Comparison of Representations of Time Series for Clustering Smart Meter Data

Peter Laurinec, and Mária Lucká

21.10.2016

Slovak University of Technology in Bratislava

Overview

- Motivation
- Data
- Proposed approach
- Representations of time series
- Results
- Conclusion

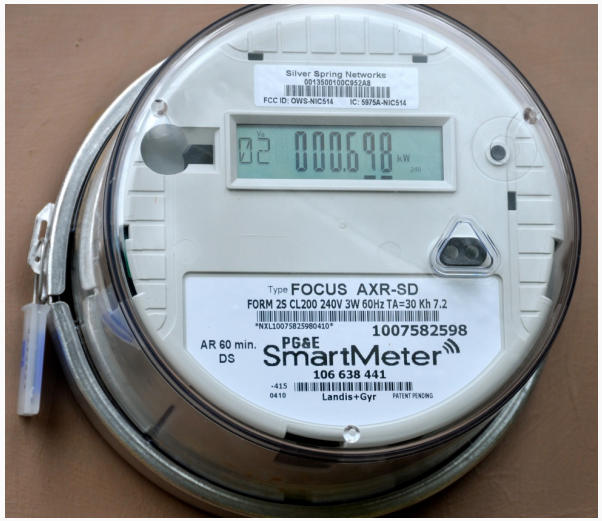
Motivation

More accurate forecast of electricity consumption is needed due to:

- Optimization of electricity consumption.
- Production of electricity. Overvoltage in grid.
- Distribution (utility) companies. Deregulation of the market. Purchase and sale of electricity.

Smart grids:

- Intelligent networks.
- Smart meters (every consumer has them).
- Usually 96 measurements per day.
- Advanced methods of forecast.



Forecast methods

Factors influencing electricity load:

- Seasonality (daily, weekly, monthly, ...)
- Weather (temperature, humidity, ...)
- Holidays
- Random effects

Methods:

- Time series analysis
- Regression
  - Linear model
  - AI methods
- Time series data mining + clustering

Clustering and forecast of load

Hypothesis:

Clustering and forecast of load

Hypothesis:

- Clustering of consumers (creation of more predictable groups of consumers) improves forecast accuracy against simple aggregate forecast.

Clustering and forecast of load

Hypothesis:

- Clustering of consumers (creation of more predictable groups of consumers) improves forecast accuracy against simple aggregate forecast.
- Classification of new consumers to the existing clusters does not affect accuracy of forecast.

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

- 3639 consumers. Residences.

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

- 3639 consumers. Residences.
- 48 measurements per day.

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

- 3639 consumers. Residences.
- 48 measurements per day.
- Test set from two months (February and September).

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

- 3639 consumers. Residences.
- 48 measurements per day.
- Test set from two months (February and September).

Slovakia

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

- 3639 consumers. Residences.
- 48 measurements per day.
- Test set from two months (February and September).

Slovakia

- 11281 consumers. Enterprises.

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

Ireland

- 3639 consumers. Residences.
- 48 measurements per day.
- Test set from two months (February and September).

Slovakia

- 11281 consumers. Enterprises.
- 96 measurements per day.

Data from smart meters

Available data come from smart meters from Ireland and Slovakia.

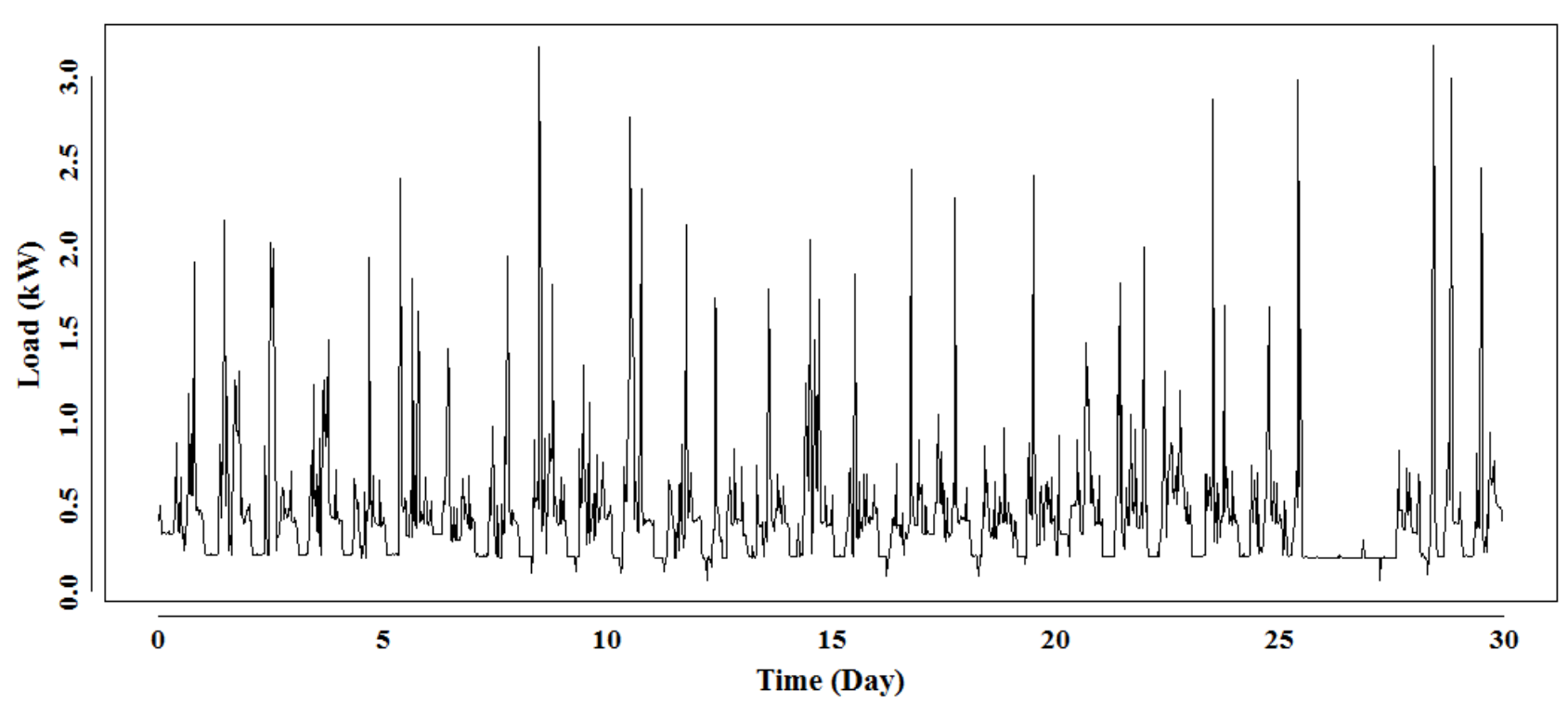
Ireland

- 3639 consumers. Residences.
- 48 measurements per day.
- Test set from two months (February and September).

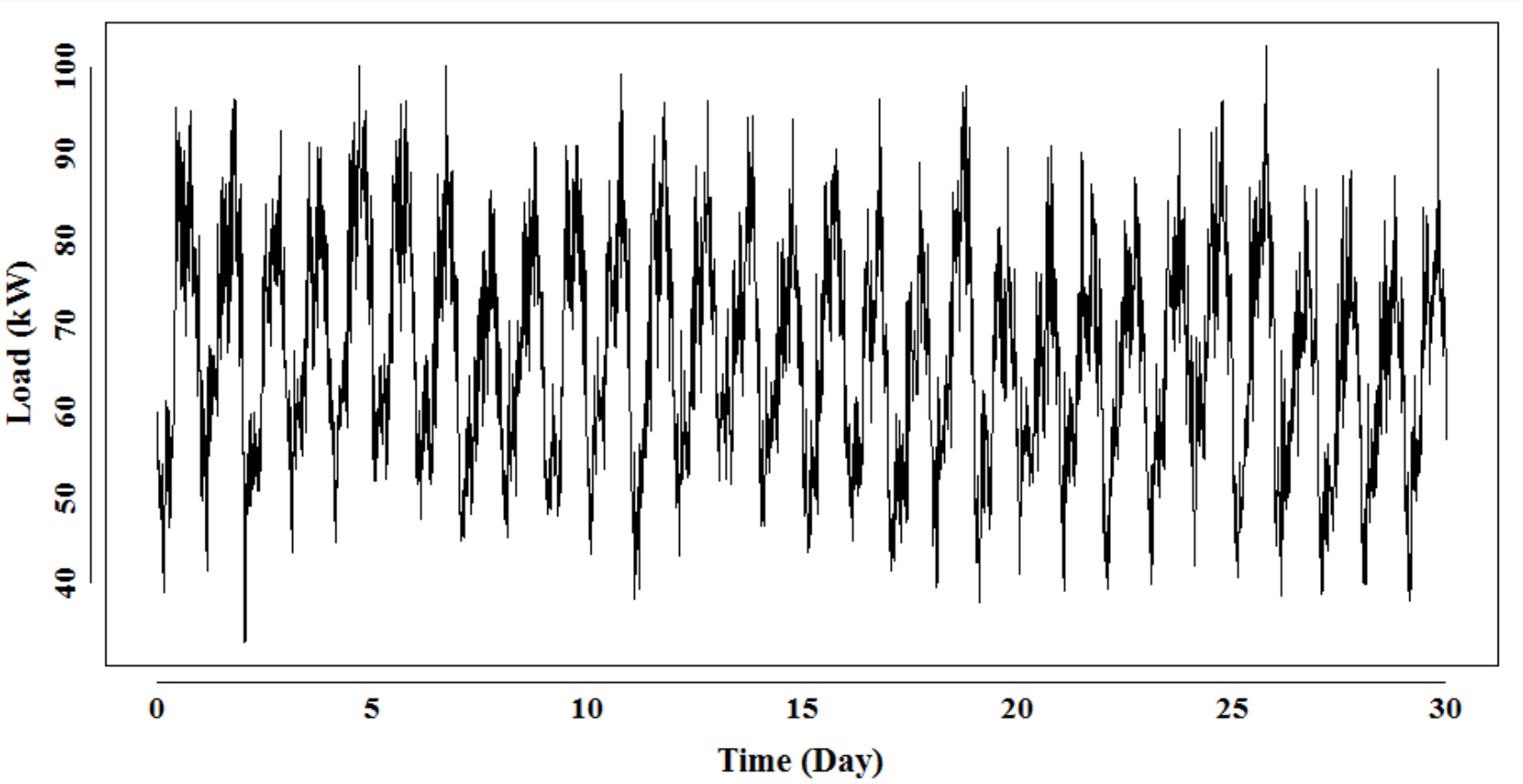
Slovakia

- 11281 consumers. Enterprises.
- 96 measurements per day.
- Test set from one month (September).

Example of consumer from Ireland

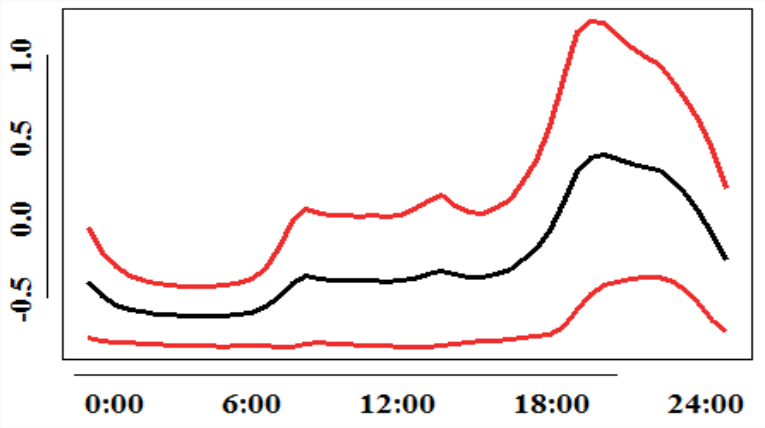


Example of consumer from Slovakia

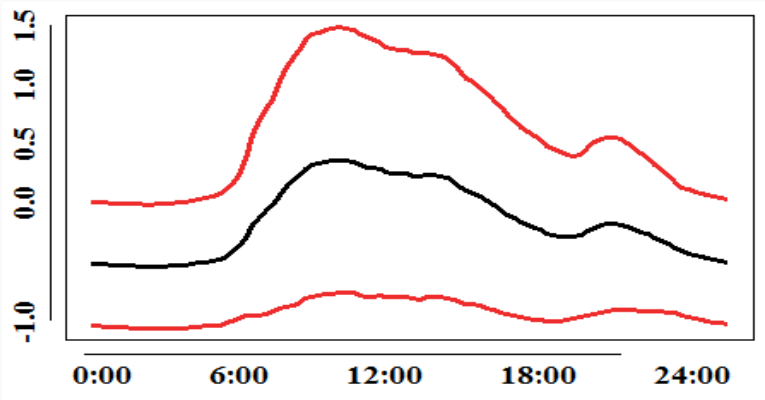


Median daily profiles

Ireland



Slovakia



Median

Median absolute deviation

Proposed approach

Aggregation with clustering

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)
5. Clustering of representations (K-means)

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)
5. Clustering of representations (K-means)
6. Summation of K time series by found clusters

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)
5. Clustering of representations (K-means)
6. Summation of K time series by found clusters
7. Training of K forecast models and the following forecast

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)
5. Clustering of representations (K-means)
6. Summation of K time series by found clusters
7. Training of K forecast models and the following forecast
8. Summation of forecasts

Aggregation with clustering and classification

Proposed approach

Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)
5. Clustering of representations (K-means)
6. Summation of K time series by found clusters
7. Training of K forecast models and the following forecast
8. Summation of forecasts

Aggregation with clustering and classification

1. For new consumers, their representations are computed

Proposed approach

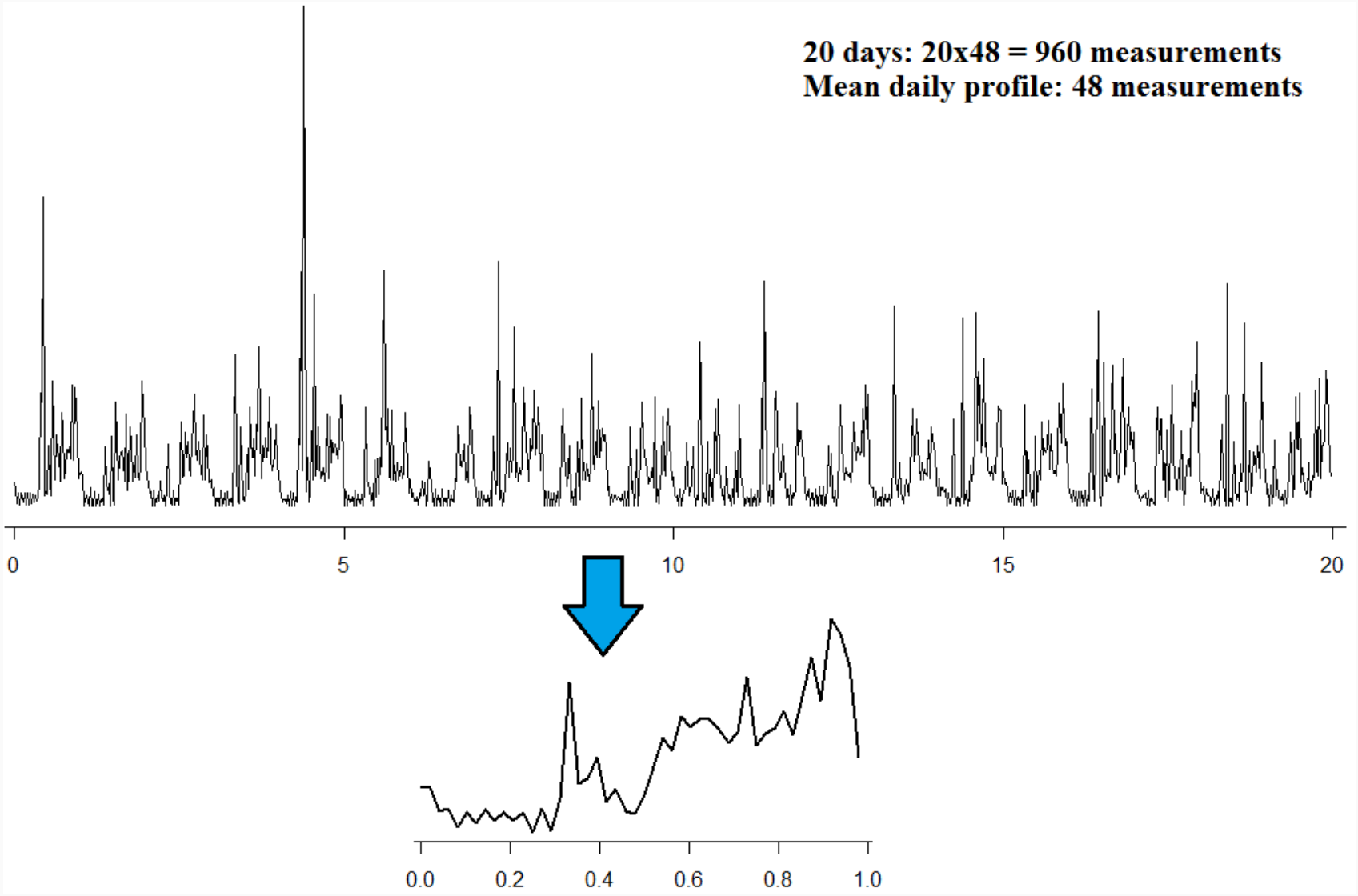
Aggregation with clustering

1. Set of time series of electricity consumption
2. Normalization (z-score)
3. Computation of representations of time series
4. Determination of optimal number of clusters K (DB-index)
5. Clustering of representations (K-means)
6. Summation of K time series by found clusters
7. Training of K forecast models and the following forecast
8. Summation of forecasts

Aggregation with clustering and classification

1. For new consumers, their representations are computed
2. They are assigned to nearest centroids and centroids are updated

Representations of time series



Representations of time series

Let T be a time series of length n, representation of T is a model $\hat{T}$ of reduced dimensionality d ($d \ll n$) such that $\hat{T}$ closely approximates T.

Why time series representations?

Three types of methods of time series representations:

Representations of time series

Let T be a time series of length n, representation of T is a model $\hat{T}$ of reduced dimensionality d ($d \ll n$) such that $\hat{T}$ closely approximates T.

Why time series representations?

1. Reduce memory load.

Three types of methods of time series representations:

Representations of time series

Let T be a time series of length n, representation of T is a model $\hat{T}$ of reduced dimensionality d ($d \ll n$) such that $\hat{T}$ closely approximates T.

Why time series representations?

1. Reduce memory load.
2. Accelerate subsequent machine learning algorithms.

Three types of methods of time series representations:

Representations of time series

Let T be a time series of length n, representation of T is a model $\hat{T}$ of reduced dimensionality d ($d \ll n$) such that $\hat{T}$ closely approximates T.

Why time series representations?

1. Reduce memory load.
2. Accelerate subsequent machine learning algorithms.
3. Implicitly remove noise from the data.

Three types of methods of time series representations:

Representations of time series

Let T be a time series of length n, representation of T is a model $\hat{T}$ of reduced dimensionality d ($d \ll n$) such that $\hat{T}$ closely approximates T.

Why time series representations?

1. Reduce memory load.
2. Accelerate subsequent machine learning algorithms.
3. Implicitly remove noise from the data.
4. Emphasize the essential characteristics of the data.

Three types of methods of time series representations:

Representations of time series

Let T be a time series of length n, representation of T is a model $\hat{T}$ of reduced dimensionality d ($d \ll n$) such that $\hat{T}$ closely approximates T.

Why time series representations?

1. Reduce memory load.
2. Accelerate subsequent machine learning algorithms.
3. Implicitly remove noise from the data.
4. Emphasize the essential characteristics of the data.

Three types of methods of time series representations:

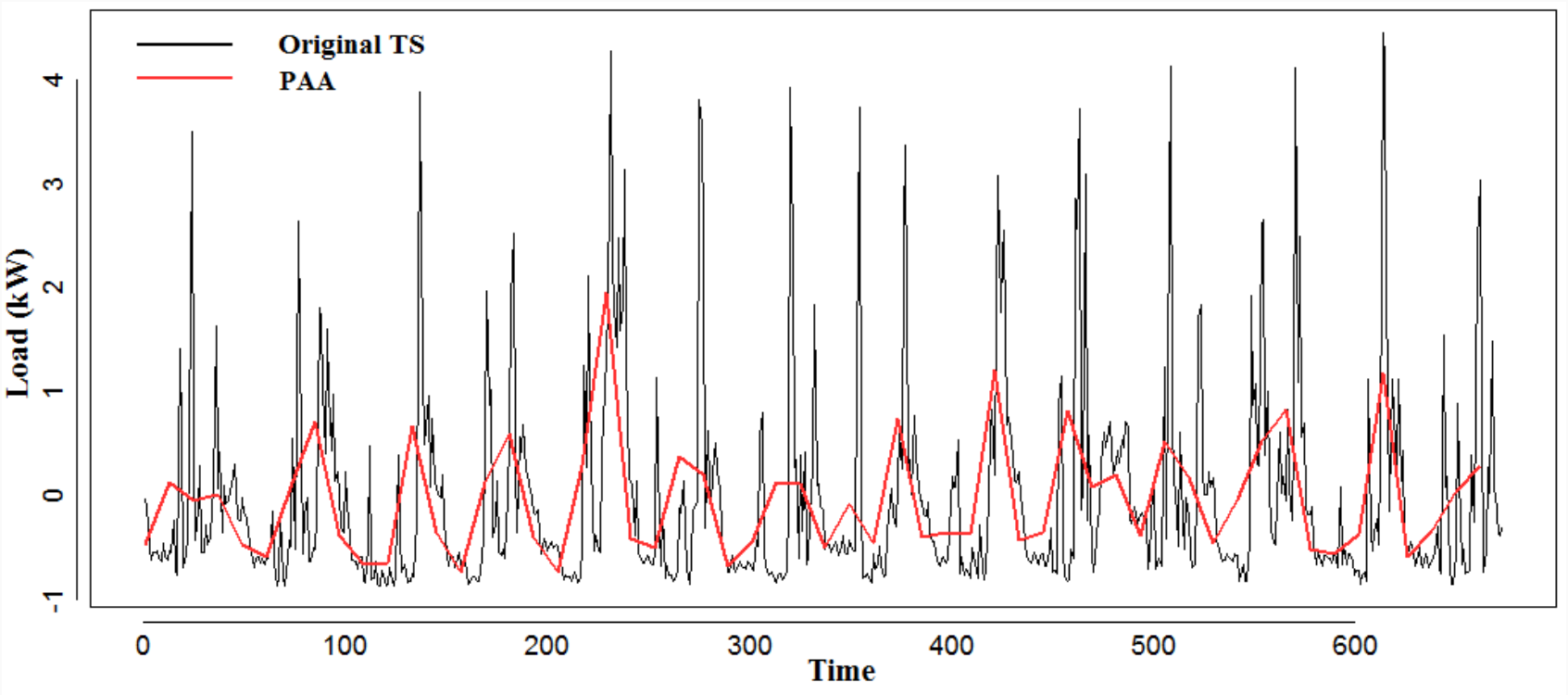
1. Nondata adaptive.
2. Data adaptive.
3. Model based.

Nondata adaptive methods

Whole time series as it is. PAA. Derivations of PAA. Discrete Wavelet Transform (DWT).

PAA - Piecewise Aggregate Approximation.

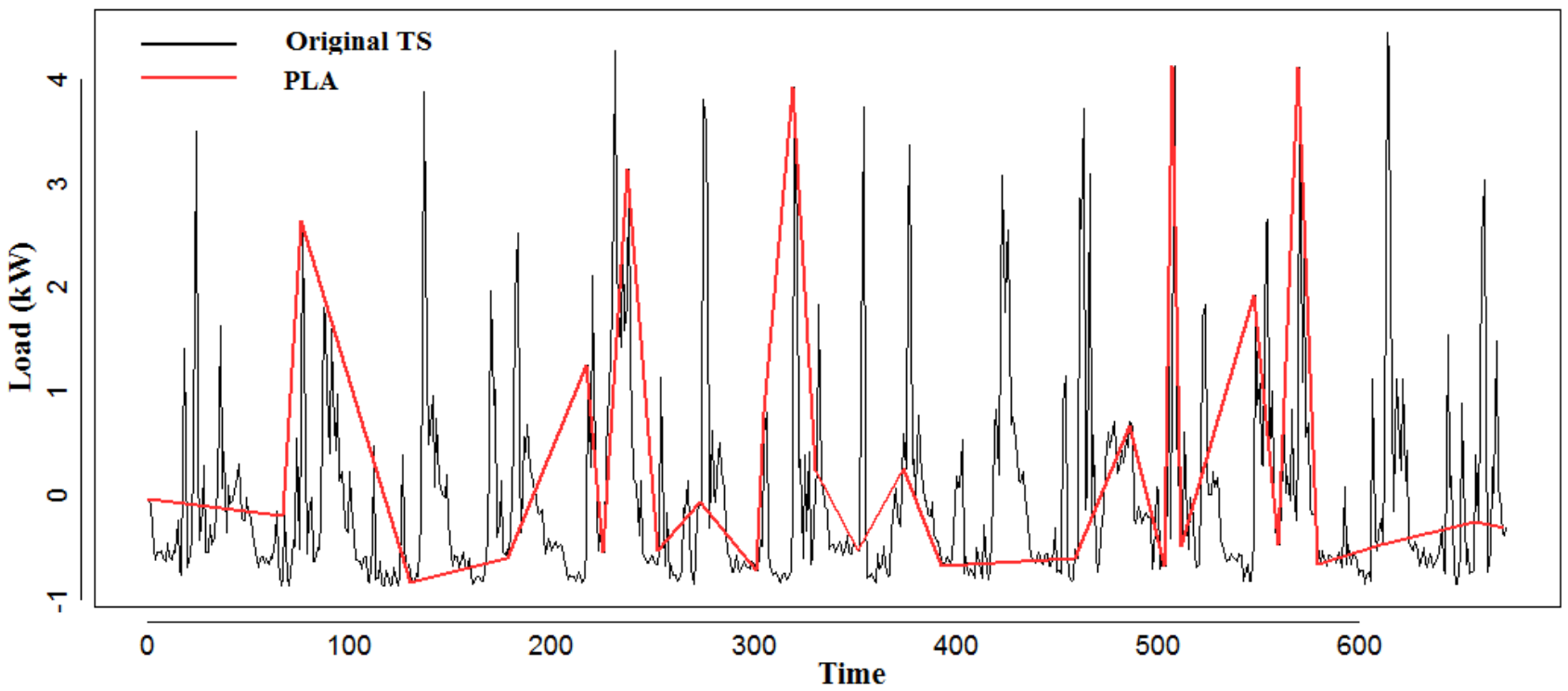
$$\hat{x}_i = \frac{d}{n} \sum_{j=(n/d)(i-1)+1}^{(n/d)i} x_j.$$



We can extract: median, standard deviation, maximum . . .

Data adaptive methods

PLA - Piecewise Linear Approximation.



Model based methods

- Representations based on statistical model.
- Extraction of regression coefficients $\Rightarrow$ creation of daily profiles.
- Creation of representation which is long as frequency of time series (48 respectively 96).

$$x_i = \beta_1 u_{i1} + \beta_2 u_{i2} + \cdots + \beta_{seas} u_{iseas} + \varepsilon_i, \ \textit{where} \ i = 1, \ldots, n$$

New representation: $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{seas})$.

Applied methods:
Multiple Linear Regression. Robust Linear Model. Generalized Additive Model.

Model based methods

- Triple Holt-Winters Exponential Smoothing. Last seasonal coefficients as representation.

    1. Manually set smoothing factors.
    2. Automatically.

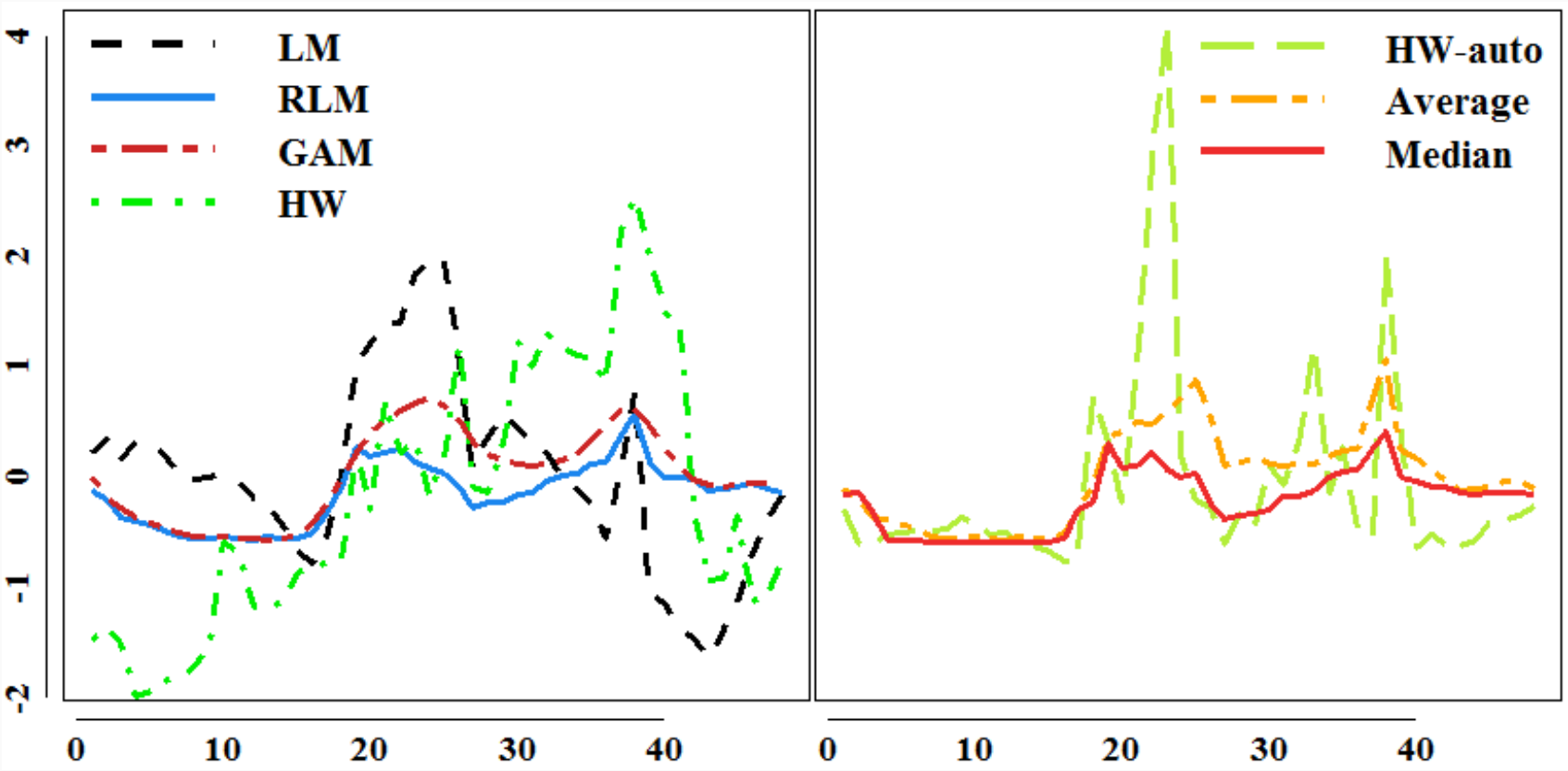
Model based methods

- Triple Holt-Winters Exponential Smoothing. Last seasonal coefficients as representation.

    1. Manually set smoothing factors.
    2. Automatically.

- Mean daily profile.
- Median daily profile.

Comparison of model based representations



LM - Multiple Linear Model, RLM - Robust Linear Model, GAM - Generalized Additive Model, HW - Holt-Winters, HW-auto - Holt-Winters, Average - Mean daily profile, Median - Median daily profile.

Forecast methods

- STL decomposition + Exponential Smoothing
- STL decomposition + ARIMA
- Support Vector Regression

The accuracy of the forecast of electricity consumption was measured by MAPE (Mean Absolute Percentage Error).

$$\text{MAPE} = 100 \times \frac{1}{n} \sum_{t=1}^{n} \frac{|x_t - \overline{x}_t|}{x_t},$$

where x_t is a real consumption, $\overline{x}_t$ is a forecasted load and n is a length of the time series.

Experiments

Random sampling of consumers from the data set.

- For the Irish data 3400 consumers from 3639 were randomly selected, whereby 300 of them have been previously classified.
- For the Slovak data 10800 consumers from 11281 were randomly selected, whereby 300 of them have been previously classified.
- Repeated 100 times and results were averaged.

One day ahead forecast. Sliding window of 10 days. Days are from Tuesday to Friday.

Results

13 representations of time series were compared.

Repres.	Ireland - February			Ireland - September		
	Classif.	Clus.	Agg.	Classif.	Clus.	Agg.
All	4.0488	4.0424	4.5055	5.0342	5.0360	5.4692
PAA	4.0408	4.0482	4.5197	4.9840	4.9907	5.4308
AVE.MAX	4.0649	4.0681	4.5099	5.0861	5.1002	5.4897
MMM	4.1054	4.0986	4.5108	5.1267	5.1404	5.4785
DWT	4.0867	4.0775	4.4903	5.0134	5.0220	5.4283
PLA	3.9248	3.9213	4.5506	4.7555	4.7603	5.4248
LM	4.0215	4.0185	4.5091	5.0542	5.0539	5.4745
RLM	3.9234	3.9176	4.5077	4.9004	4.9053	5.4222
GAM	3.9233	3.9238	4.4961	5.0132	5.0193	5.4716
Average	3.9571	3.9522	4.4888	5.0336	5.0256	5.4721
Median	3.9462	3.9497	4.5237	4.8932	4.8901	5.4160
HW	3.9215	3.9168	4.5208	4.9424	4.9439	5.4817
HW-auto	3.9368	3.9391	4.4927	5.1763	5.1722	5.4706

Results

Repres.	Slovak - September		
	Classif.	Clus.	Agg.
All	3.0145	3.0250	2.8766
PAA	3.0200	3.0206	2.8767
AVE.MAX	2.9779	2.9773	2.8839
MMM	2.9189	2.9255	2.8668
DWT	2.9760	2.9751	2.8638
PLA	2.7365	2.7369	2.8614
LM	2.7319	2.7281	2.8758
RLM	2.7396	2.7405	2.8598
GAM	2.7514	2.7531	2.8930
Average	2.7431	2.7435	2.8829
Median	2.7150	2.7164	2.8541
HW	2.7258	2.7196	2.8768
HW-auto	2.7117	2.7155	2.8718

Conclusion

- Clustering of consumers can **improve** forecast accuracy.
- Classification of new consumers was successful and does not worsen the forecast accuracy.
- We have shown that the best representations in this task are adaptive representations (PLA) and model-based representations, particularly the robust ones (RLM and Median).