

Data Mining Smart Energy Time Series

Janina POPEANGĂ
 University of Economic Studies, Bucharest, Romania
janina.popeanga@yahoo.com

With the advent of smart metering technology the amount of energy data will increase significantly and utilities industry will have to face another big challenge – to find relationships within time-series data and even more - to analyze such huge numbers of time series to find useful patterns and trends with fast or even real-time response.

This study makes a small review of the literature in the field, trying to demonstrate how essential is the application of data mining techniques in the time series to make the best use of this large quantity of data, despite all the difficulties.

Also, the most important Time Series Data Mining techniques are presented, highlighting their applicability in the energy domain.

Keywords: *Time Series Data Mining, Clustering, Classification, Motif Discovery, Data Reduction*

1 Introduction

The increasing deployment of Automated Meter Reading (AMR) systems has created new challenges for utilities industry in terms of how to utilize the recorded data, not only to improve the day-to-day operations. The smart grid will provide a large volume of sensor and meter data that will require intelligent analytics that move further than data management, querying and reporting. To make the best use of this large quantity of data, it is essential to apply data mining techniques to extract relevant information valuable to the utilities industry. [1]

Data mining is defined as a “type of database analysis that attempts to discover useful patterns or relationships in a group of data. The analysis uses advanced statistical methods, such as cluster analysis, and sometimes employs artificial intelligence or neural network techniques. A major goal of data mining is to discover previously unknown relationships among the data, especially when the data come from different databases.” [2]

The paper is structured as follows: Section 2 focuses on the concept of Time-Series Data Mining. Section 3 presents the most important techniques

widely used in this domain. Section 4 makes a review of relevant literature. Section 5 presents some software tools for time-series data mining, while Section 6 ends this article presenting the conclusions.

2. The Time Series Data Mining Concept

Time series data type, also called chronological series or simply time series represent results of measurements made on the characteristics of a unit of population studied, over time, at successive moments of its evolution in some time intervals.

A time series T of size n is an ordered set of n real-value variables, where

$$T = (t_1, t_2, \dots, t_n).$$

A subsequence of length p of time series T is a sampling of length $p < n$ of adjacent positions from T , where

$$T_{i,p} = (t_i, t_{i+1}, \dots, t_{i+p-1}),$$

for $1 \leq i \leq n - p + 1$.

The time intervals for which smart energy measurements are made may include: hours or fractions of hours, days, weeks, decades, months, quarters, semesters, years. Since the intervals are equal and represent the passage of time, the observations resulting from these measurements are successive usually

equidistant in time.

There are two ways to analyze time series data:

- **Time domain analysis** which studies how a time series process evolves through time;
- **Frequency domain analysis** (spectral analysis) which studies how periodic components at different frequencies illustrate the evolution of a time series.

According to Hui et. al., Data Mining “*is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses, or other information repositories.*” The most used data mining functions in commercial and research analysis are association, classification, prediction and clustering.

These data mining techniques have also been applied to other types of data such as time-series, telecommunications, web, spatial, and multimedia data. [3]

Applying the principles and techniques of classical data mining in the time series analysis had led to the concept called **Time Series Data Mining**.

Energy data (production and consumption) recorded over a period of time at fixed intervals is a classic time series data mining problem.

The steps taken in the entire process are:

- collect data from multiple sources: web, text, databases, data warehouses, sensors, smart devices;
- data filtering by eliminating errors. When using a data warehouse, this process is removed because a process of extraction, transformation and loading (ETL) was already applied on the data;
- establishing key data attributes that will participate in the DM process, by selecting those properties that interest the

analysis;

- application of templates and detection/analysis of new knowledge;
- visualization, validation and evaluation of results.

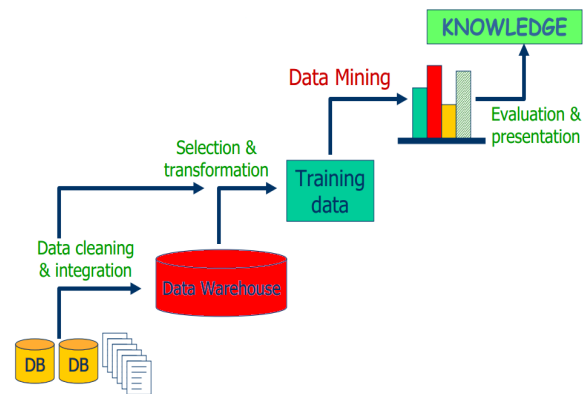


Fig. 1. Data Mining process

With the increasing deployment of a large number of sensors and other smart devices, the amount of time-series data is growing rapidly, often in the order of gigabytes per day or even per minute. The big challenge is to find correlation relationships within time-series data and even more - to analyze such huge numbers of time series to find similar or regular patterns and trends with fast or even real-time response.

3. Time Series Data Mining Techniques

Classification and clustering have the effect of dividing the objects in some classes, but the difference is how this division is made. The classification classes are predetermined, and at clustering they are built taking into account the characteristics of the analyzed objects.

The *time series clustering technique* can be divided into three main categories:

- *whole series clustering* – the purpose is to regroup complete time series into clusters so that the time series are as similar to each other as possible within each cluster;
- *subsequence clustering* – the purpose is to find similarity among different time subsequences and to group them in the same cluster. It is used

for discovering structures or patterns in time series data;

- *time point clustering* – the purpose is to find clusters of the time point. It is similar to time series segmentation, however, time point clustering is different from segmentation by the fact that not all points must be assigned to the cluster, some of points being considered noise.

The *time series classification technique* can be divided into two types of classification:

- *time series classification* – the purpose is to mapped complete time series into predefined classes. It is used to match a whole sequence to a class of other sequences;
- *subsequence classification* – the purpose is to check if a subsequent belongs to a certain sequence.

Association finds some rules and models in data analyzed and sequence analysis seeks some models on some data ordered in time, trying to determine a way to order these models, assuming that between events there are certain implications.

According to the number of involved time series, the *association rules technique* can be divided into two categories:

- *association rules mining from single series* – the purpose is to find rules and sequential patterns;
- *association rules from multiple series* – it is divided into two types:
 - *intra-transactional association rules mining* – the purpose is to make known the co-relations of multiple time series at same time. Taking the renewable energy production as an example, we can find rules like “If wind speed goes up and sun intensity raises then the renewable energy production

increases on the same day”;

- *inter-transactional association rules* – the purpose is to make known the co-relations of different series at different time. Taking the energy consumption and renewable energy production as an example, we can find rules like “If the number of solar panels installed increases on the first day and the number of hours of sunlight increases on the second day then the energy consumption from the electricity common network will decrease on the third day”.

Regression treats a general trend that, if carried out in time, can be used at forecast, while the **forecast** function takes into account other factors, including cyclical events etc. Many time series prediction application involve regression analysis, for example, predicting future energy consumption based on historical data and other information like outside temperature, house characteristics, living habits, number of person hours spent at home, electrical appliances and devices, number of electric heaters in the home, etc.

Exception analysis may be related to clustering and classification, which may find exceptions either observing that certain groups are composed of a single object very different from the other (grouping) or when an object does not fall into any class (classification).

The purpose of detecting anomalies in a time series is to find abnormal subsequences in that series, which means to find subsequences that do not follow the model of a series normal behavior.

Another data mining technique, available for time series is **motif discovery**. The purpose of this task is to find every sequence that appears repeatedly in a time series. The sequence can be known from the beginning or not. Given a sequence as pattern, this technique performs a search to find other sequences that are similar with the pattern, but the search for unknown motifs is a more complex problem because all subsequences of all possible lengths have to be compared.

In my opinion, data mining techniques can be used in energy management to solve, for example, problems such as:

- **Classification** – determining consumer profiles based on different variables, determining the possibility of purchase and install special equipment for renewable energy generation, based on user profile;
- **Regression** – time trend analysis of the energy consumption and production, monitoring the effect of energy policies and measures;
- **Forecast** – predicting future energy production and consumption;
- **Anomaly detection** – consumer fraud detection, network intrusion and other unusual and rare events that are hard to find;
- **Motif discovery** – identify energy relationships that can aid in the process of forecasting, identify patterns that can be used to predict customers behavior.
- **Association rules** – analyzing the links between certain factors that could cause increased/decreased energy consumption or production;
- **Clustering** – locating fraud or high energy consumption, getting a group of them around certain areas.

Contrasting the traditional techniques for the time series analysis and limiting assumptions, the methods in the Time Series Data Mining network can be successfully applied to identify the complex characteristics, and to predict the non-periodic, non-linear, irregular and chaotic time series.

4. Research Trends and Issues

Data Mining techniques have been used by many researchers in energy system applications.

Tso and Yau have used three data mining techniques (regression analysis, decision

tree and neural networks) for understanding energy consumption patterns and predicting electricity consumption. [4]

Azadeh et al. have proposed an integrated fuzzy system, data mining and time series framework to estimate and predict electricity demand for seasonal and monthly changes in electricity consumption especially in developing countries such as China and Iran with non-stationary data. [5]

Kusiak et. al. have applied eight data-mining algorithms to model the nonlinear relationship among energy consumption, control settings (supply air temperature and supply air static pressure), and a set of uncontrollable parameters, for minimization of the energy to air condition a typical office-type facility. [6]

Yu et al. have demonstrated that the use of decision tree method can classify and accurately predict building energy demand levels, using real data from Japanese residential buildings. The competitive advantage of decision tree method over other broadly used modelling techniques, such as regression method and ANN method, lies in the ability to generate accurate predictive models with interpretable flowchart-like tree structures that enable users to quickly extract useful information. [7]

Figueiredo et. al. have presented an electricity consumer characterization framework based on DM techniques. Using real data of consumers from the Portuguese distribution company, the load profiling module created a set of consumer classes using a clustering operation and the representative load profiles for each class. The classification module builded a classification model able to assign different consumers to the existing classes. [8]

Dent et. al. have described the bottom up and clustering methods for defining representative load profiles for domestic electricity users in the UK. Investigation of electricity consumption in order to determine similarities between types of consumers required that the day's usage pattern to be summarized so it can be

compared with others. [9]

Time Series Data Mining has been an ever-growing and stimulating field of study that has continuously raised challenges and research issues over the past years.

As one of the most prominent issues arises from the high dimensionality of time series data, researchers have proposed methods of how to reduce the size of the data without substantial loss of information.

Data reduction is often the first step to combating massive time series data, as it will provide a summary of the data.

As Miles et. al. [10] clarify, “data reduction is not something separate from analysis. It is part of analysis. The researcher’s decisions—which data chunks to code and which to pull out, which evolving story to tell—are all analytic choices. Data reduction is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that final conclusions can be drawn and verified”.

Discrete Fourier Transform, Singular Value Decomposition, Discrete Wavelet Transform and Random Projection (Sketches) are the most frequently used dimensionality reduction techniques.

Discrete Fourier Transform, first introduced by Agrawal et. al., is used to transform time series into frequency domain and selects the most important coefficients.

Giving the fact that energy storage can be a practical solution to balance energy production against its consumption, Makarov et. al. have proposed to use Discrete Fourier Transform to decompose the required balancing power into different time-varying periodic components, i.e., intra-week, intra-day, intra-hour, and real-time. This approach was used in a study conducted for the 2030 Western Electricity Coordinating Council (WECC) system model and was considered a success. [11]

Xiao et. al. have proposed to use Discrete

Fourier Transform (DFT) for coordinated sizing of energy storage and diesel generators in an isolated microgrid. The DFT-based coordinated dispatch strategy allocates balance power between the two components through frequency-time domain transform. [12]

Saenthon et. al. have demonstrated the effectiveness and high accuracy of the Discrete Fourier Transform to automatically identify load type of electrical appliances by transforming time domain to frequency domain. [13]

Discrete Wavelet Transform is used to approximate the data by dividing the sequences into equal-length sections and saving the weighted value of these sections.

Chang has used the Wavelet Transform (WT) of the time-frequency domain to analyze and detect the transient physical behavior of loads during the load identification. According to his research the discrete wavelet transform (DWT) is more suitable than short-time Fourier transform for transient load analyses. [14]

Nieto-Hidalgo et. al. have demonstrated that the Wavelet Transform could be used to identify simpler electrical consumption patterns as a part of total consumption curve. They have proposed an innovative method based on WT to decompose the global power consumption in elemental loads corresponding to each appliance. [15]

Abu-Shikhah and Elkarmi have proposed the using of the SVD (singular value decomposition) technique in a new combination which uses hourly loads of successive years to predict hourly loads and peak load for the next selected time span. After filtering out the load trend of the Jordanian power system, they have applied the SCV technique to de-noise the resulting signal. [16]

Dieb Martins and Gurjao have applied the random projection technique in order to obtain a reduced version (sketch) of smart meters' original data, thus increasing the processing throughput of the utility. Using real smart meters measurements, they have demonstrated that processing using sketches

sized 50% smaller than original data can achieve a 2% average relative error while presenting greater data rates. [17]

5. Time Series Data Mining Tools

In recent years, Oracle offered more and better statistical and time series analysis capabilities through *Oracle Data Mining*. ODM provides several kinds of data-

mining algorithms for functions such as clustering, classification, regression, anomaly detection, association rules, sequence similarity etc.

Working directly with ODM to analyze time series data stored in Oracle databases or data warehouses, can reduce resources consumption and the time needed for data transmission.

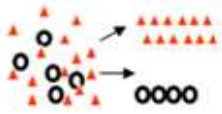

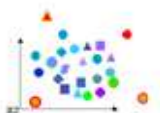
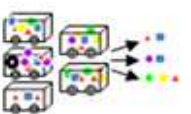
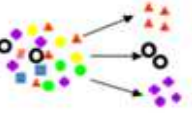
Function	Algorithm
<p>Classification</p> 	<p>Logistic Regression</p> <p>Decision Trees</p> <p>Naïve Bayes</p> <p>Support Vector Machine</p>
<p>Regression</p> 	<p>Multiple Regression</p> <p>Support Vector Machine (SVM)</p>
<p>Anomaly Detection</p> 	<p>One Class SVM</p>
<p>Association Rules</p> 	<p>Apriori</p>
<p>Clustering</p> 	<p>K-Means</p> <p>Hierarchical O-Cluster</p>

Fig.2. SQL Data Mining Algorithms

Microsoft SQL Server Analysis Services (SSAS) contains online analytical processing (OLAP) and data mining functionality for business intelligence applications.

SSAS provides data mining techniques like: classification, regression, clustering,

association algorithms, sequence analysis etc.

Mining historical data using SSAS can offer new visions and create a basis for forecasting, and this may be extremely interesting for analysis of time series energy data.

The *Microsoft Time Series* algorithm offers

regression algorithms that are optimized to forecast continuous values, such as energy production or consumption, over time.

The Time Series Data Mining nodes in *SAS Enterprise Miner* (Time Series Data Preparation Node, Time Series Similarity Node, and Time Series Exponential Smoothing Node) significantly improve the time series analysis and data preparation capabilities of the data miner. Finding time series that unveil similar statistical characteristics permits identifying customer behaviors in large volumes of time series energy data.

Also, with the large volumes of energy consumption and production data stored in time series, the power to integrate this data into analysis workflows will help utilities to build valuable models more easily.

R is a free software with statistical and graphical capabilities which runs on a wide variety of UNIX, Windows and MAC OS platforms. *R* provides time-series analysis and data mining techniques such as: TS forecasting, TS clustering, classification and association rules etc. The most popular models used to predict energy consumption with *R* based on known past data are: autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA).

Weka's time series framework proposes a data mining approach to model time series by transforming the data into a form that can be processed by standard data mining algorithms. The solution is to remove the temporal ordering of individual input examples by encoding the time dependency via additional input fields. These fields are sometimes referred to as "lagged" variables. Various other fields are also computed automatically to allow the algorithms to model trends and seasonality. After the data has been transformed, any method capable of predicting a continuous target can be applied. This approach to time

series analysis and forecasting is often more powerful and more flexible than classical statistical techniques such as ARMA and ARIMA. [18]

Handling time series analysis in a tool like *RapidMiner* requires advanced skills, but this did not prevent its success. *RapidMiner* has over 3 million total downloads and has over 200,000 users including eBay, Intel, PepsiCo and Kraft Foods as paying customers.

According to Bloor Research, *RapidMiner* provides 99% of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors.

RapidMiner's approach to time series is based on two main data transformation processes [19]:

- Windowing to transform the time series data into a generic data set
- Applying any of the data mining algorithms to predict the target variable and thus predict the next time step in the series.

For the second year in a row, Gartner Research has placed *RapidMiner* in the Leaders Quadrant in the Magic Quadrant for Advanced Analytics Platforms, describing this tool as a "*platform that supports an extensive breadth and depth of functionality, and with that it comes quite close to the market Leaders.*" [20]

6. Conclusions

The concept behind this paper, Time Series Data Mining, can be defined as the process of discovering useful patterns and significant structures, unknown associations and relationships, anomalies and motifs, which can be used to predict future events and behaviors.

We have made a small review of the most important techniques, highlighting our proposal of their applicability in the energy domain.

As can be seen from the research trends presented, Time Series Data Mining methods have been applied successfully in a wide range of energy system applications and for issues arisen from the high

dimensionality of time series data, have been proposed various methods of how to reduce the size of the data without substantial loss of information.

Acknowledgment

This work was cofinanced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/142115

„Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”

References

- [1] J. Popeangă, *Building Consumer Profiles through Time-Series Data Mining Techniques*, Abstracts Book, 4th World Conference on Business, Economics and Management (BEM-2015), 30 April - 02 May 2015.
- [2] Sunita, Prachi, *Efficient Cloud Mining Using RBAC (Role Based Access Control) Concept*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013, page 679.
- [3] Hui, S. C., Jha, G., *Data mining for customer service support*, Information & Management, Volume 38, Issue 1, 2000, pp. 1-13.
- [4] G. K. F. Tso, K. K. W. Yau, *Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks*, Energy 32, 2007, 1761–1768.
- [5] A. Azadeh, M. Saberi, S. F. Ghaderi, A. Gitiforouz, V. Ebrahimipour, *Improved estimation of electricity demand function by integration of fuzzy system and data mining approach*, Energy Conversion and Management 49 (2008) 2165–2177.
- [6] A. Kusiak, M. Li, F. Tang, *Modeling and optimization of HVAC energy consumption*, Applied Energy 87, 2010, 3092–3102.
- [7] Z. Yua, F. Haghghata, B.C.M. Fungb, H. Yoshinoc, *A decision tree method for building energy demand modelling*, Energy and Buildings 42 (2010) 1637–1646.
- [8] V. Figueiredo, F. Rodrigues, J. G. Gouveia, *An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques*, IEEE Transactions on Power Systems, 20(2005) 596-602.
- [9] I. Dent, U. Aickelin and T. Rodden, *The application of a data mining framework to energy usage profiling in domestic residences using UK data*, Proceedings of the Research Students' Conference on “Buildings Don't Use Energy, People Do?” – Domestic Energy Use and CO2 Emissions in Existing Dwellings 28 June 2011, Bath, UK.
- [10] M. B. Miles and M. A. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook* (2nd ed.), Thousand Oaks, Calif.: Sage, 1994.
- [11] Y. V. Makarov, M. C.W. Kintner-Meyer, P. Du, C. Jin, and H. F. Illian, *Sizing Energy Storage to Accommodate High Penetration of Variable Energy Resources*, IEEE Transactions on Sustainable Energy, (Volume:3, Issue: 1), ISSN :1949-3029, 2011.
- [12] J. Xiao, L. Bai, F. Li, H. Liang, C. Wang, *Sizing of Energy Storage and Diesel Generators in an Isolated Microgrid Using Discrete Fourier Transform (DFT)*, IEEE Transactions on Sustainable Energy, ISSN :1949-3029, 2014.
- [13] A. Saenphon and S. Kaitwanidvilai, *Load Identification in Household Apparatus Equipment using Discrete Fourier Transform with Proper Window Function*, Proc. Power and Energy Systems (AsiaPES 2013), track 800-144, 2013.
- [14] H. Chang, *Non-Intrusive Demand Monitoring and Load Identification for Energy Management Systems Based on*

- Transient Feature Analyses, Energies*, 2012, vol. 5, issue 11, pages 4569-4589.
- [15] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, J. M. García-Chamizo, V. Romacho-Agud and F. Flórez-Revuelta, *Using Wavelet Transform to Disaggregate Electrical Power Consumption into the Major End-Uses*, Paper presented at the 8th International Conference on Ubiquitous Computing & Ambient Intelligence UCAmI 2014.
- [16] N. Abu-Shikhah and F. Elkarmi, *Medium-term electric load forecasting using singular value decomposition*, *Energy*, Volume 36, Issue 7, July 2011, Pages 4259-4271.
- [17] A. D. Martins and E.C. Gurjao, *Processing of smart meters data based on random projections*, 2013 IEEE PES Conference On Innovative Smart Grid Technologies Latin America (ISGT LA), ISBN: 978-1-4673-5272-7, Sao Paulo, 2013.
- [18] Pentaho, *Time Series Analysis and Forecasting with Weka*, 2015, Link: <http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka>
- [19] Simafore, *Using RapidMiner for time series forecasting in cost modeling: 1 of 2*, 2012, Link: <http://www.simafore.com/blog/bid/106430/Using-RapidMiner-for-time-series-forecasting-in-cost-modeling-1-of-2>
- [20] RapidMiner, *RapidMiner: Leader in Gartner Research Magic Quadrant for Advanced Analytics Platforms*, Gartner, 2015



Janina POPEANGĂ graduated in 2010 the Faculty of Cybernetics, Statistics and Economic Informatics, Economic Informatics specialization. The title of her Bachelor's thesis is "*Distributed Databases*". In 2012, she graduated the Databases for Business Support master program with the thesis "*Monitoring and management of electric power consumption using sensorial data*". Janina's interests are broadly in the fields of databases and distributed systems. Since 2012 she is a Ph.D. Student in the Doctoral School of Bucharest Academy of Economic Studies. Her research focuses on real-time database systems,

business intelligence analytics, sensor data management, smart grid and renewable energy.