Time Series Clustering Methods

With Applications In Environmental Studies

Xiang Zhu

The University of Chicago

xiangzhu@uchicago.edu
https://sites.google.com/site/herbertxiangzhu

September 5, 2013
Argonne National Laboratory

Overview

1. Short survey of time series clustering
2. High-dimensional time series clustering via factor modelling
   - Factor modelling for high-dimensional time series
   - A new clustering approach based on factor modelling
   - Example: channel selection in hyper-spectral imagery
3. Shapelet-based feature extraction for long time series
   - Motivating example: Indoor Environmental Time Series
   - Unsupervised-shapelets algorithm for time series clustering
   - Indoor environmental data revisit: shapelets extraction

Motivation

Example (Hyper-spectral Imagery)

Investigate the **spectral-temporal** relationship

- Number of wave bands: 100–1000 (900 in the 2008 dataset)
- Number of measurements: ≤ 10 (6 in the 2008 dataset)
- More complicated if multiple pixels included (spatial)

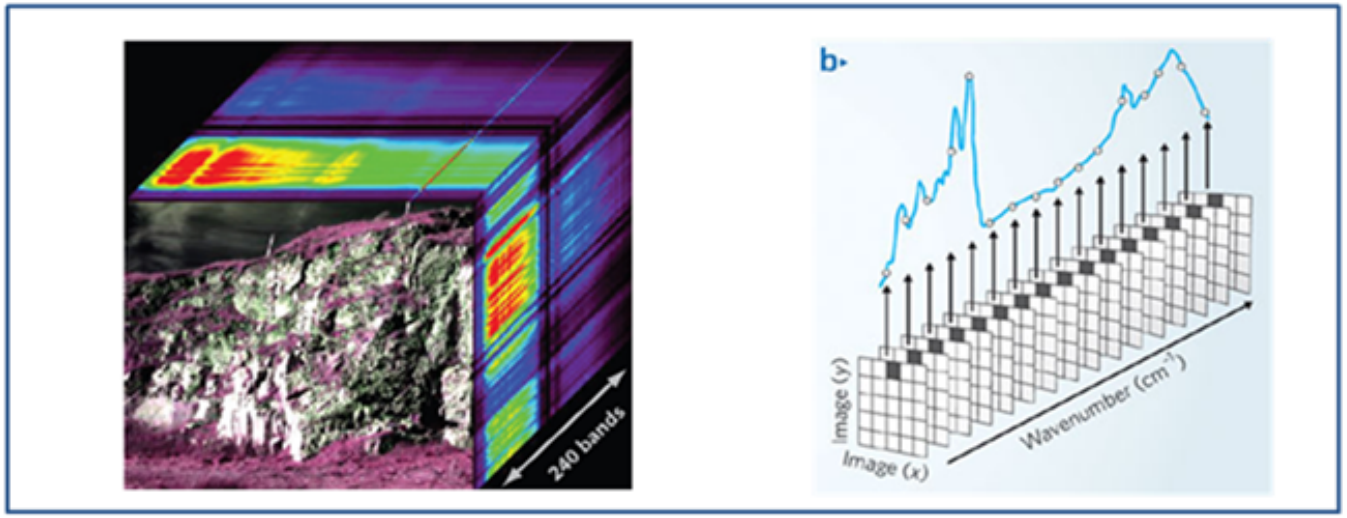


Figure: Hyper-spectral cube with 240 bands (left) and a conceptual diagram of hyper-spectral image layers and spectral reflectance plot from a single pixel (right)

Motivation

Choose a subset of channels to study.

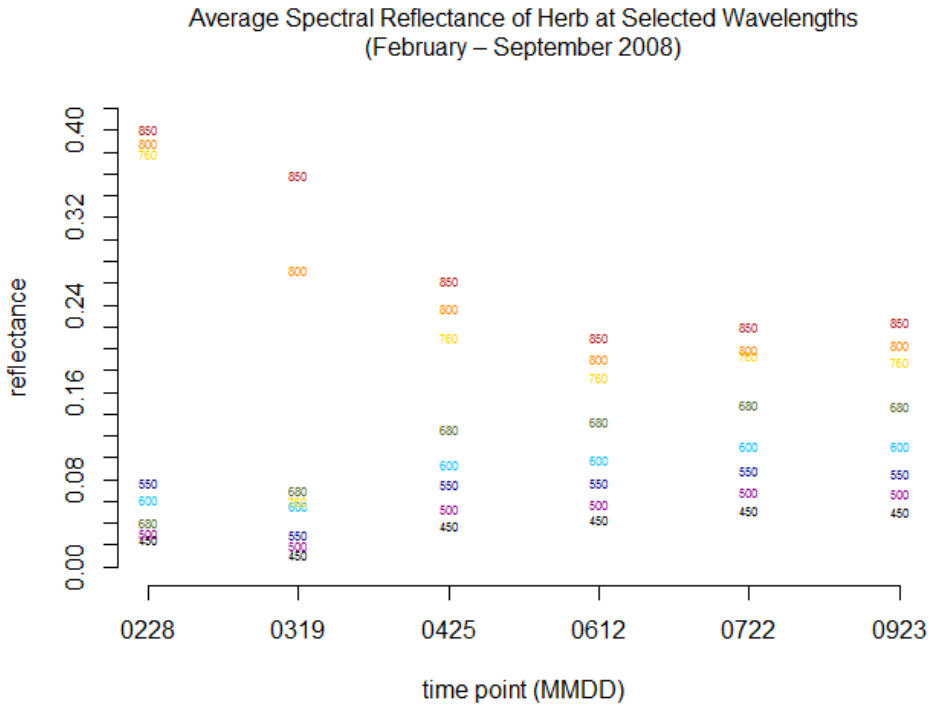


Figure: Eight wavelengths were selected via expert elicitation

Can we select the channels purely based on the data?

Motivation

Example (Environmental Sensor Networks)

To detect events based on different properties, various types of sensors are deployed on different occasions. How to decide when to turn on them?

- Empirical knowledge from professional technicians
- Hidden patterns discovered from historical dataset

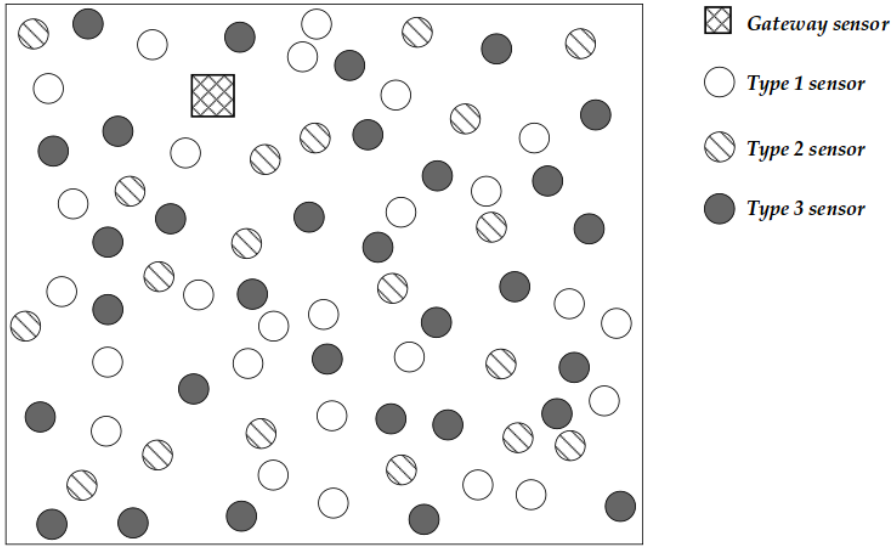


Figure: Composite Event Detection in Wireless Sensor Networks, *T. Vu et al. (2007)*

Time Series Clustering

Objective

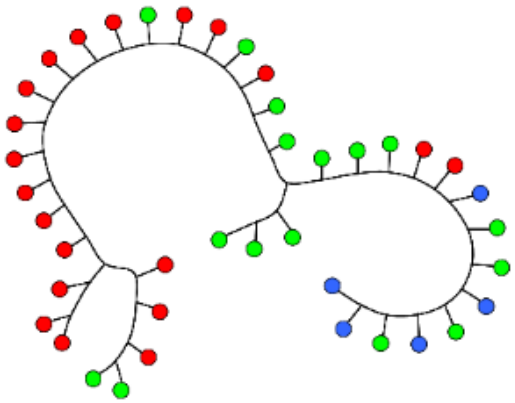
Clustering is an **unsupervised** learning method

- In general: To identify hidden structures and similarities
- For time series: To discover similar stochastic dynamics

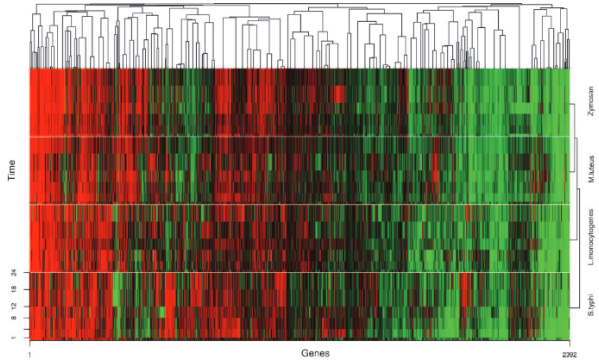
Applications

- Genetics
- Economics & Finance
- Remote Sensing
- Engineering
- Meteorology
- .....

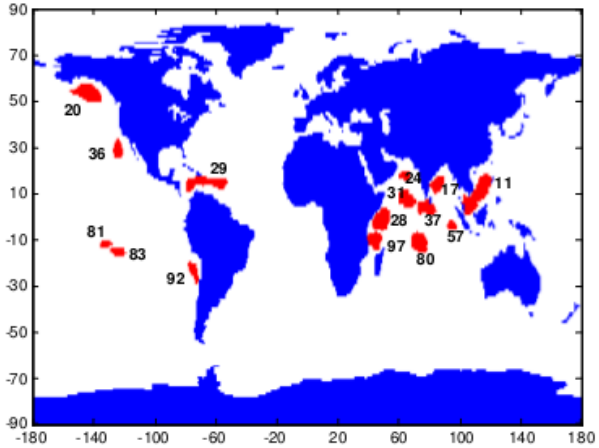
Time Series Clustering



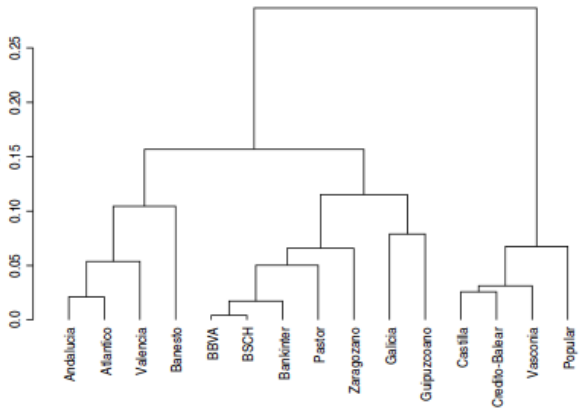
(a) Rock Categorization, *Cerra et al. (2011)*



(b) Gene Expression, *Heard et al. (2005)*



(c) Climate Index, *Steinbach et al. (2003)*



(d) Bank Evaluation, *Vilar et al. (2009)*

Time Series Clustering

Basics

- General-purpose clustering algorithms
- Evaluation criteria of the clustering performance
- Similarity measures between two time series

Approaches

- Raw-data-based
- Feature-based
- Model-based
- Prediction-based

Literatures

For an extensive review, refer *Liao, 2005*.
A very active area: many new works have been done.

High-dimensional Time Series

Example:
Daily smoothed sulfate levels for 98 US counties, 2002-2005, *Peng (2008)*.
Matrix Representation: $\mathbf{Y}_{N \times p}$

$$
\begin{array}{rcccc}
 & \mathbf{Var}_1 & \mathbf{Var}_2 & \cdots & \mathbf{Var}_p \\
\mathbf{Time}_1: & y_{11} & y_{12} & \cdots & y_{1p} \\
\mathbf{Time}_2: & y_{21} & y_{22} & \cdots & y_{2p} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{Time}_N: & y_{N1} & y_{N2} & \cdots & y_{Np}
\end{array}
$$

High-dimension: $p > N$; really common in environmental sciences.

High-dimensional Time Series

Major Difficulties

- Visualization
  - One variate: 2-D plot
  - Two variates: 3-D plot
  - More than three variates?
- Identification
  - Three equations
  - Ten unknown quantities
  - Can we solve them?
- Computation

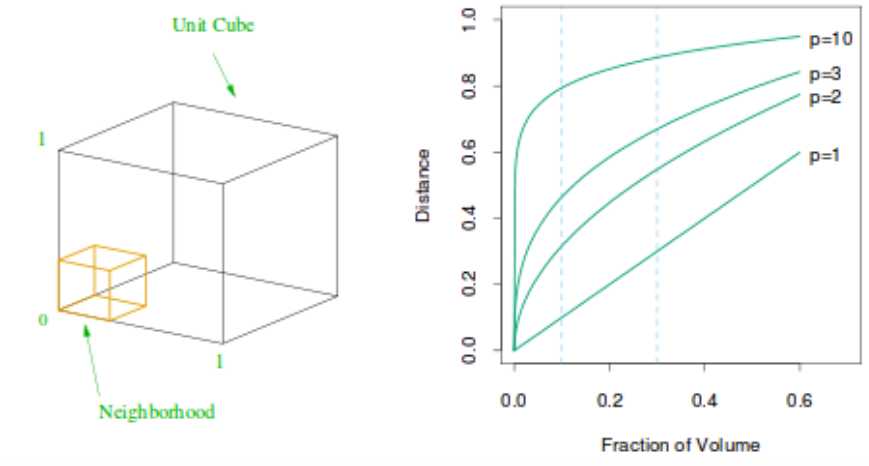


Figure: Curse of Dimensionality, *Hastie et al. (2009)*

High-dimensional Time Series

Reduce the dimensionality

- **Variable Selection**: select a subset of original attributes

$$\texttt{Micro-organism} = \texttt{Nitrogen} + \texttt{Moisture} + \texttt{Grazing} + \texttt{Wind} + \texttt{Temperature} + \texttt{Fertilizer} + \texttt{Solar} + \texttt{Policy}$$

- **Feature Extraction**: extract a set of new features from the original attributes via functioning mapping
  - Principal Component Analysis (the most renown one)
  - Wavelet Transformation (widely embraced by information scientists)
  - Factor analysis (originated in psychometrics; economist's favorite)

Factor Modelling for Time Series

Factor Model

Decompose $\mathbf{y}_t \rightsquigarrow$ latent factor process $\mathbf{x}_t \bigoplus$ white noise $\boldsymbol{\epsilon}_t$

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_t,$$

where $\mathbf{x}_t$: $r \times 1$ factor process with unknown $r \leq p$, $\mathbf{A}$: $p \times r$ unknown constant matrix.

Identification assumptions

1. No linear combinations of $\mathbf{x}_t$ are white noise.
2. The rank of $\mathbf{A}$ is r and its columns are orthonormal.

The keys for the inference of factor models

K1 determine the number of factors r

K2 estimate the $p \times r$ factor loading matrix $\mathbf{A}$

Factor Modelling for Time Series

Lam and Yao (2012)

Conducted simple eigen-analysis of a non-negative definite matrix.

Time lag information accumulator **M**

For prescribed integer $k_0 \geq 1$, define

$$\mathbf{M} = \sum_{k=1}^{k_0} \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y^{\mathsf{T}}(k), \ \mathbf{\Sigma}_y(k) = \mathrm{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t), \ k \in \mathbb{N}.$$

Why consider **M**?

Under mild **stationary** assumptions

- the number of non-zero eigenvalues of $\mathbf{M}$ is r (**K1**)
- the factor loading space $\mathcal{M}(\mathbf{A})$ is spanned by the eigenvectors of $\mathbf{M}$ corresponding to its non-zero eigenvalues (**K2**).

Factor Modelling for Time Series

Estimated time lag information accumulator

In practice, $\mathbf{M}$ can be estimated by

$$\hat{\mathbf{M}} = \sum_{k=1}^{k_0} \hat{\mathbf{\Sigma}}_y(k)\hat{\mathbf{\Sigma}}_y^{\mathsf{T}}(k)$$

where $\hat{\mathbf{\Sigma}}_y(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(\mathbf{y}_{t+k} - \bar{\mathbf{y}}) \cdot (\mathbf{y}_t - \bar{\mathbf{y}})^{\mathsf{T}}$, $\bar{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t$.

Tuning parameter k_0

- smaller values are favored
  - autocorrelation often at its strongest for small time lags
  - estimation of $\mathbf{\Sigma}_y(k)$ is less accurate when k is larger
- result is robust to the choice of k_0

Factor Modelling for Time Series

Estimation Procedures (*Lam and Yao, 2012*)

1. A ratio-based estimator $\hat{r}$ for the number of factors:

$$\hat{r} = \arg\min_{1 \le i \le R} \frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i}, \quad \textbf{(Intuition?)}$$

   where $\hat{\lambda}_1 \ge \ldots \ge \hat{\lambda}_p$ are the eigenvalues of $\hat{\mathbf{M}}$ and $r \le R \le p$ is a constant. (In practice take $R = p/2$.)

2. The columns of $\hat{\mathbf{A}}$ are the $\hat{r}$th orthonormal eigenvectors of $\hat{\mathbf{M}}$ corresponding to its $\hat{r}-$largest eigenvalues.

3. $\hat{\mathbf{x}}_t = \hat{\mathbf{A}}^{\mathsf{T}}\mathbf{y}_t$, $\hat{\epsilon}_t = (\mathbf{I}_p - \hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathsf{T}})\mathbf{y}_t$ and $\hat{\mathbf{y}}_t = \hat{\mathbf{A}}\hat{\mathbf{x}}_t$.

Clustering based on Factor Modelling

Zhu (2013) [my working paper at ANL this summer]

Basic Idea

Cluster the variates of multiple time series:
$\rightsquigarrow$ How similar they are in terms of temporal dynamics?

1. Dynamic structures extraction: $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_t$
2. Similarity measure: compare loadings in $\mathbf{A}$
3. Variate clustering: clustering rows of $\mathbf{A}$

Clustering based on Factor Modelling: 1.Temporal Dynamics Extraction

The loading matrix **A** conveys all useful information about stochastic dynamics of $\mathbf{Y} = \{\mathbf{y}_t\}_{t\in[N]}$.

$$\begin{pmatrix} y_{t1} \\ y_{t2} \\ \vdots \\ y_{tp} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pr} \end{pmatrix} \cdot \begin{pmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tp} \end{pmatrix} + \begin{pmatrix} \epsilon_{t1} \\ \epsilon_{t2} \\ \vdots \\ \epsilon_{tp} \end{pmatrix}.$$

Use *Lam and Yao (2012)*'s method to solve the factor modelling problem.

Clustering based on Factor Modelling
2. Measure of Similarity

Given two variates i and j, at time t,

$$y_{ti} = a_{i1}x_{t1} + a_{i2}x_{t2} + \ldots + a_{ir}x_{tr},$$
$$y_{tj} = a_{j1}x_{t1} + a_{j2}x_{t2} + \ldots + a_{jr}x_{tr}.$$

A metric of similarity between variate i and j is defined as

$$d(i,j) = \sqrt{\sum_{k=1}^{r}(a_{ik} - a_{jk})^2}$$

Clustering based on Factor Modelling
3. Loading Matrix is Enough

Mapping:

- each variate of $\mathbf{Y}_t$
- each row of $\mathbf{A}$

One simple choice:
Pass $\mathbf{A}$ into

- k-means
- hierarchical
- .....

More 'complex':
still working on.....

Multiple time series $\mathbf{Y}_{N\times p}$

$$\begin{pmatrix} \mathbf{y}_1^\mathsf{T} \\ \mathbf{y}_2^\mathsf{T} \\ \vdots \\ \mathbf{y}_N^\mathsf{T} \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Np} \end{pmatrix}$$

Data matrix $\mathbf{A}_{p\times r}$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pr} \end{pmatrix}$$

Clustering based on Factor Modelling

This new method is validated by

- Monte-Carlo experiments
- Classical test datasets in time series clustering
  http://www.cs.ucr.edu/~eamonn/time_series_data

Extensive comparisons are made.

Five evaluation criterion

- Jaccard Score
- Rand Index
- Folkes and Mallow Index
- Cluster similarity measure
- Normalized mutual information

Three existing methods

- Directly with raw data
- Wavelet feature extraction
- AR(∞) approximation

[Details are reported in my working paper.]

Example: Hyper-spectral Imagery

Hyper-spectral Imagery Dataset (From 08/17 To 08/21 in 2013)

- Focus on a single pixel (spatial variation ignored)
- Daily temporal profiles: 08:30–17:00, every 30 minutes
- Wide range of wave bands: 374.26–1038.6 nm, 128 in total

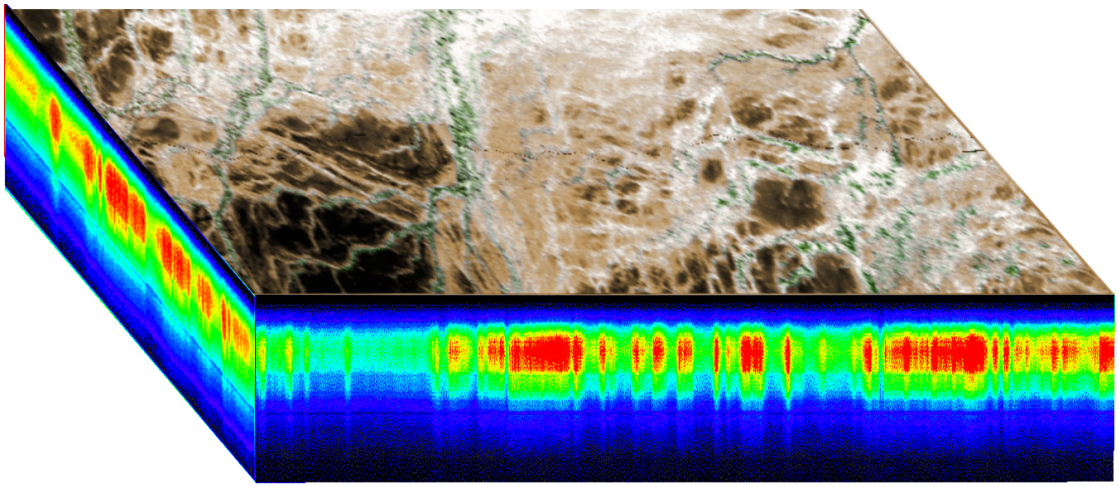
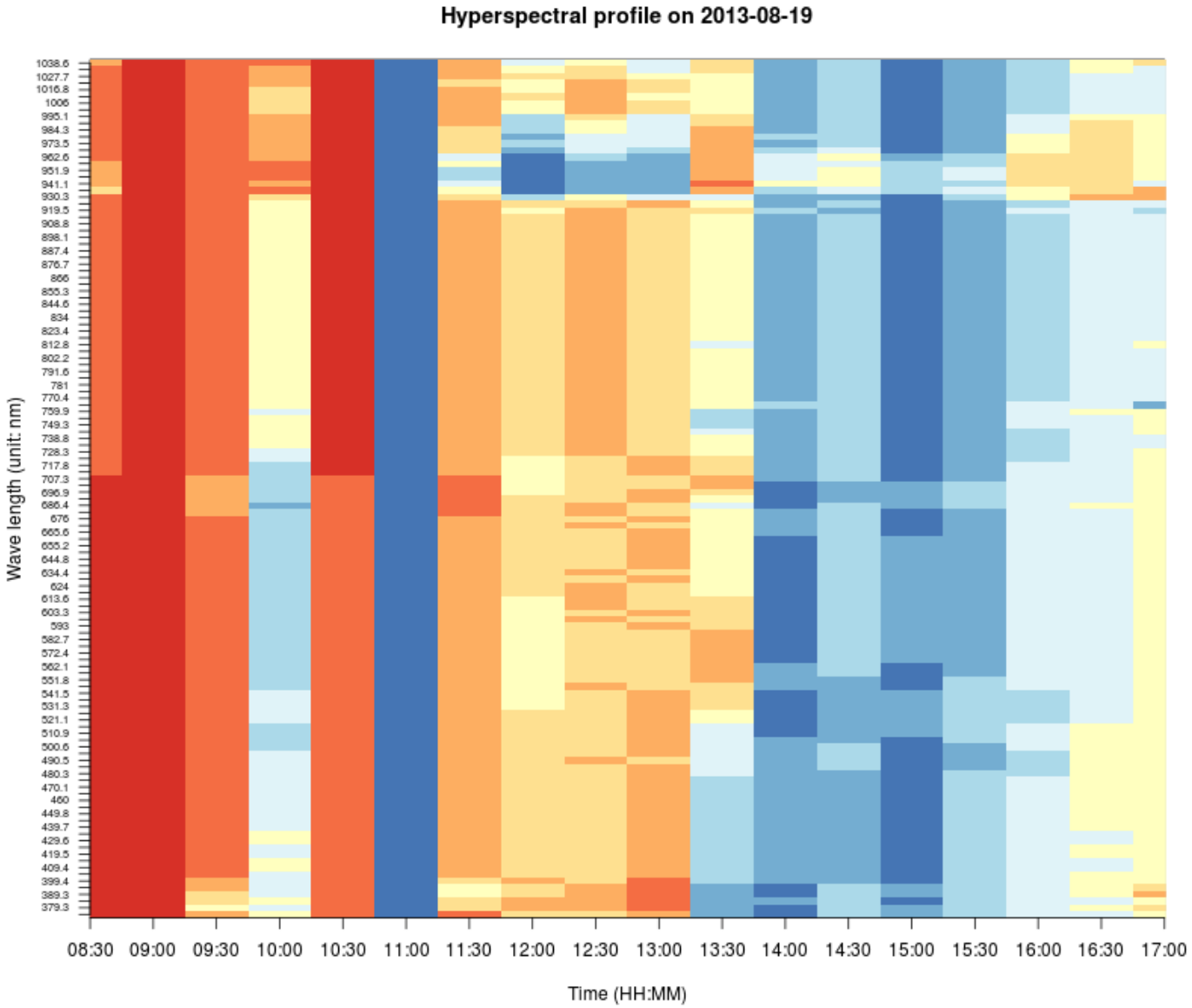


Figure: Hyper-spectral cube collected from an aircraft using SOC710 (*Hamada 2012*). The photo represents landscape of the study area and the profile represents the spectral reflectance of the 128 bands.

Example: Hyper-spectral Imagery



Example: Hyper-spectral Imagery

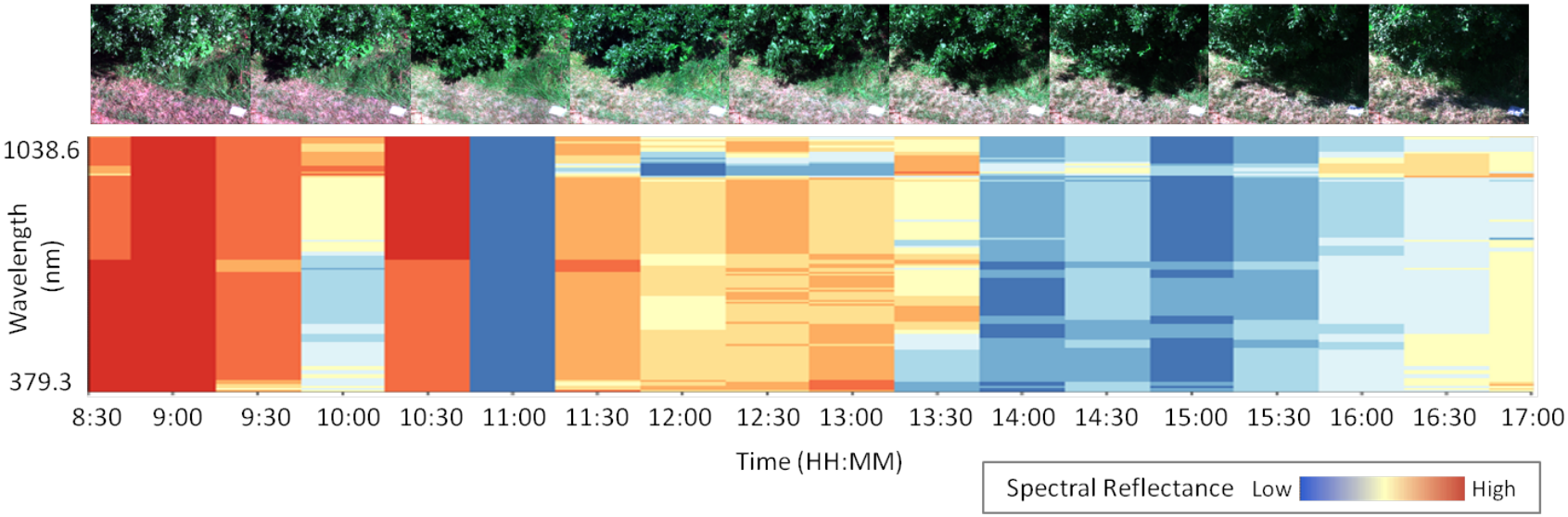
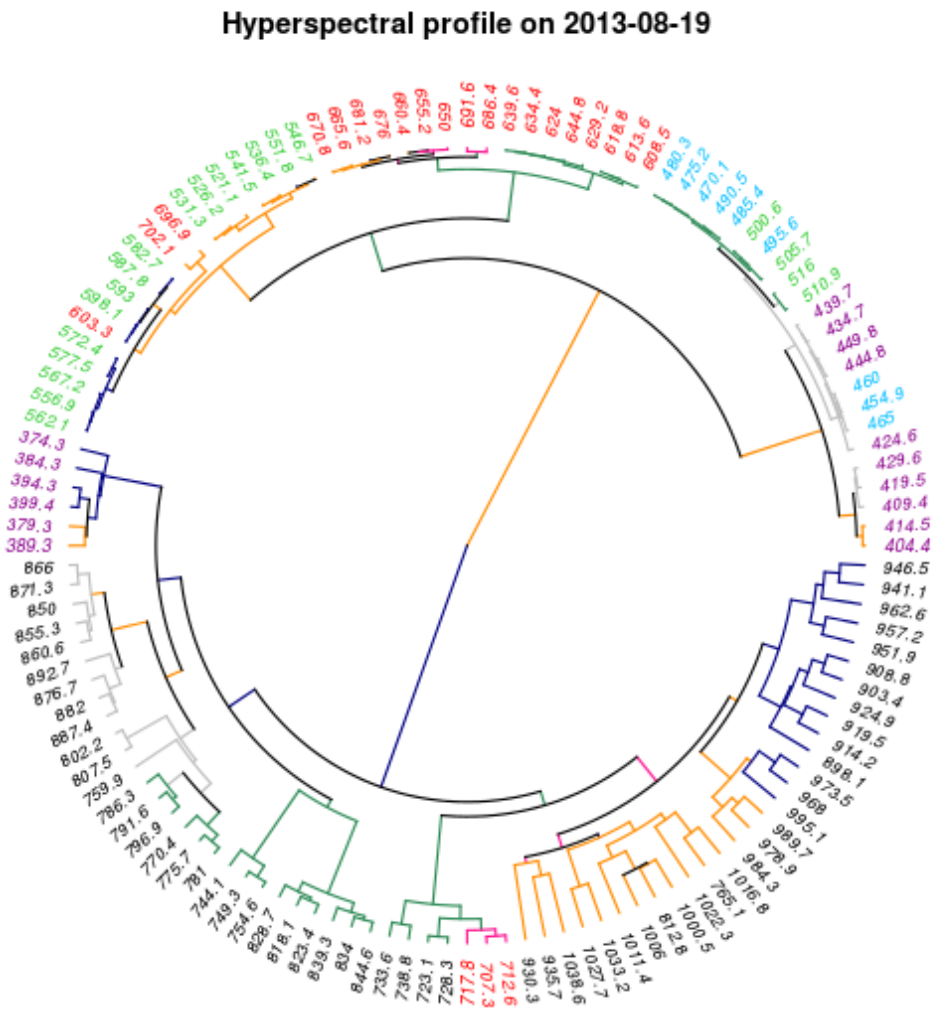


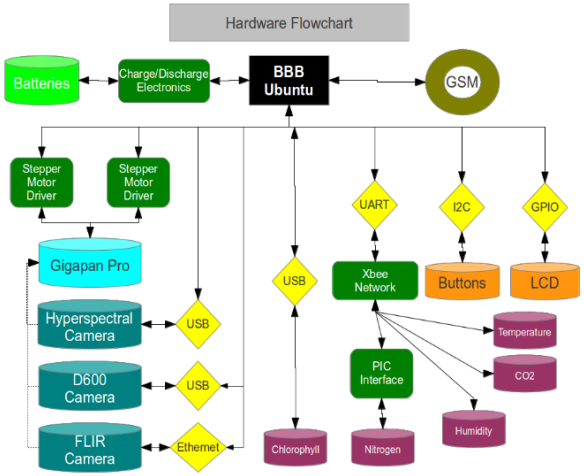
Figure: Reflectance time series from band 374.26 nm to 1038.60 nm on 2013-08-19, *Hamada (2013)*

Example: Hyper-spectral Imagery

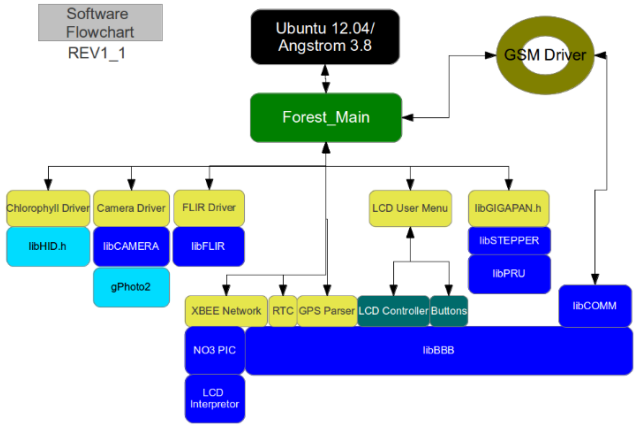
wavelength palette: <450 nm purple; 450-500 nm blue; 500-600 nm green; 600-720 nm red; >720 nm grey



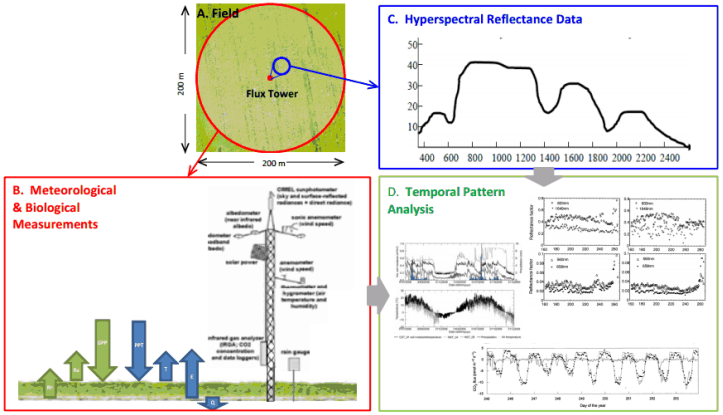
Environmental Sensor Network Systems



(a) Hardware Network, *Strunk (2013)*



(b) Software Network, *Strunk (2013)*

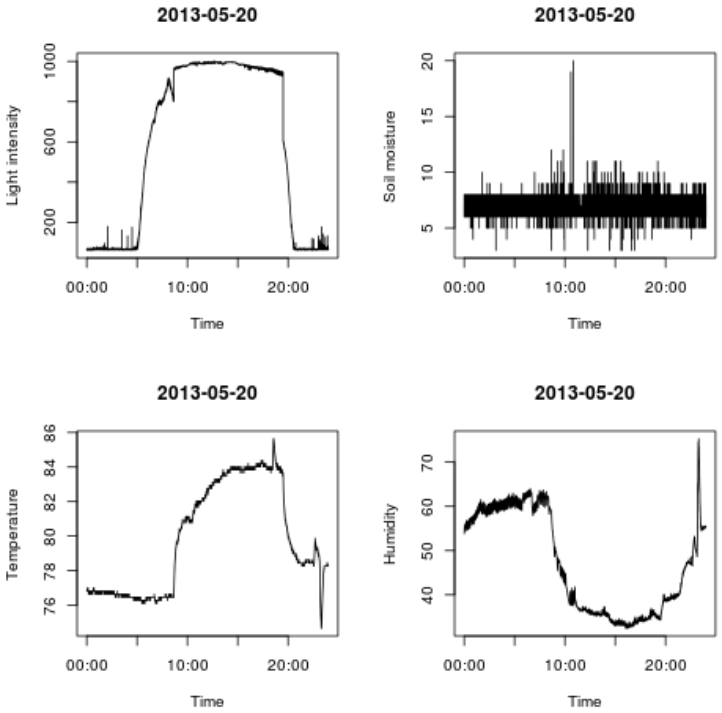


(c) Conceptual Diagram of Data Collection and Analysis, *Hamada (2013)*

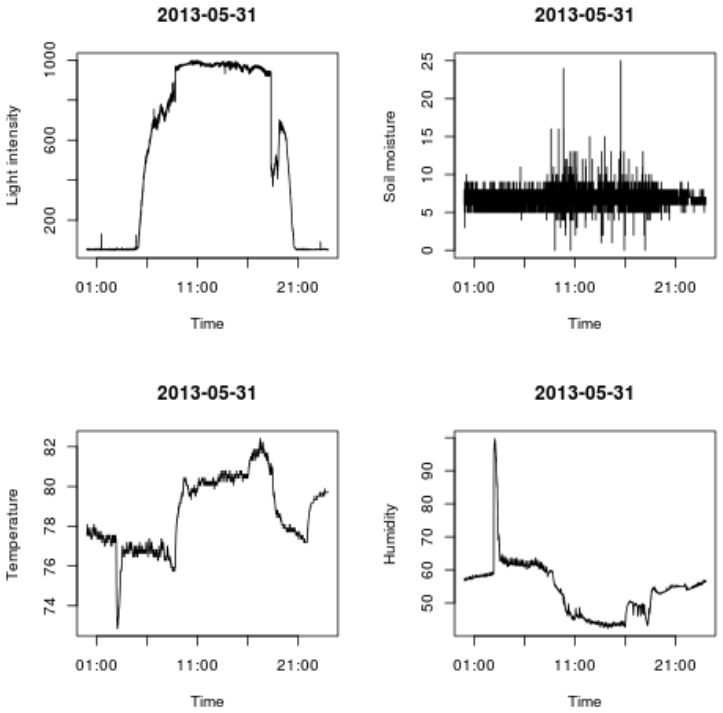
Indoor Environmental Time Series

Variables:

Temperature, Humidity, Soil moisture and Light intensity



(d) 2013-05-20



(e) 2013-05-31

Figure: Sensory Profiles of Plant on Two Days

Indoor Environmental Time Series

Inconsistent in temporal manner: feature of sensory datasets

```
"2013-05-16 11:09:38" "2013-05-16 11:11:32" "2013-05-16 11:11:35"
"2013-05-16 11:11:38" "2013-05-16 11:11:53" "2013-05-16 11:12:08"
"2013-05-16 11:12:53" "2013-05-16 11:13:01" "2013-05-16 11:13:05"
"2013-05-16 11:13:11" "2013-05-16 11:13:14" "2013-05-16 11:13:17"
"2013-05-16 11:13:23" "2013-05-16 11:15:18" "2013-05-16 11:15:27"
"2013-05-16 11:15:31" "2013-05-16 11:15:48" "2013-05-16 11:15:55"
"2013-05-16 11:16:04" "2013-05-16 11:16:10" "2013-05-16 11:16:13"
```

(a) 2013-05-16

```
"2013-05-17 03:21:34" "2013-05-17 04:00:53" "2013-05-17 04:15:26"
"2013-05-17 04:18:32" "2013-05-17 04:18:35" "2013-05-17 04:19:35"
"2013-05-17 04:19:50" "2013-05-17 04:22:44" "2013-05-17 04:23:53"
"2013-05-17 04:25:50" "2013-05-17 04:26:53" "2013-05-17 04:27:05"
"2013-05-17 04:28:20" "2013-05-17 04:29:05" "2013-05-17 04:30:56"
"2013-05-17 04:31:14" "2013-05-17 04:32:41" "2013-05-17 04:32:59"
"2013-05-17 04:33:18" "2013-05-17 04:34:48" "2013-05-17 04:38:27"
```

(b) 2013-05-17

```
"2013-05-18 06:06:14" "2013-05-18 07:01:00" "2013-05-18 07:31:45"
"2013-05-18 07:32:00" "2013-05-18 07:38:27" "2013-05-18 07:41:01"
"2013-05-18 07:45:55" "2013-05-18 07:47:07" "2013-05-18 07:47:13"
"2013-05-18 07:47:49" "2013-05-18 07:49:28" "2013-05-18 07:51:25"
"2013-05-18 07:51:31" "2013-05-18 07:52:13" "2013-05-18 07:53:04"
"2013-05-18 07:54:19" "2013-05-18 08:04:40" "2013-05-18 08:04:43"
"2013-05-18 08:04:46" "2013-05-18 08:04:49" "2013-05-18 08:04:55"
```

(c) 2013-05-18

Figure: First 20 time points in three different daily profiles

Data Preprocessing

`NA`: No data available; `DC`: Data collected

	T_1	T_2	T_3	T_4	...	T_l	...	T_k	...	T_{N-2}	T_{N-1}	T_N
Day 1	NA	DC	DC	NA	...	NA	...	DC	...	NA	DC	NA
Day 2	NA	NA	DC	DC	...	DC	...	DC	...	DC	DC	NA
Day 3	DC	NA	DC	DC	...	DC	...	DC	...	DC	DC	DC
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Day $s-1$	DC	DC	DC	DC	...	DC	...	NA	...	NA	DC	NA
Day s	NA	NA	NA	DC	...	DC	...	NA	...	DC	DC	NA
Day $s+1$	DC	DC	NA	DC	...	DC	...	NA	...	DC	DC	DC
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Day $M-2$	NA	DC	DC	NA	...	DC	...	DC	...	DC	DC	NA
Day $M-1$	DC	DC	DC	DC	...	DC	...	NA	...	DC	DC	DC
Day M	NA	DC	DC	DC	...	DC	...	NA	...	DC	DC	NA

Indoor Environmental Time Series Tell Us

For most real world problems regarding time series

- Contamination: Significant noise, Extraneous data, ...
- Series from the same class might be various in length.

General solution

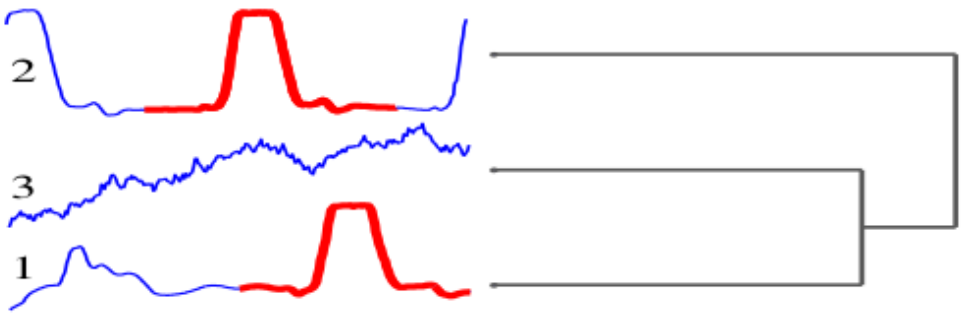
Refine the raw data and Extract the most useful information

Zakaria et al. (2012) extended a recently-introduced concept in data mining, **shapelets**, and proposed a new method for time series clustering based on shapelets.

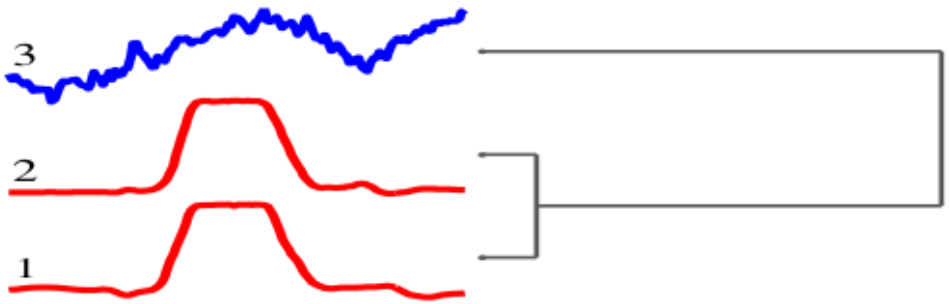
Unsupervised-shapelets

Basic Idea

Ignore part of the raw data and only use some **local patterns**



(a) Clustering of the whole data



(b) Clustering after ignoring some data

Figure: *Zakaria et al. 2012*

Unsupervised-shapelets

Heuristic Definition

Given a collection of time series **DATA**

- a subsequence $\mathbf{S}^\star$ of a time series **T** in **DATA**
- an effective separation for the entire dataset **DATA**:
  - Group $\mathbf{D}_A \Rightarrow d_A$ '*distance*' between $\mathbf{S}^\star$ and $\mathbf{D}_A$
  - Group $\mathbf{D}_B \Rightarrow d_B$ '*distance*' between $\mathbf{S}^\star$ and $\mathbf{D}_B$
  - Effective separation: d_A '*much smaller/larger than*' d_B

Figure: Illustration of *effective separation* of $\mathbf{D}_A$ and $\mathbf{D}_B$ (*Zakaria et al. 2012*)

Unsupervised Shapelets

Three kinds of distances between series

1. $dist(\cdot, \cdot)$: length-normalized Euclidean distance
2. Distance between shorter series $\mathbf{S}^\star$ and longer series $\mathbf{T}$

$$sdist(\mathbf{S}, \mathbf{T}) = \min_{1 \le i \le n-m} dist(\mathbf{S}, \mathbf{T}_{i,m})$$

3. Distance between subsequence $\mathbf{S}^\star$ and a group of series $\mathbf{D}_A$

$$d_A := \{sdist(\mathbf{S}^\star, \mathbf{T}) : \mathbf{T} \in \mathbf{D}_A\}$$

Two more question

- How to decide the division of **DATA** to obtain $\mathbf{D}_A$ and $\mathbf{D}_B$?
- What if there are more than one unsupervised shapelet?

They will be answered when you see ♣ and ♣♣ later on.

Unsupervised Shapelets Clustering

Algorithm Overview

1. Extracting all shapelets
2. Pruning away some shapelets
3. Clustering time series

Unsupervised Shapelets Clustering: 1. Shapelet Extraction

Separation Measure *Gap*

Given a shapelet $\mathbf{S}^\star$. Suppose that $\mathbf{S}^\star$ can divide **DATA** into two groups of time series $\mathbf{D}_A$ and $\mathbf{D}_B$. Let

	$sdist(\mathbf{S}^\star, \mathbf{D}_A)$	$sdist(\mathbf{S}^\star, \mathbf{D}_B)$
mean	μ_A	μ_B
variance	σ_A^2	σ_B^2

The separation measure is defined as

$$Gap(\mathbf{D}_A, \mathbf{D}_B, \mathbf{S}^\star) = (\mu_B - \sigma_B) - (\mu_A + \sigma_A)$$

Essence of Unsupervised Shapelet Extraction

Greedy search to maximize *Gap* between any dichotomy of **DATA**

Unsupervised Shapelets Clustering: Shapelet Extraction

Shapelets Extraction

Initially, **T**←the first series in time series dataset **DATA**
while true
 for each subsequence length L
  for each subsequence $\mathbf{T}_{L,k}$ in a time series **T**
   ♣ Find the $\mathbf{D}_A$ and $\mathbf{D}_B$ s.t. $\text{gap}(\mathbf{D}_A, \mathbf{D}_B, \mathbf{T}_{L,k})$ is maximized
  end for
  Store the maximum of all gap values as **GAP**
 end for
 ♠ Find maximum of **GAP** and the corresponding subsequence $\mathbf{S}^\star$
 update T to be the most 'dissimilar' series to $\mathbf{S}^\star$ (Shapelet)
 remove 'some' series from **DATA** which are 'similar' to $\mathbf{S}^\star$
 stop until there is only one series 'similar' to $\mathbf{S}^\star$
end while
return the set of all shapelets

Unsupervised Shapelets Clustering: 2. Shapelet Selection

- Too many shapelets extracted
- Recall ♠ in the last page
- Selection $\rightsquigarrow$ Stopping criterion

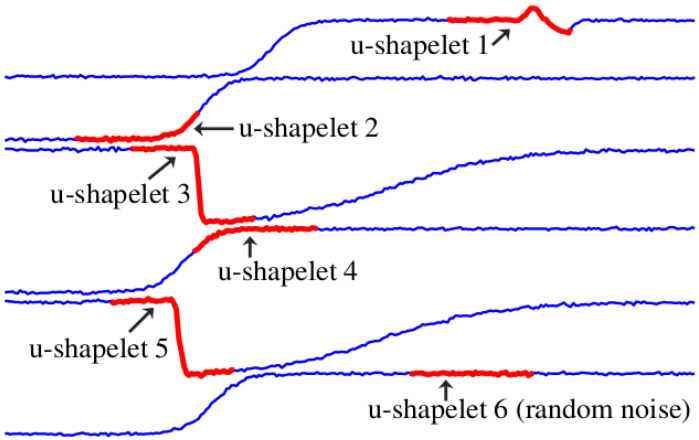


Figure: *Zakaria et al. 2012*

Unsupervised Shapelets Clustering: 2. Shapelet Selection

When to stop the searching process?

for i from 1 to the total number of all shapelets
  do clustering based on shapelet 1, ..., shapelet i (HOW?)
  do clustering based on shapelet 1, ..., shapelet $(i+1)$ (HOW??)
  CHANGE:=$Eva(i,i)$-$Eva(i,i+1) \stackrel{\star}{=} 1$-$Eva(i,i+1)$
end for
return the i minimizing **CHANGE** as the stopping step

Unsupervised Shapelets Clustering: 3. Clustering

Recall the red **HOW**s in last page.
Also recall the second question: what if more than one shapelets?

Distance Map

m shapelets in a dataset of N time series

$$\begin{pmatrix} sdist(\mathbf{S}_1^\star, \mathbf{T}_1) & sdist(\mathbf{S}_2^\star, \mathbf{T}_1) & \dots & sdist(\mathbf{S}_m^\star, \mathbf{T}_1) \\ sdist(\mathbf{S}_1^\star, \mathbf{T}_2) & sdist(\mathbf{S}_2^\star, \mathbf{T}_2) & \dots & sdist(\mathbf{S}_m^\star, \mathbf{T}_2) \\ \vdots & \vdots & \vdots & \vdots \\ sdist(\mathbf{S}_1^\star, \mathbf{T}_N) & sdist(\mathbf{S}_2^\star, \mathbf{T}_N) & \dots & sdist(\mathbf{S}_m^\star, \mathbf{T}_N) \end{pmatrix}_{N \times m} \quad (\clubsuit\clubsuit)$$

Pass the Distance Map into any off-the-shelve algorithms:

- $k-$means clustering
- hierarchical clustering
- .....

Revisit: Indoor Environmental Time Series

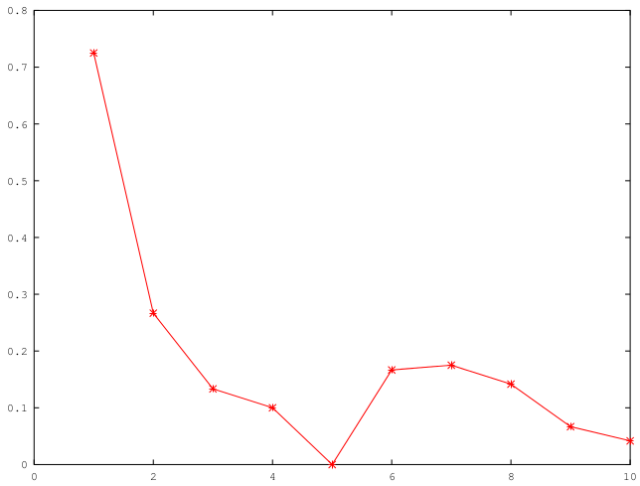
Data

- 16 daily sensory profiles
  - starting: 2013-05-18
  - ending: 2013-06-02
- 4 variates in each profile
  - light intensity
  - soil moisture
  - temperature
  - humidity

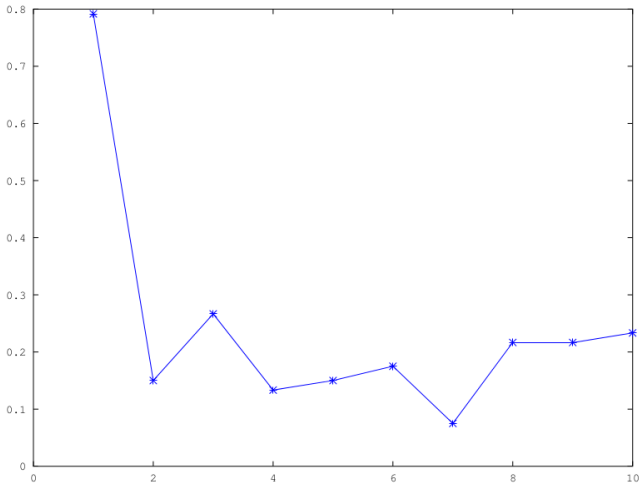
Results

variable	# of shapelets
light intensity	**10**
soil moisture	**10**
temperature	3
humidity	4

Revisit: Indoor Environmental Time Series



(a) Light intensity



(b) Soil moisture

Figure: Plots of Changes in the Values of Rand Index

Revisit: Indoor Environmental Time Series

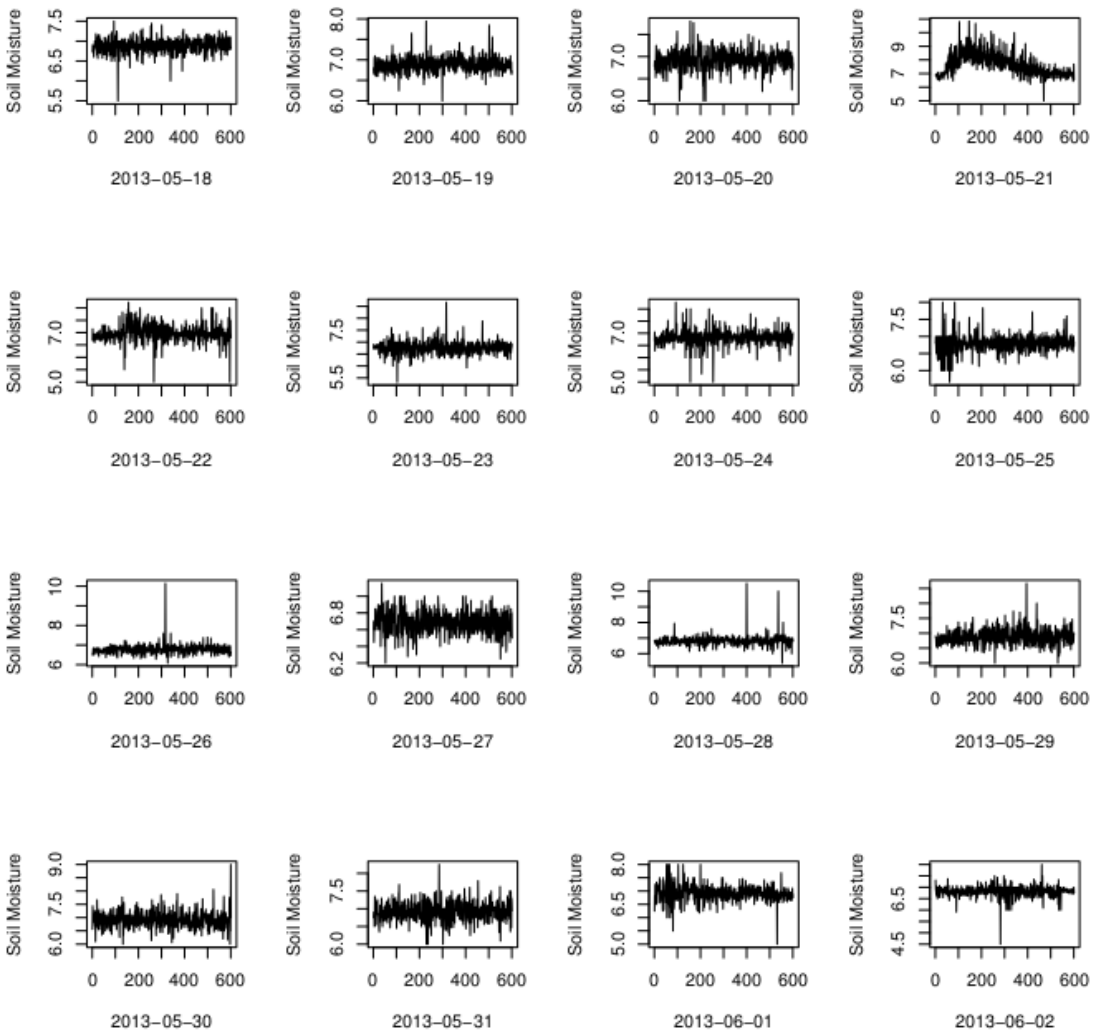


Figure: Time series plots for 16 daily soil moisture profiles

Revisit: Indoor Environmental Time Series

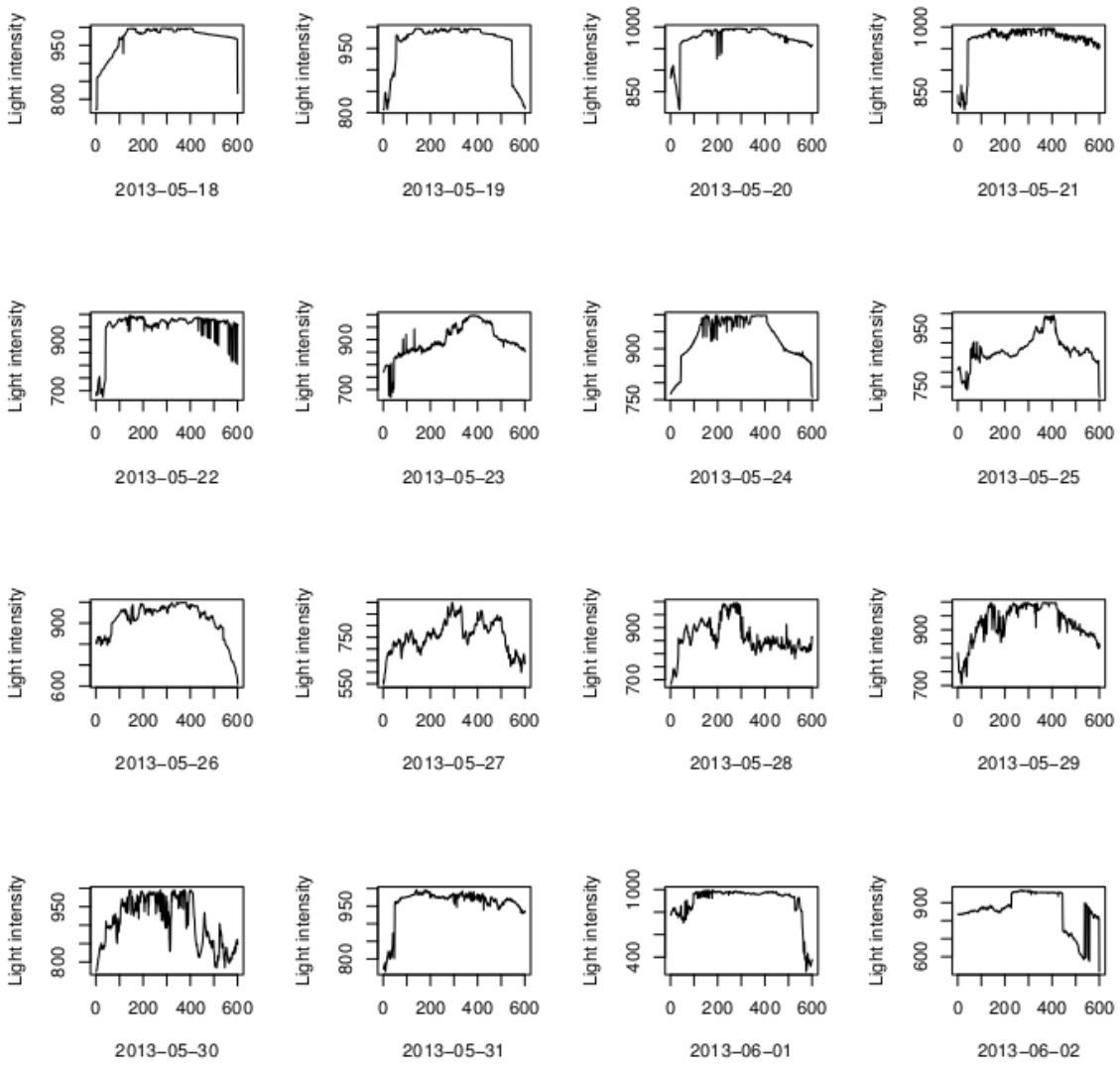
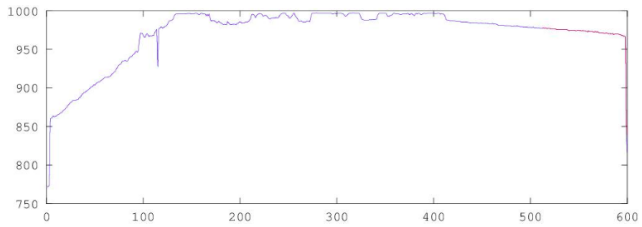
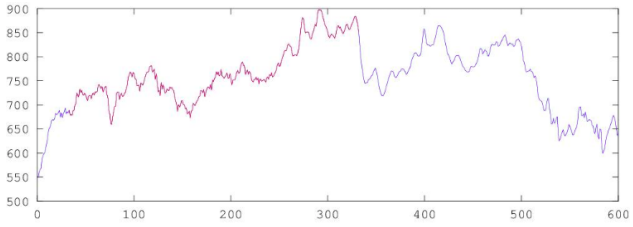


Figure: Time series plots for 16 daily light intensity profiles

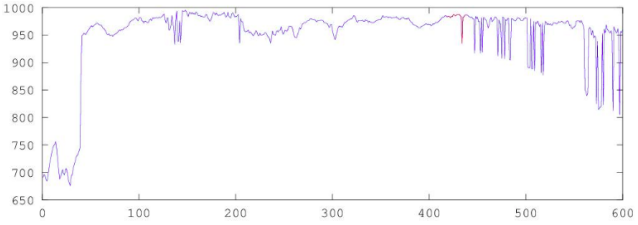
Revisit: Indoor Environmental Time Series



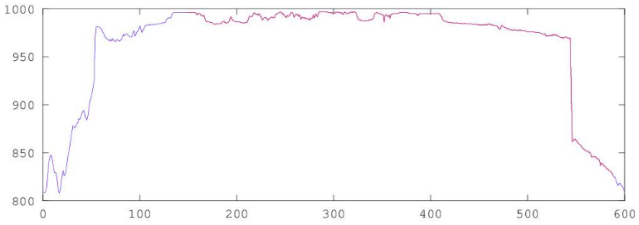
(a) 1st shapelet



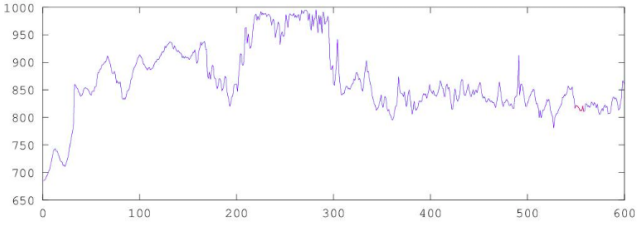
(b) 2nd shapelet



(c) 3rd shapelet



(d) 4th shapelet



(e) 5th shapelet

Figure: Selected 5 shapelets for light intensity profiles

Conclusion

Factor-modelling-based Clustering

- Effective in the high-dimensional settings
- Can be applied the 'too-many-variables' case
- How to generalize to non-stationary case?

Unsupervised-shapelet-based Clustering

- Series can be different in lengths
- Cannot handle series with missing values
- Open a door for research on event detection

Acknowledgement

- FOREST Group, Beckman, P. et al., ANL, 2013
  - Hyper-spectral: Hamada, Y. & Keever, E.
  - Sensor: Sankaran, R.
  - Field site: Adamiec, J.
- External Help
  - Keogh E., University of California, Riverside
  - Bagnall, A, The University of East Anglia, UK
  - Vilar, A., Universidade da Coruña, Spain

This summer is merely a start!

Ongoing Sub-projects

- Shape-based clustering for short time series
- Missing-value-tolerant feature extraction algorithm
- Borrowing strengths: co-clustering across all measurements
- Multi-dimensional time series clustering
- .....

Update will be posted on

- STAT page on FOREST blog:
  http://press3.mcs.anl.gov/forest/stat
- Personal page:
  https://sites.google.com/site/herbertxiangzhu/research

Thank you!