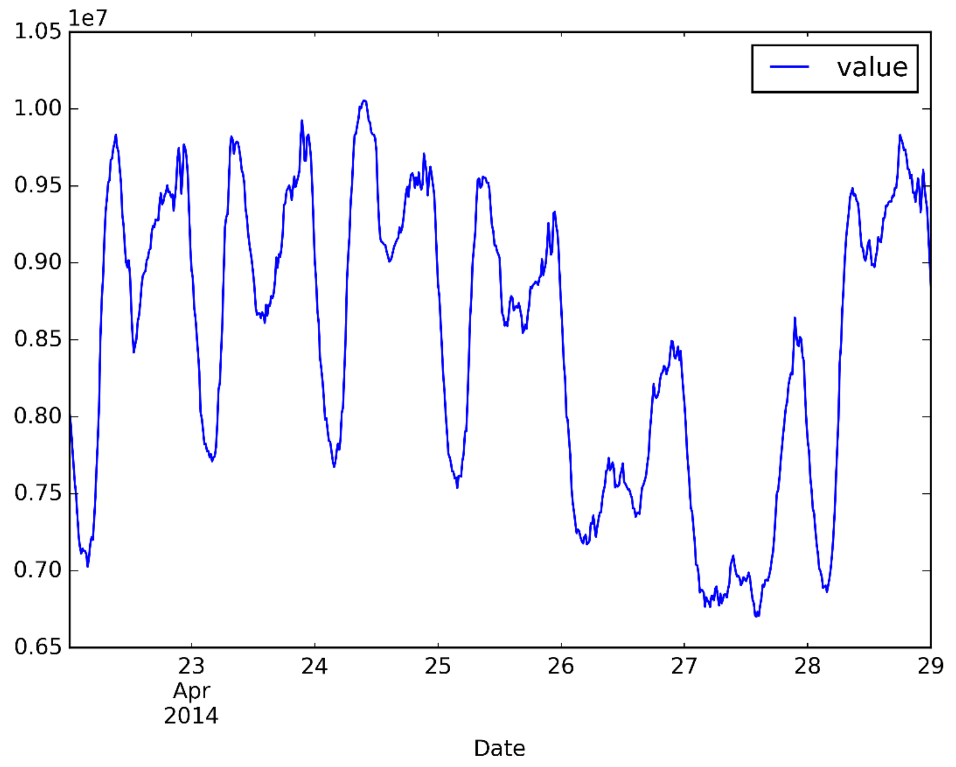
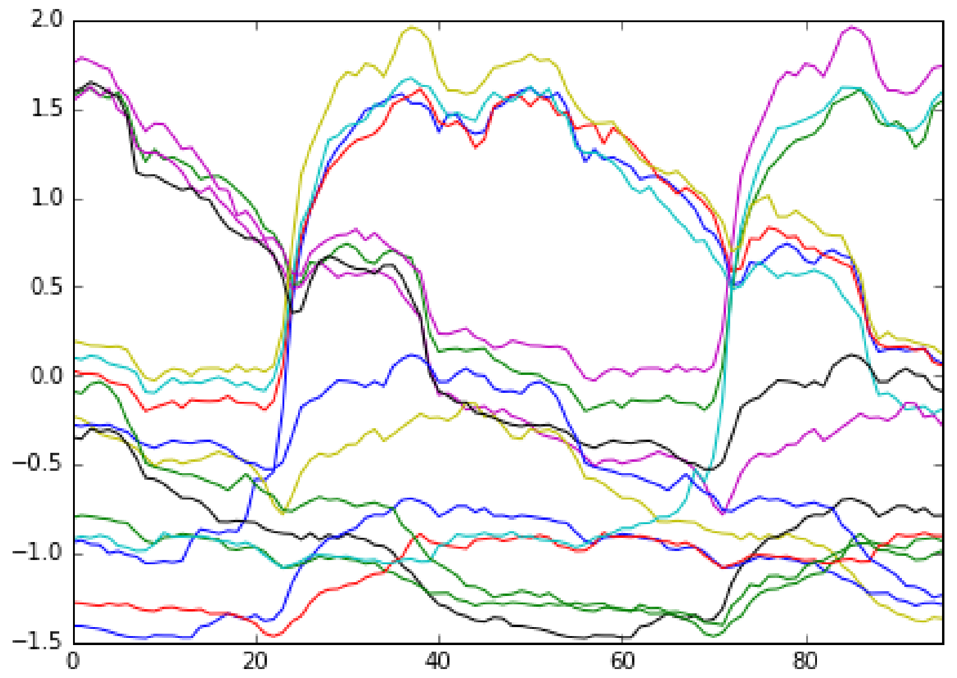
Alphabet Size Reduction for Symbolic Time Series Representation

Author: Jakub Ševcech
Supervisor: Mária Bieliková

Transforming Time Series Into Symbols

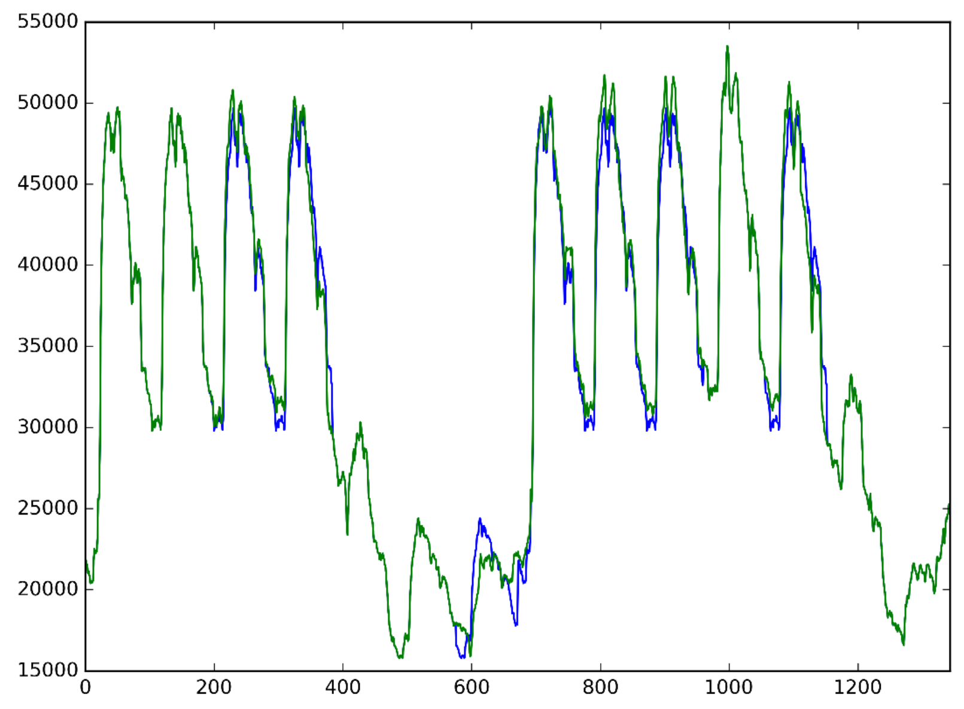
Many seasonal time series are composed of repeating shapes. Electricity consumption data alternates several shapes.



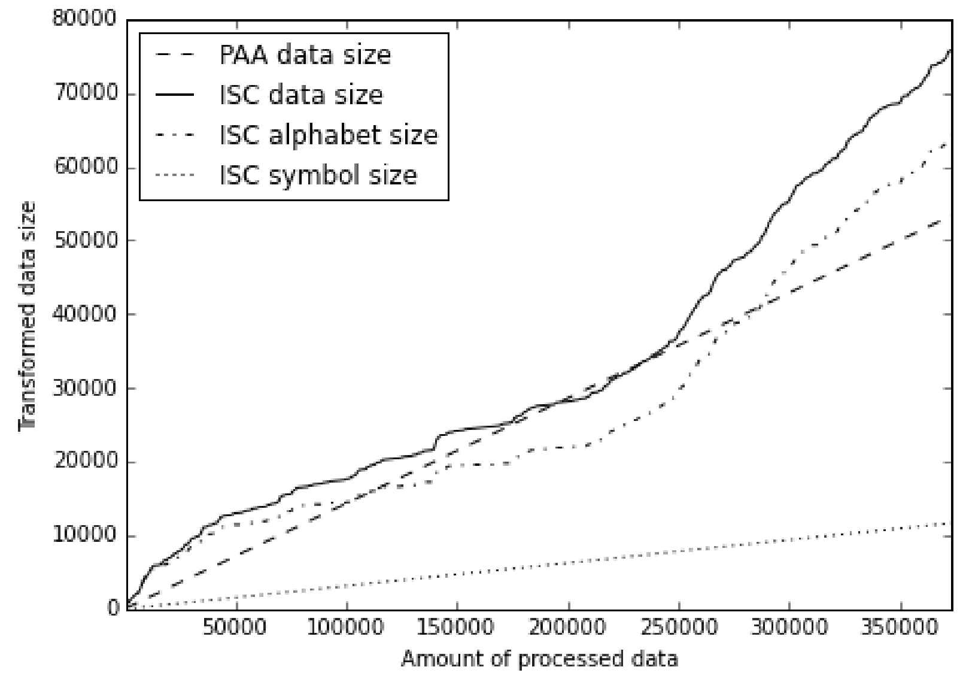


These shapes can represent the time series. We can transform them into an alphabet of symbols and a sequence of identifiers.

Alphabet of symbols and sequence of their identifiers can be used to approximately reconstruct the data.

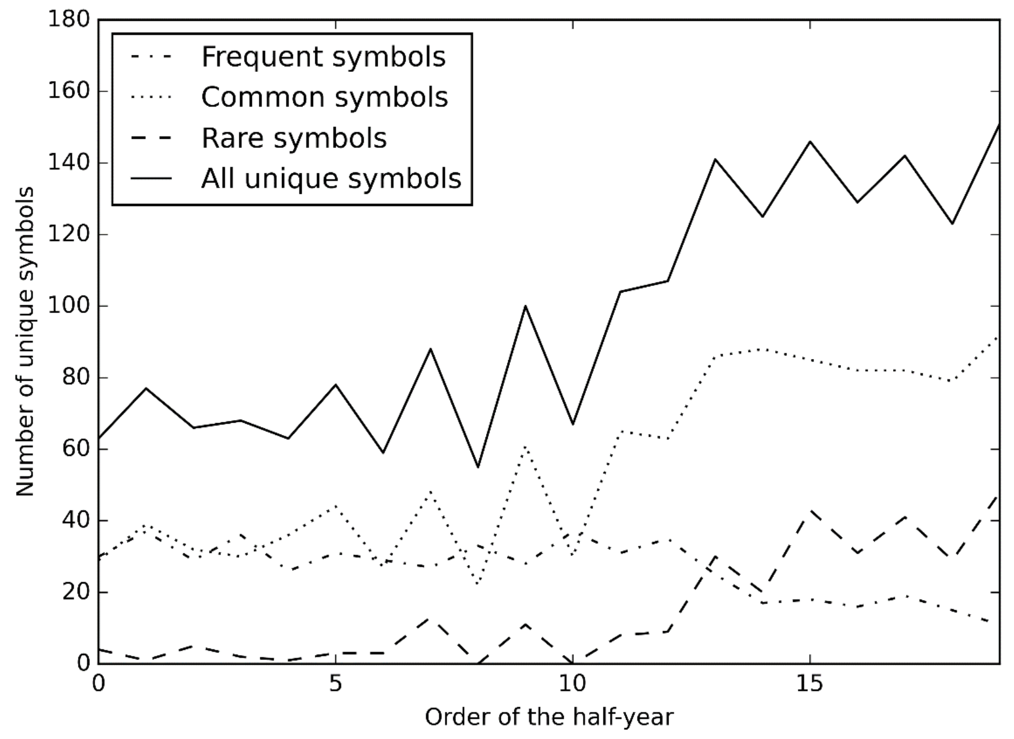


Problem of Ever Growing Alphabet



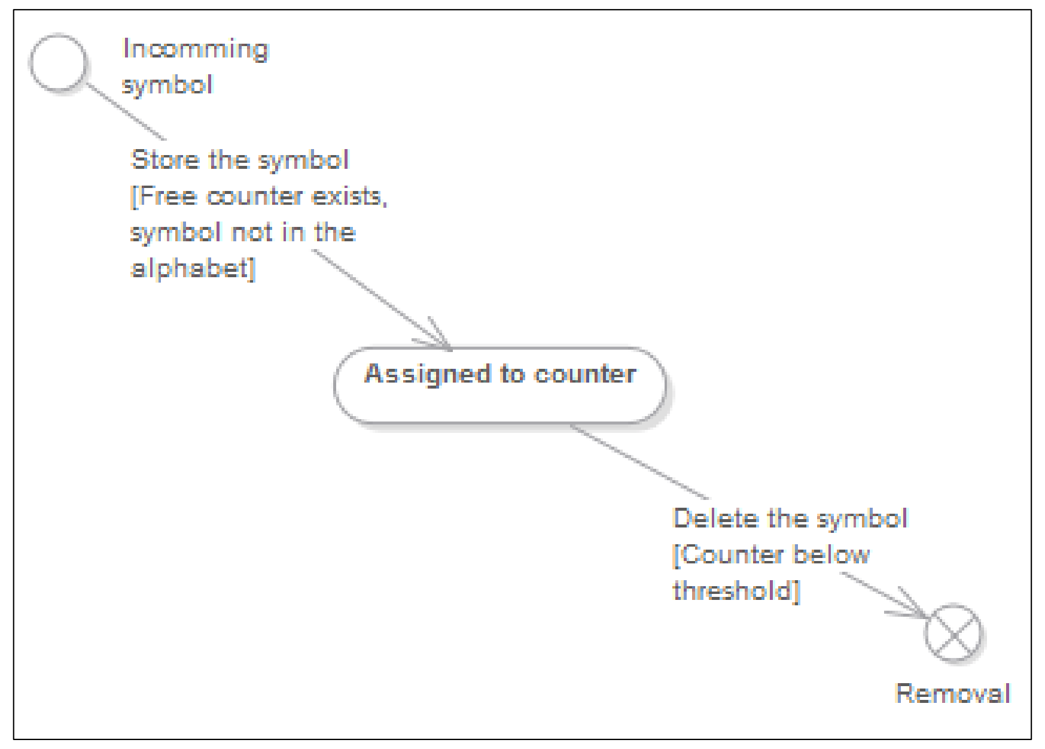
When we transform very long time series, symbols change and the alphabet **grows**. This is **inacceptable** in stream data processing.

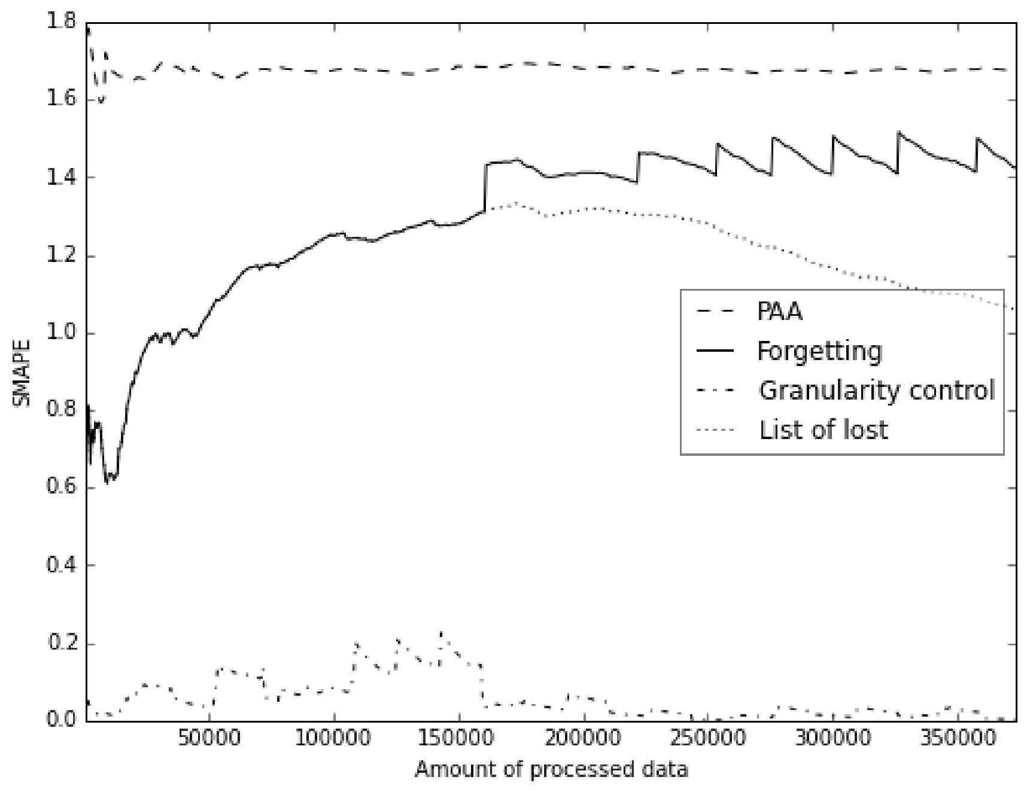
Alphabet Size Reduction by Forgetting



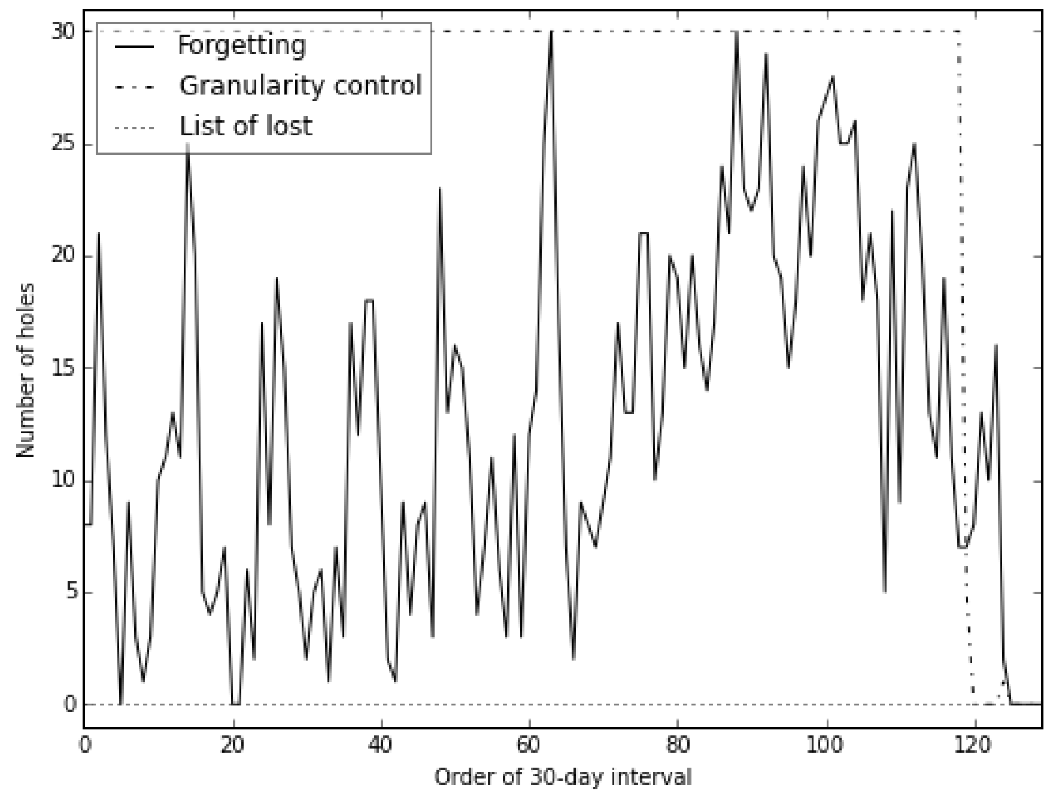
Older symbols are not used in the last sections of the dataset. If we are more interested in recent parts of the time series, we could remove them.

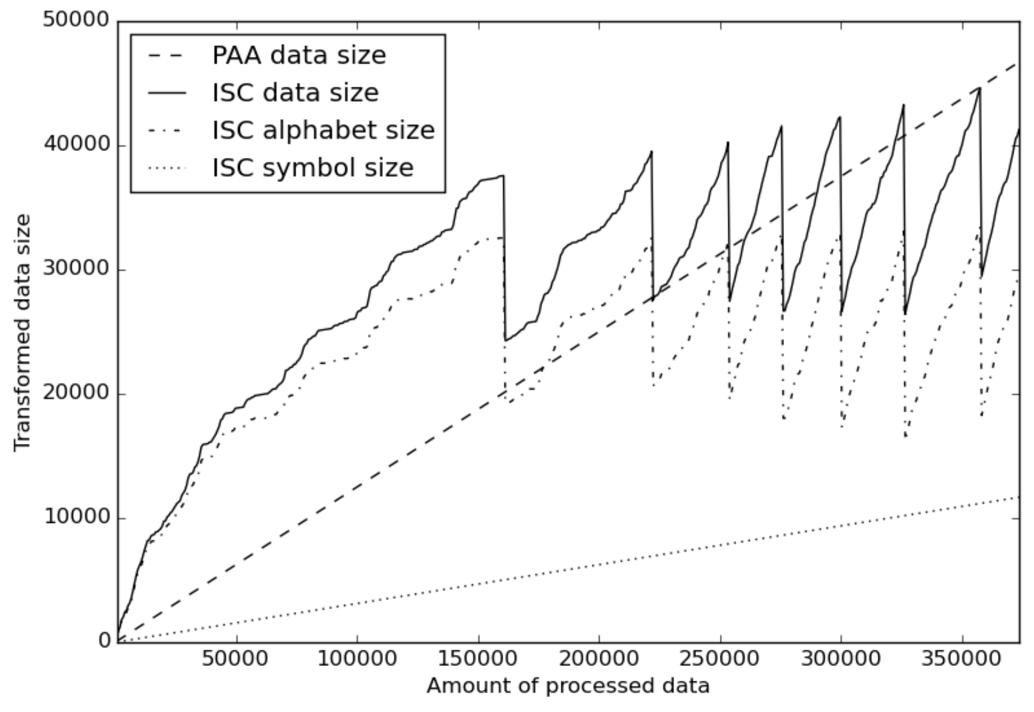
We use symbol forgetting to remove obsolete symbols with 3 improvements: closest symbols returned instead of holes, granularity control and list of lost symbols.


Incomming symbol
Store the symbol [Free counter exists, symbol not in the alphabet]
Assigned to counter
Delete the symbol [Counter below threshold]
Removal



All approaches produced lower reconstruction error than PAA, with different number of holes.





All of them were able to maintain constant alphabet size with the best reconstruction error and number of holes achieved using List of lost improvement. Multiple improvements can be combined for even better results.

email: jakub.sevcech@stuba.sk