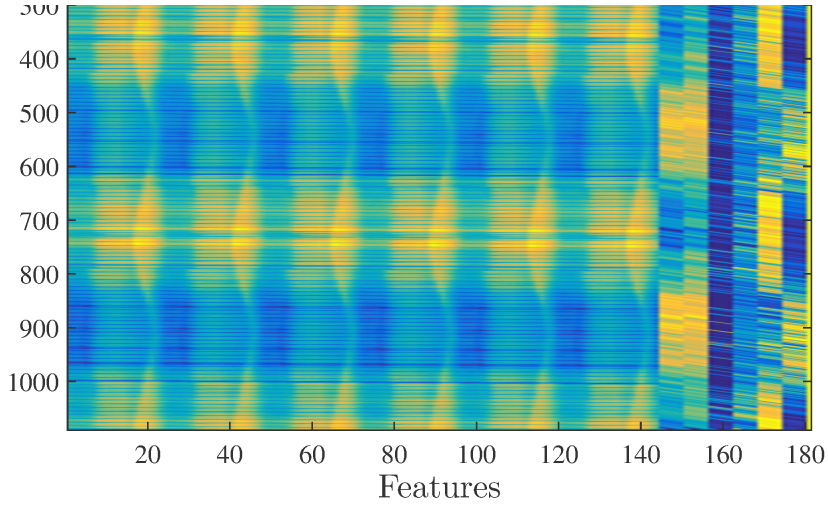
6. Data set with artificially varied sampling rate.



Data



Target variables.



The design matrix.

Models and features

Models:

- Baseline method: $\hat{s}_i = s_{i-1}$.
- Multivariate linear regression (MLR) with l_2-regularization. Regularization coefficient: 2
- SVR with multiple output. Kernel type: RBF, p_1: 2, p_2: 0, γ: 0.5, λ: 4.
- Feed-forward ANN with single hidden layer, size: 25
- Random forest (RF). Number of trees: 25 , number of variables for each decision split: 48.

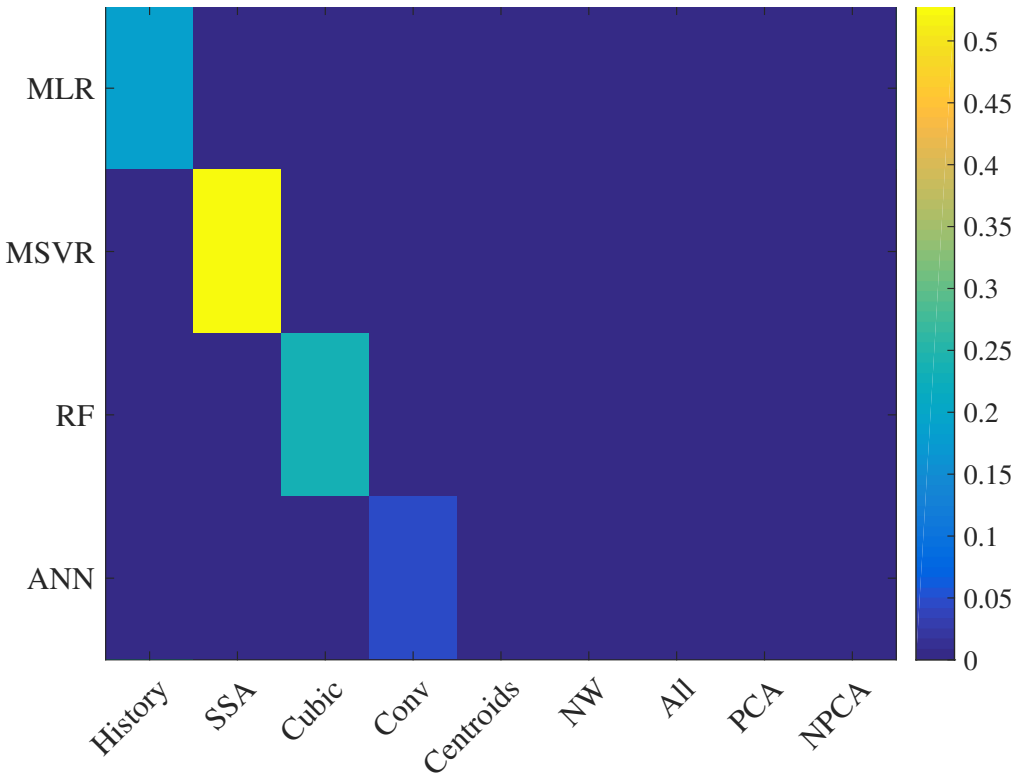
Feature combinations:

- History: the standard regression-based forecast with no additional features.
- SSA, Cubic, Conv, Centroids, NW: history + a particular feature.
- All: all of the above, with no feature selection.
- PCA and NPCA: all generation strategies with feature selection.

Forecasting errors, SMAPE

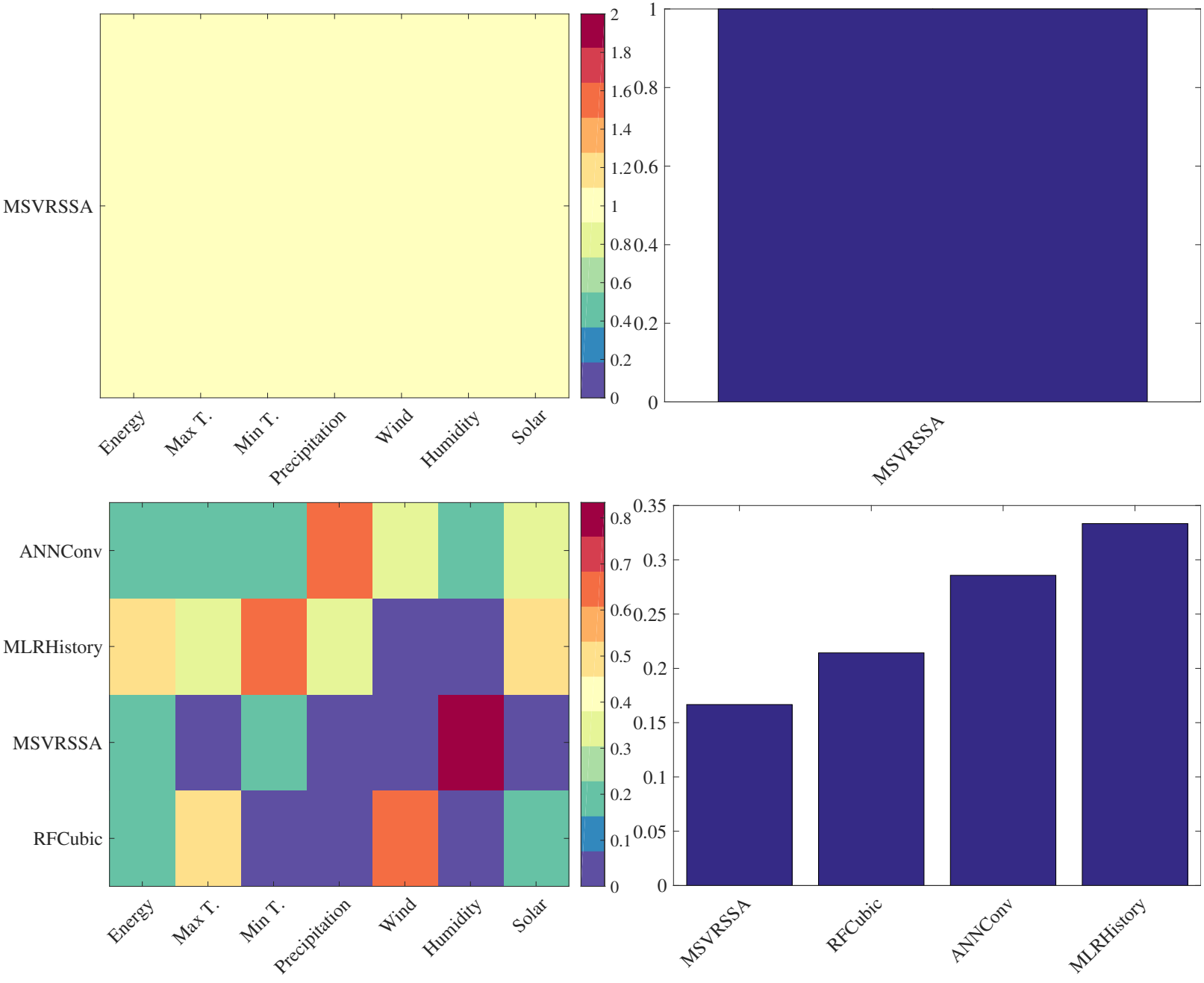
Data	Energy	Max T.	Min T.	Precip.	Wind	Humid.	Solar
Test							
orig	0.111	0.127	0.111	1.222	0.396	0.201	0.495
0.01	0.230	0.185	0.129	1.028	0.397	0.254	0.577
0.03	0.231	0.191	0.137	1.026	0.396	0.253	0.591
0.05	0.230	0.200	0.141	1.017	0.390	0.250	0.592
0.1	0.247	0.198	0.151	1.192	0.381	0.225	0.562
varying	0.124	0.139	0.102	1.232	0.395	0.219	0.489
Train							
orig	0.031	0.073	0.057	0.848	0.111	0.051	0.267
0.01	0.034	0.055	0.040	0.595	0.111	0.055	0.253
0.03	0.034	0.057	0.042	0.595	0.110	0.055	0.249
0.05	0.034	0.060	0.043	0.592	0.109	0.054	0.246
0.1	0.031	0.081	0.063	0.743	0.102	0.051	0.272
varying	0.027	0.057	0.044	0.888	0.112	0.055	0.272

Feature analysis

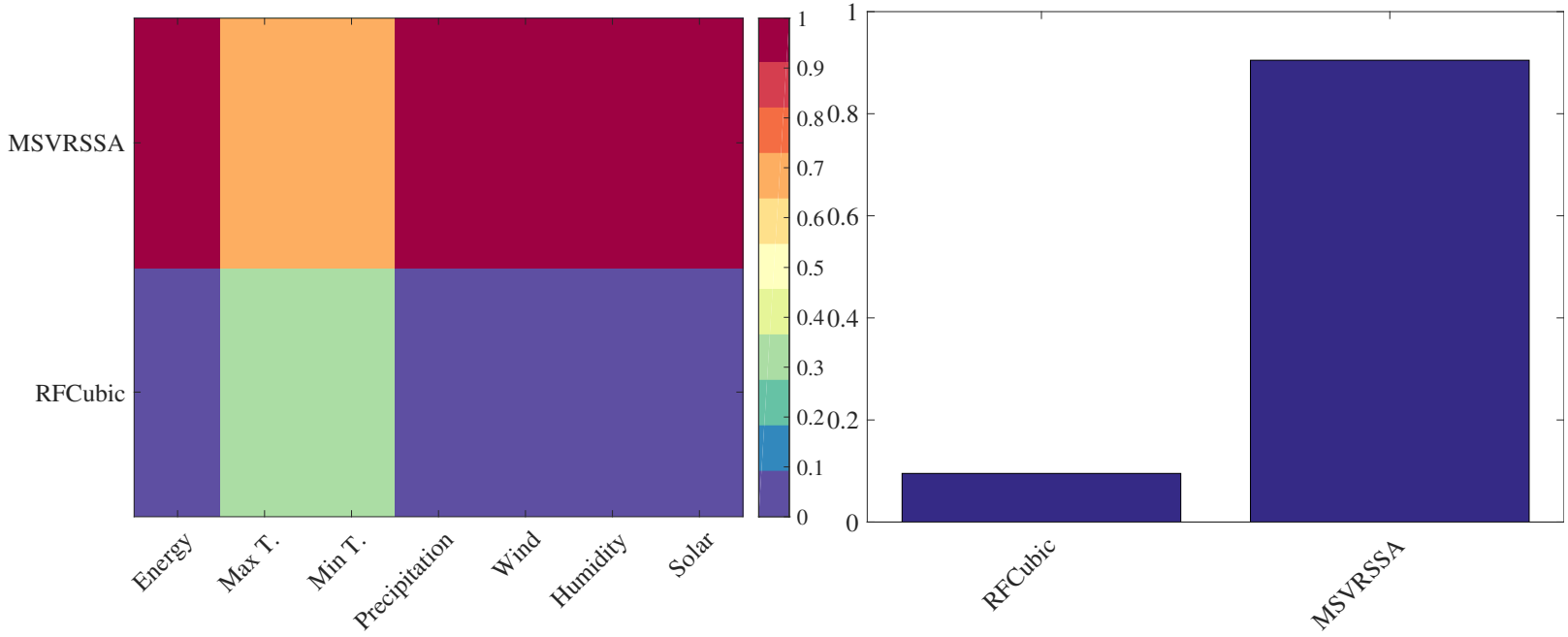


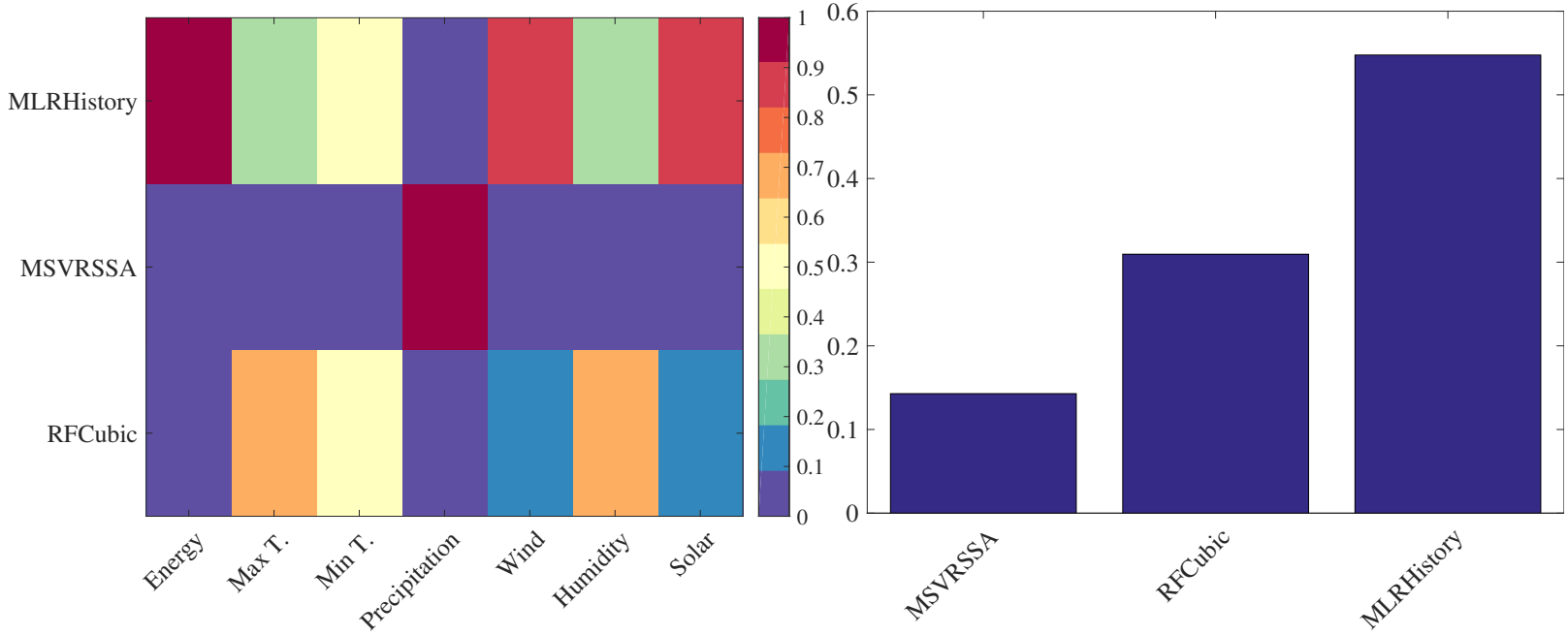
Ratio of times each combination of model and feature performed best for at least one of the time series (7) or error functions (6), all (6) data sets ($6 \times 7 \times 6 = 252$ cases).

Best models, train/test residues

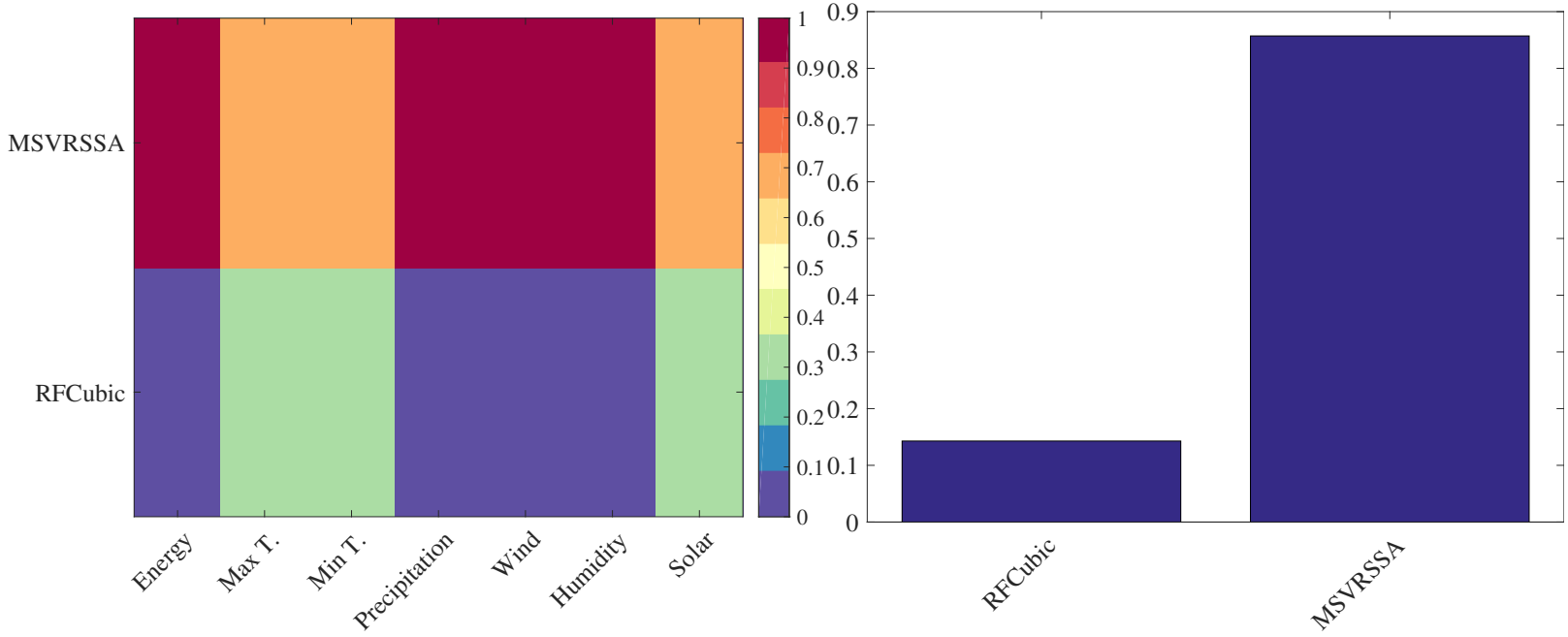


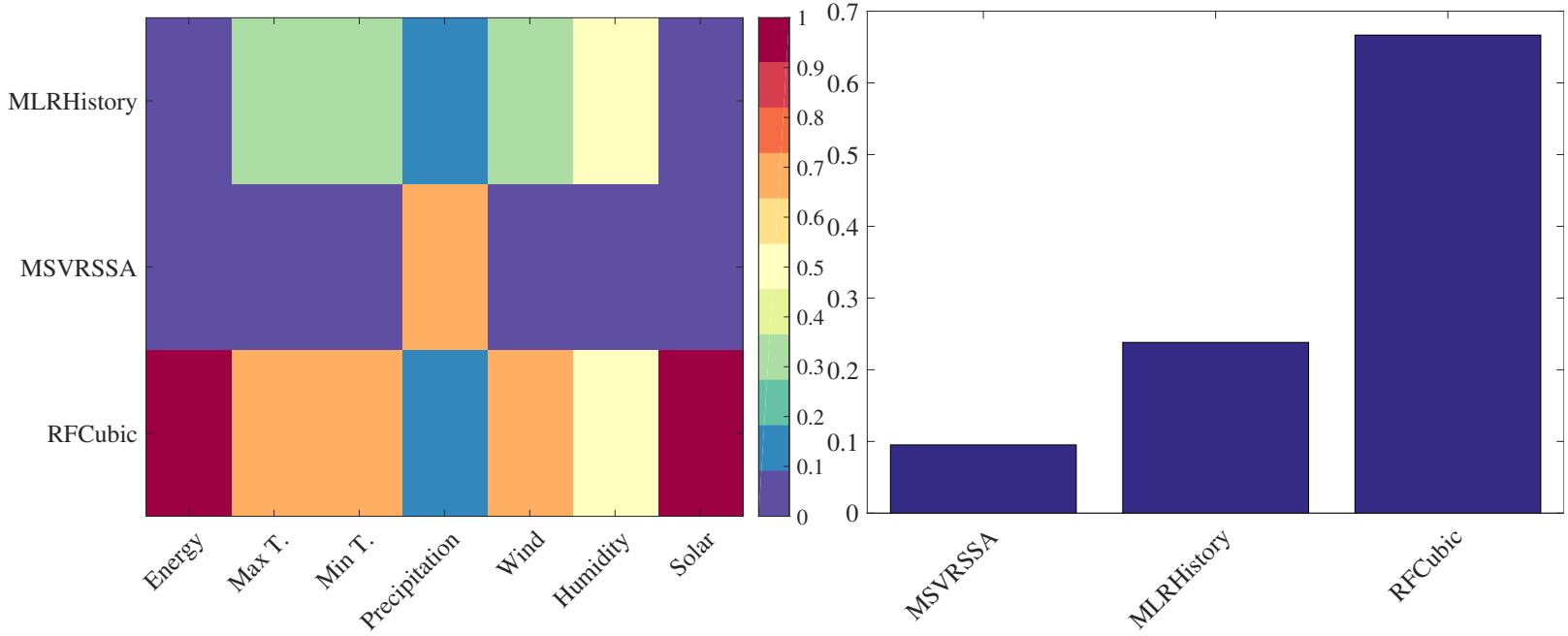
Best models, standard deviations (train/test residues)





Best models, train/test SMAPE





Resume: consequent model generation

Resume: model generation and selection cheat-sheet

A non-exhaustive list

What are hypothesis on data set?

- Set prior and posterior distribution hypothesis and construct internal criterion
- Assume one model or a mixture
- Assume outliers, class imbalances

How we generate models?

- Set a universal model
- Use primitive functions and rules of generation
- Forecast a model

How we select an optimal model?

- Use feature selection algorithms
- Use hyper-parameter analysis
- Run exhaustive search or genetic algorithm

How we check the model has the optimal structure?

- Use external criterions: AUC, BIC, Cp, Complexity, Stability