Time Series Data Mining - from PhD to Startup

Peter Laurinec

October 27, 2018

POWEREX

Highlights

Time series data mining - from PhD to start-up:

Highlights

Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),

Highlights

Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),
    - TS data mining methods,

Highlights

Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),
  - TS data mining methods,
- PhD. study thesis - combining and developing TS data mining methods,

Highlights

Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),
  - TS data mining methods,
- PhD. study thesis - combining and developing TS data mining methods,
- TSrepr R package - TS representations,

Highlights

Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),
  - TS data mining methods,
- PhD. study thesis - combining and developing TS data mining methods,
- TSrepr R package - TS representations,
- Work after Phd - energy start-up,

Highlights

Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),
  - TS data mining methods,
- PhD. study thesis - combining and developing TS data mining methods,
- TSrepr R package - TS representations,
- Work after Phd - energy start-up,
  - Differences and my thoughts,

Highlights

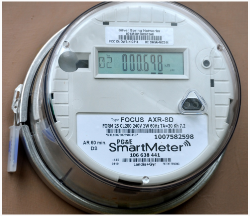
Time series data mining - from PhD to start-up:

- Problems and solutions for using large amount of long time series (TS),
    - TS data mining methods,
- PhD. study thesis - combining and developing TS data mining methods,
- TSrepr R package - TS representations,
- Work after Phd - energy start-up,
    - Differences and my thoughts,
    - What we do there...

Time Series Data in Energetics

Smart metering

- Measuring electricity consumption or production (photovoltaic panels) from every consumer or producer (together prosumer) every 5, 15, or 30 minutes,
- This creates a large amount of time series data,
- 3 years of data from consumer 96*365*3 = 105120...from 10 thousand consumers... > 1 billion rows of multiple columns,
- Smart grid - set of consumers and producers,



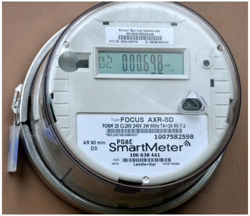
Time Series Data in Energetics

Smart metering

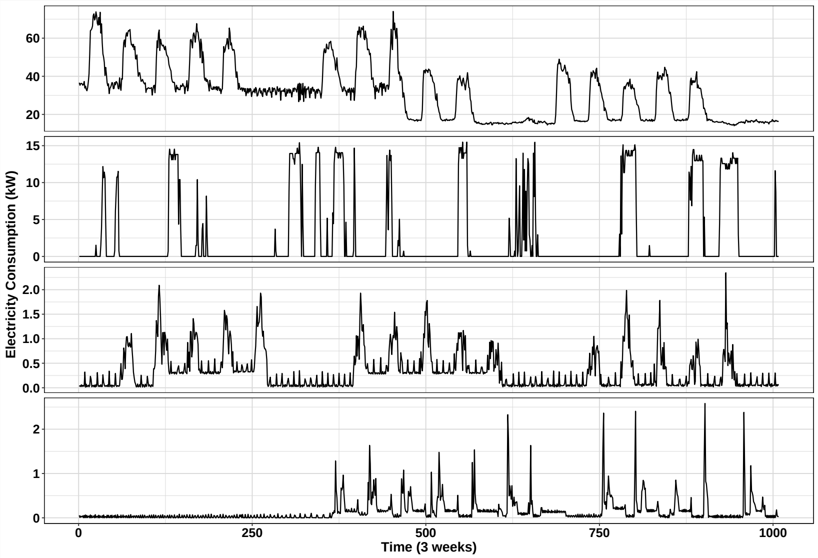
- Measuring electricity consumption or production (photovoltaic panels) from every consumer or producer (together prosumer) every 5, 15, or 30 minutes,
- This creates a large amount of time series data,
- 3 years of data from consumer 96*365*3 = 105120...from 10 thousand consumers... > 1 billion rows of multiple columns,
- Smart grid - set of consumers and producers,

Characteristics:

- High-dimensionality,
- Multiple seasonalities (daily, weekly, yearly),
- Large amount of stochastic factors as: weather, holidays, black-outs, changes on market etc.



Examples of Consumers TS



Typical Use Cases

- Forecasting el. consumption or production - market planning, black-outs prevention etc.,

Typical Use Cases

- Forecasting el. consumption or production - market planning, black-outs prevention etc.,
- Extract typical profiles of consumption - changes in tariffs, create new ones etc.,

Typical Use Cases

- Forecasting el. consumption or production - market planning, black-outs prevention etc.,
- Extract typical profiles of consumption - changes in tariffs, create new ones etc.,
- Optimizing electricity consumption of some consumer,

Typical Use Cases

- Forecasting el. consumption or production - market planning, black-outs prevention etc.,
- Extract typical profiles of consumption - changes in tariffs, create new ones etc.,
- Optimizing electricity consumption of some consumer,
- Optimizing whole smart grid,

Typical Use Cases

- Forecasting el. consumption or production - market planning, black-outs prevention etc.,
- Extract typical profiles of consumption - changes in tariffs, create new ones etc.,
- Optimizing electricity consumption of some consumer,
- Optimizing whole smart grid,
- Monitoring smart grid,

Typical Use Cases

- Forecasting el. consumption or production - market planning, black-outs prevention etc.,
- Extract typical profiles of consumption - changes in tariffs, create new ones etc.,
- Optimizing electricity consumption of some consumer,
- Optimizing whole smart grid,
- Monitoring smart grid,
- Anomaly detection.

TS Data Mining Methods

- Methods for working with TS:

TS Data Mining Methods

- Methods for working with TS:
  - TS representations,

TS Data Mining Methods

- Methods for working with TS:
  - TS representations,
  - TS distance measures,

TS Data Mining Methods

- Methods for working with TS:
  - TS representations,
  - TS distance measures,
- Tasks:

TS Data Mining Methods

- Methods for working with TS:
  - TS representations,
  - TS distance measures,
- Tasks:
  - TS classification,
  - TS clustering,
  - TS forecasting,
  - TS anomaly detection,
  - TS indexing.

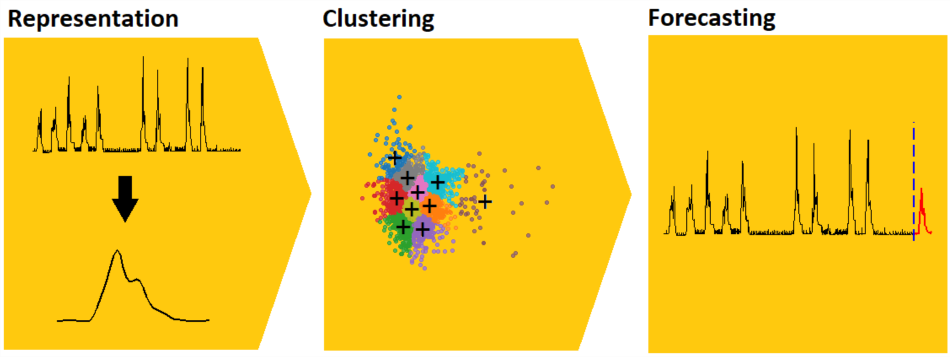
PhD. Thesis Goals

- The thesis had the goal to investigate, in the broader context, the **usage of time series data mining (analysis) methods** in order to **improve the predictive performance of machine learning methods** and its combinations.

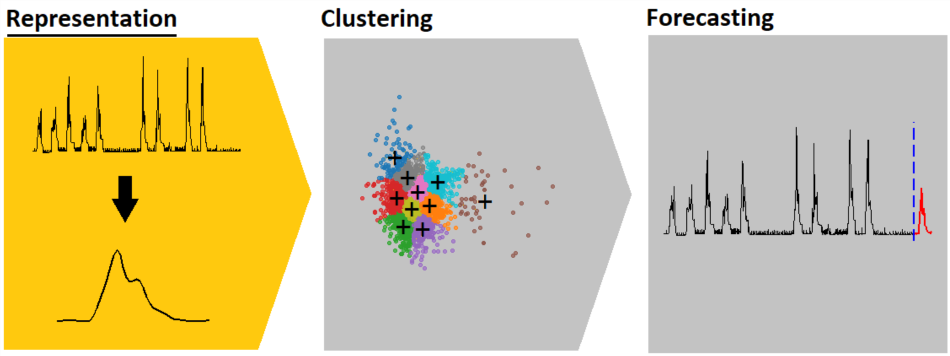
PhD. Thesis Goals

- The thesis had the goal to investigate, in the broader context, the **usage of time series data mining (analysis) methods** in order to **improve the predictive performance of machine learning methods** and its combinations.
- In more detail, the goal was to investigate the usage of various **time series representations** for seasonal time series, **clustering**, and **forecasting methods** for electricity consumption forecasting accuracy improvement.

Approach Overview



I. Time Series Representations



I. Time Series Representations

What can we do for solving problems with high-dimensional TS?

I. Time Series Representations

What can we do for solving problems with high-dimensional TS?

- Use time series representations!

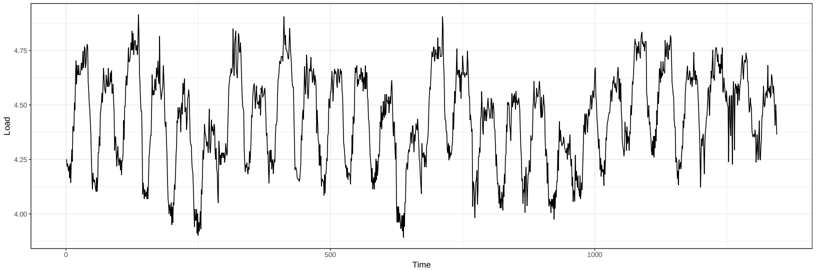
I. Time Series Representations

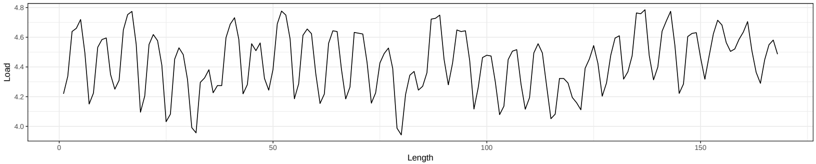
What can we do for solving problems with high-dimensional TS?

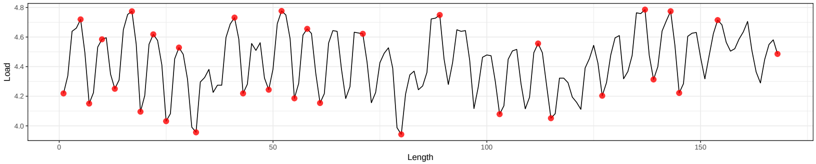
- **Use time series representations!**

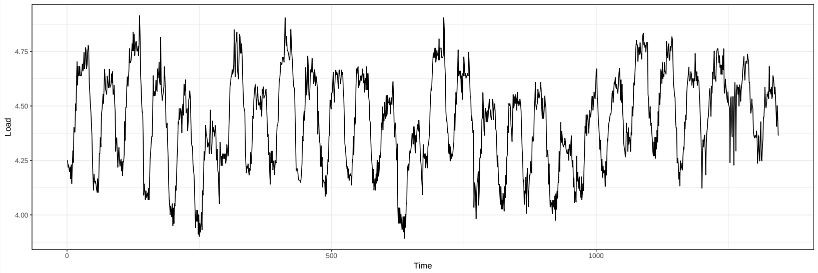
They are excellent to:

- Reduce memory load.
- Accelerate subsequent machine learning algorithms.
- Implicitly remove noise from the data.
- Emphasize the essential characteristics of the data.
- Help to find patterns in data (or motifs).

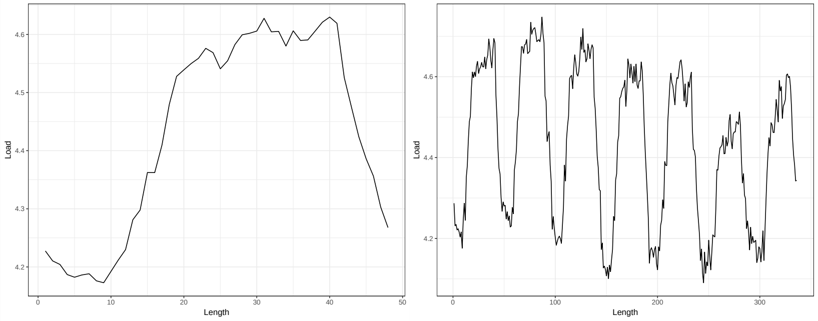

4.75
4.50
4.25
4.00
Load
0
500
1000
Time


4.8
4.6
4.4
4.2
4.0
Load
0
50
100
150
Length


4.8
4.6
4.4
4.2
4.0
Load
0
50
100
150
Length



Load
Time
4.75
4.50
4.25
4.00
0
500
1000



Load
Length
4.6
4.5
4.4
4.3
4.2
0
10
20
30
40
50
4.6
4.4
4.2
0
100
200
300

I. Time Series Representations [1]

I used TS representations for:

[1] Laurinec P., Lucká M., Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016.

I. Time Series Representations [1]

I used TS representations for:

- Dimensionality reduction (curse of dimensionality),

[1] Laurinec P., Lucká M., Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016.

I. Time Series Representations [1]

I used TS representations for:

- Dimensionality reduction (curse of dimensionality),
- Emphasising the main characteristics of data,

[1] Laurinec P., Lucká M., Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016.

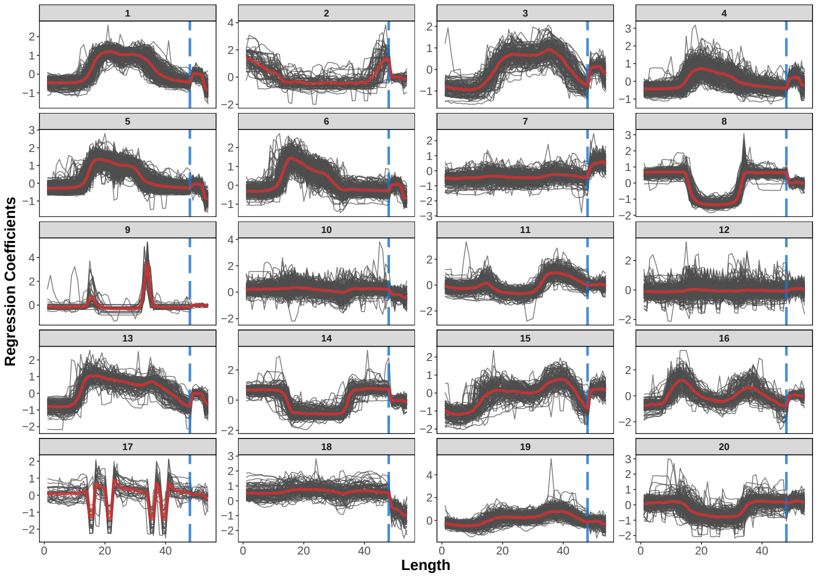
I. Time Series Representations [1]

I used TS representations for:

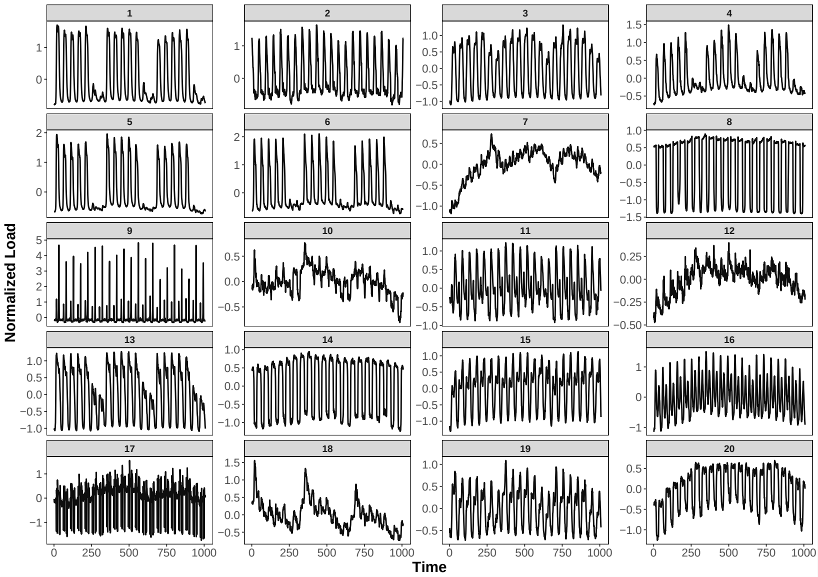
- Dimensionality reduction (curse of dimensionality),
- Emphasising the main characteristics of data,
- More accurate clustering of consumers TS to create more predictable (forecastable) groups of aggregated TS of electricity consumption.

[1] Laurinec P., Lucká M., Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016.

Clustered TS Representations



Groups of Aggregated TS



TSrepr

TSrepr - CRAN[2], GitHub[3]

- R package for time series representations computing
- Large amount of various methods are implemented
- Several useful support functions are also included
- Easy to extend and to use

```
data <- rnorm(1000)
repr_paa(data, func = median, q = 10)
```

[2] https://CRAN.R-project.org/package=TSrepr
[3] https://github.com/PetoLau/TSrepr/

All type of time series representations methods are implemented, so far these:

- PAA - Piecewise Aggregate Approximation (`repr_paa`)
- DWT - Discrete Wavelet Transform (`repr_dwt`)
- DFT - Discrete Fourier Transform (`repr_dft`)
- DCT - Discrete Cosine Transform (`repr_dct`)
- PIP - Perceptually Important Points (`repr_pip`)
- SAX - Symbolic Aggregate Approximation (`repr_sax`)
- PLA - Piecewise Linear Approximation (`repr_pla`)
- Mean seasonal profile (`repr_seas_profile`)
- Model-based seasonal representations based on linear model (`repr_lm`)
- FeaClip - Feature extraction from clipping representation (`repr_feaclip`)

Additional useful functions are implemented as:

- Windowing (`repr_windowing`)
- Matrix of representations (`repr_matrix`)
- Normalisation functions - z-score (`norm_z`), min-max (`norm_min_max`)

Usage of TSrepr

```
mat <- "some matrix with lot of time series"
```

```
mat_reprs <- repr_matrix(mat, func = repr_lm,
      args = list(method = "rlm", freq = c(48, 48*7)),
      normalise = TRUE, func_norm = norm_z)
```

```
mat_reprs <- repr_matrix(mat, func = repr_feaclip,
      windowing = TRUE, win_size = 48)
```

```
clustering <- kmeans(mat_reprs, 20)
```

Simple Extensibility of TSrepr

Example #1:

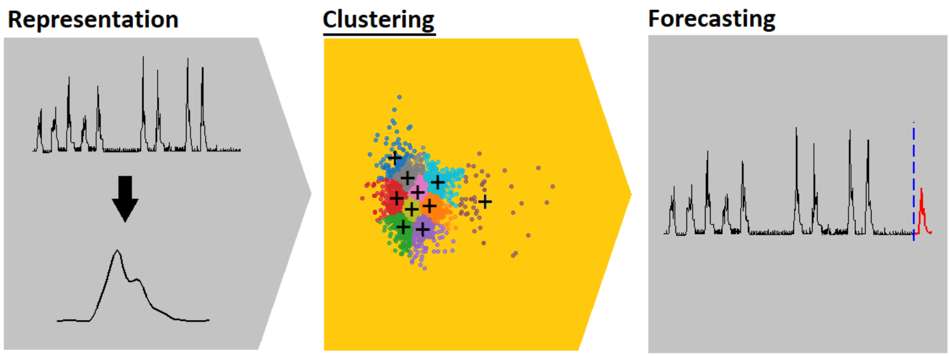
```
library(moments)
data_ts_skew <- repr_paa(data, q = 48, func = skewness)
```

Example #2:

```
repr_fea_extract <- function(x)
      c(mean(x), median(x), max(x), min(x), sd(x))

data_fea <- repr_windowing(data,
      win_size = 100, func = repr_fea_extract)
```

II. Time Series Clustering



II. Clustering Multiple Data Streams [4]

Motivation:

[4] https://github.com/PetoLau/ClipStream/

II. Clustering Multiple Data Streams [4]

Motivation:

- Deal with velocity of data coming,

[4] https://github.com/PetoLau/ClipStream/

II. Clustering Multiple Data Streams [4]

Motivation:

- Deal with velocity of data coming,
- Dynamic change of number of clusters,

[4] https://github.com/PetoLau/ClipStream/

II. Clustering Multiple Data Streams [4]

Motivation:

- Deal with velocity of data coming,
- Dynamic change of number of clusters,
- Automatic anomaly detection (anomalous consumers),

[4] https://github.com/PetoLau/ClipStream/

II. Clustering Multiple Data Streams [4]

Motivation:

- Deal with velocity of data coming,
- Dynamic change of number of clusters,
- Automatic anomaly detection (anomalous consumers),
- Automatic change detection.

[4] https://github.com/PetoLau/ClipStream/

II. Clustering Multiple Data Streams [4]

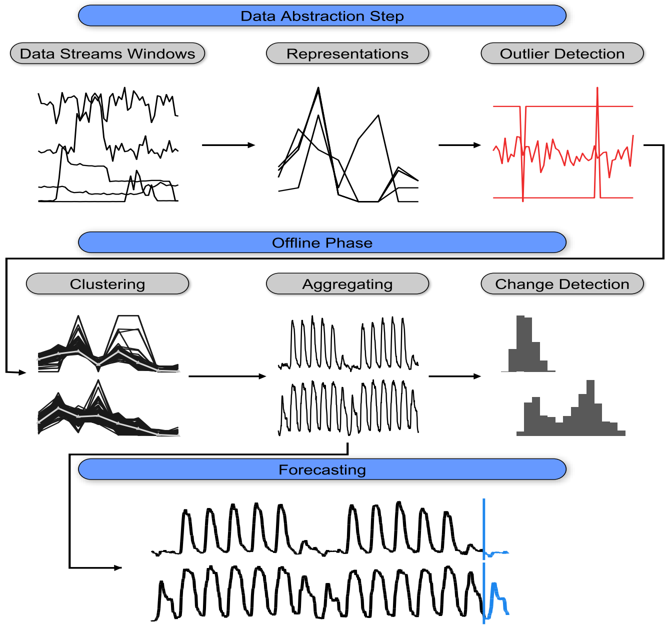
Motivation:

- Deal with velocity of data coming,
- Dynamic change of number of clusters,
- Automatic anomaly detection (anomalous consumers),
- Automatic change detection.

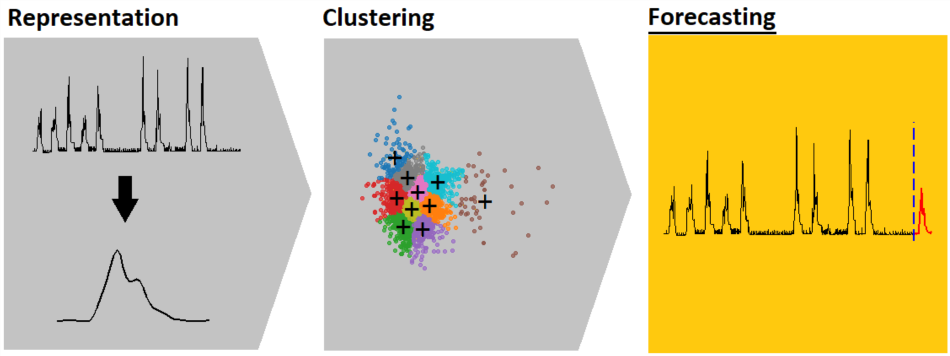
Approach:

- Take advantage of incrementality of clipped representation (windowing),
- Fast detection of anomalous consumers from extracted features from clipping,
- Change detection by Anderson-Darling test.

[4] `https://github.com/PetoLau/ClipStream/`


Data Abstraction Step
Data Streams Windows
Representations
Outlier Detection
Offline Phase
Clustering
Aggregating
Change Detection
Forecasting

III. Time Series Forecasting



III. Time Series Forecasting

Large number of methods suitable for forecasting:

- Time series analysis methods:
  - ARIMA,
  - Exponential smoothing,
  - Theta,

III. Time Series Forecasting

Large number of methods suitable for forecasting:

- Time series analysis methods:
  - ARIMA,
  - Exponential smoothing,
  - Theta,
- Regression methods:
  - Linear regression, GAM,
  - SVR, Gaussian process,
  - Regression trees, Bagging, Random Forest, Boosting,
  - Artificial Neural Networks.

III. Time Series Forecasting [5]

Finding the most suitable forecasting methods **with clustering**...

- STL+ARIMA, Exponential smoothing, Tree-based methods, Advanced ANNs (S2S + LSTM nets).

[5] https://github.com/PetoLau/TSMedianBasedEnsembleLearning/, https://github.com/PetoLau/UnsupervisedEnsembles/, https://github.com/PetoLau/DensityEnsembles/

III. Time Series Forecasting [5]

Finding the most suitable forecasting methods **with clustering**...

- STL+ARIMA, Exponential smoothing, Tree-based methods, Advanced ANNs (S2S + LSTM nets).

The problem of choosing the most suitable method among the set of methods...

Solution:

- **Ensemble learning** - combining forecasts.

[5] https://github.com/PetoLau/TSMedianBasedEnsembleLearning/, https://github.com/PetoLau/UnsupervisedEnsembles/, https://github.com/PetoLau/DensityEnsembles/

Life after PhD

- I was happy to be hired by start-up PowereX.
- We solve problems strongly related with my thesis.

Life after PhD

- I was happy to be hired by start-up **PowereX**.
- We solve problems strongly related with my thesis.

PowereX

- **P2P energy sharing** - commodity and also capacity,
- Analysis of consumers smart meter data,
- Forecasting and modelling maximal load (hourly, daily, etc.).

Differences between PhD and Business

PhD:

- Strong focus on accuracy measures - $\downarrow$ % of Mean Absolute Percentage Error, or internal validation indexes for clustering...

Differences between PhD and Business

PhD:

- Strong focus on accuracy measures - $\downarrow$ % of Mean Absolute Percentage Error, or internal validation indexes for clustering...
- Many times working with poor academic datasets.

Differences between PhD and Business

PhD:

- Strong focus on accuracy measures - ↓ % of Mean Absolute Percentage Error, or internal validation indexes for clustering...
- Many times working with poor academic datasets.

Business:

- Finding real value for customers,
- Accuracy is not that important,
- Working on real rich data.

Differences between PhD and Business

PhD:

- Strong focus on accuracy measures - $\downarrow$ % of Mean Absolute Percentage Error, or internal validation indexes for clustering...
- Many times working with poor academic datasets.

Business:

- Finding real value for customers,
- Accuracy is not that important,
- Working on real rich data.

But...they are also related and need each other...

Conclusions

TS data mining:

Conclusions

TS data mining:

- TS representations are our fiends in clustering, forecasting, classification etc.,

Conclusions

TS data mining:

- TS representations are our fiends in clustering, forecasting, classification etc.,
- Implemented in TSrepr package,

Conclusions

TS data mining:

- TS representations are our fiends in clustering, forecasting, classification etc.,
- Implemented in **TSrepr** package,
- PhD study is great practice before work.

Conclusions

TS data mining:

- TS representations are our fiends in clustering, forecasting, classification etc.,
- Implemented in **TSrepr** package,
- PhD study is great practice before work.

Questions: Peter Laurinec `laurinec.peter@gmail.com`
Code: `https://github.com/PetoLau/`
More research: `https://petolau.github.io/research`
Blog: `https://petolau.github.io`