NEWSLETTER

May 2022

MORE

Management of Real-time Energy Data

Dear reader,

I welcome you to the first newsletter of the MORE: Management of Real-Time Energy DATA. MORE is a Research and Innovation project funded by the European Commission under the topic "ICT-51-2020 - Big Data technologies and extreme-scale analytics " (Grant Agreement: 957345). MORE aims at strengthening the leadership of the EU in two of the most important industries: big data management and renewable energy production. MORE creates the techniques and tools that unlock the potential of the data produced by wind and solar parks. These tools will allow to increase the efficiency of renewable energy sources and will contribute to a greener environment. MORE states as its main objective to *"allow stakeholders in industry sectors with huge volumes of sensor data, especially the Renewable Energy industry, to: a) scale the management of streaming and historical time series beyond an order of magnitude beyond the state-of-art and b) to perform forecasting, prediction and prediction and diagnostics using the whole data that is available to them with accuracy that outperforms existing approaches."*

The work and progress in the first half of the project has been exciting! Scientific and industry partners worked closely together to understand and document the problems and challenges in managing data from wind and solar parks, and created well defined requirements and technical specifications for the MORE platform. We designed a data processing architecture that combines edge and cloud computing allows computationally cheap decisions to be performed on the edge, and complex analysis to take place in the cloud. Following this architecture the MORE platform will be able to take advantage of all the data that are produced in the edge, and provide scalable processing in an unpreceded scale. A key idea for increasing scalability in MORE is to employ lossy compression on the data at the edge, before transmitting them to the cloud. The data volume is greatly reduced, while the compression impact remains low. Accurate prediction, forecasting and diagnostics are enabled by novel Incremental Machine Learning algorithms and by identifying patterns in the monitoring data that are associated with important conditions in the renewable energy parks (e.g. malfunctions).

In this newsletter, we are happy to bring you some of the most interesting outcomes of the project. ENGIE Laborelec explains the potential gains to the wind energy production by having the suitable data analysis tools, and the University of Aalborg investigates the impact of compressing monitoring data from the wind and solar parks. IBM presents the Streams and Incremental Learning (SAIL) library created in the context of MORE, and Athena Research details the method for detecting soiling of solar parks.

In the following months we will keep you updated with additional newsletters, but if you want frequent updates please visit our website or follow us in Twitter (@MOREAnalytic) and LinkedIn (https://www.linkedin.com/company/more_h2020_project/). We are looking forward for your feedback and your participation in the dissemination events that will take place in the next period!

Yours sincerely,
Manolis Terrovitis
Principal Researcher, Research Center Athena
Coordinator of the MORE project

Investigating The Effect of Data Compression on Time Series Forecasting

Carlos Enrique Muniz Cuza, Jonas Brusokas, Søren Kejser Jensen, Kaixuan Chen, Nguyen Thi Thao Ho and Torben Bach Pedersen Aalborg University

Lossy compression methods effectively reduce the storage footprint on the edge devices, e.g., wind turbines, and the amount of data sent through the network to the operators. Particular types of lossy compression for time series data can restrain the decompression error to a specific error bound. For example, an error bound of 10% means that any point of the time series after decompression will be within a 10% difference from the original point. This opens the door to many opportunities in the RES industry as operators can control the error bound at any specific point. However, there is one important gap limiting its broad adoption. Using lossy compression can introduce undesired systematic distortions after decompression (e.g., over-smoothing), affecting the reliability of the data analytics. At the same time, reliable and precise prediction of relevant target variables like wind power is crucial for the dispatch, unit commitment, and stable functioning of power systems. In this context, applying lossy compression without understanding its impact on future predictions is impossible. Our research aims to accelerate the broad application of lossy compression in the RES industry by developing a profound understanding of its effects on the time series forecasting analytics.

More specifically, we want to answer the following three questions:

- What is the expected accuracy lost (concerning the baseline) as the operator increases the error bound of the lossy compression algorithm? Answering this question is essential as the operator of the RES park will be tempted to increase the error bound to increase the compression further, thus reducing the storage. Then, expecting how much accuracy will be lost as this error bound increases is of extreme relevance. The baseline is the best-performing forecasting model trained and tested on the original time series.
- What is the best lossy compression algorithm? Usually, the only optimization variable of lossy compression algorithms is the compression ratio, i.e., how much they can compress. However, in our use case, we also need to consider the accuracy of the forecasting analytics as an optimization variable. Then, understanding which compression algorithm provides the best trade-off between forecasting accuracy and compression ratio is crucial.
- What is the most resilient forecasting model? There are no extended experiments in the literature on measuring which forecasting model is more resilient to systematic distortions of the data like those inserted during the reconstruction after compression. Benchmarks and comparisons between forecasting models usually are performed on the original time series. Thus, pointing out the most resilient forecasting model from the beginning can avoid operational costs.

These three questions have not been answered in the literature and are extremely important to accelerate the application of lossy compression in the RES industry. We are performing a series of experiments combining multiple state-of-the-art forecasting models and multiple time series lossy compression algorithms in different datasets related to the RES domain. Taking as a baseline the accuracy results of the forecasting models on the original time series, we are compressing the data and calculating the difference in the accuracy of the models using the reconstructed time series for multiple error bounds. To generalize our findings, we are performing thousands of experiments exploring different training and testing scenarios, as well as different random initialization of the model's parameters. This project will contribute to the RES industry and have a significant impact on the scientific community, promoting the development of new lossy compression algorithms and different use-cases in various industry sectors.

Incremental models for time-series data

Seshu Tirupathi and Dhaval Vinodbhai Salwala
IBM

Wind turbines generate optimal power output when the turbine blades are perpendicular to the wind direction. However, due to technical errors or malfunctioned sensors, this is not always the case - when a wind turbine does not face the wind, it is defined as yaw misalignment. Detecting and correcting yaw misalignment directly impacts the bottom line of wind farm companies. In the past, physical models, machine learning models, and hybrids of these two approaches have been used to detect yaw misalignment. However, it has been identified as a difficult challenge to solve with 5-10 minute aggregated data, which is the usual frequency at which data is stored in wind farms. Data at higher velocity/frequency is now being analyzed to detect yaw misalignment.

Similarly, very short-term forecasting for energy bidding also requires data at a high frequency. There is a growing demand for machine learning algorithms on time-series data across various domains like finance, inventory management, etc. where high-velocity or high-volume data needs to be analyzed. Further complications, like pre-processing noisy streaming data, memory constraints of models that need to be deployed on the edge, and AutoML models that need to be constantly updated due to concept drifts, arise in the incoming data and target variables.

We have developed the **Streams and Incremental Learning (SAIL)** library as part of the MORE (www.more2020.eu) project to address some of these issues. SAIL includes a standard set of incremental machine learning algorithms and wrapper functionality to existing libraries (River, Scikit-Multiflow, Scikit-Learn, Skorch) with standard Scikit-Learn APIs for incremental models through the partial_fit function. Ensemble learning for incremental models and a combination of batch and incremental models are part of SAIL. Elementary distributed computing functionality for model selection in SAIL has also been set up with Ray.

Data-driven soiling detection

Ioannis Psarros, Alexandros Kalimeris , Giorgos Giannopoulos,
Giorgos Papastefanatos and Manolis Terrovitis
Athena RC

Soiling is the accumulation of dirt on the surface of the solar panels, which results in a gradual loss of solar energy production. To reduce the effect of soiling, the panels must be cleaned on strategically chosen dates to reduce the cost induced by energy loss while considering cleaning costs. The problem is quite challenging because soiling monitoring systems are often unreliable or costly. Ideally, decisions should be taken based only on measurements of a minimal number of variables, which are considered reliable, e.g., power output, irradiance, module temperature, and precipitation.

Therefore, accurately quantifying soiling loss is an important step toward mitigating the financial loss caused by the underperformance of solar panels due to soiling. We have developed a framework that operates on a time series defined by a minimal set of variables measured by the sensors of a solar park and extracts models that accurately predict losses due to soiling. These models can be used in historical data to detect periods during which soiling severely affected power production. Detecting such periods can assist the parks' administrators in efficiently planning future cleanings of the park. Moreover, these models can be used in a real-time scenario where the park administrators can detect that the production loss is severe and a manual cleaning must be scheduled immediately.

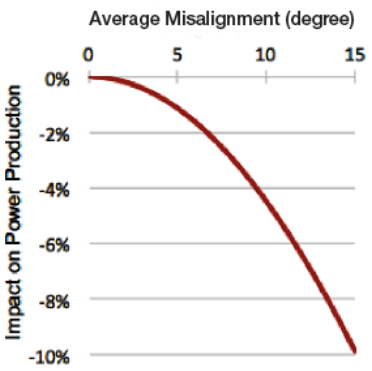
Our approach works in two steps: first, we detect cleaning events in the input time series, i.e., classify rains or manual cleanings that effectively wash the solar panels. Then, using these cleaning events, we define periods that represent the "clean" performance of solar panels, and we use them for training regression models capturing the power yield of clean solar panels. The main advantages of our approach are that it does not require labeled data, it is not based on the accuracy of an analytical formula for the optimal energy output of the park, and it is agnostic to the type of solar panels employed. It requires no knowledge about the model-specific parameters and the specifications associated with the solar panels. As a purely data-driven approach, it solely depends on the availability of data and a minimal set of generally available variables. Moreover, the proposed method exhibits multifold functionality; it identifies cleaning events, models the real expected power output, i.e., power output when solar panels are clean, and computes a ranked list of periods with severe soiling losses. Our experiments indicate that our models are more accurate in predicting "clean" power output than standard generic analytical formulas, which constitute the main alternative modeling approach agnostic to the type of solar panels.

To evaluate our method, we have used a publicly available dataset that contains a set of current-voltage (I-V) curves and associated meteorological data for solar panels in three different locations and climates for approximately one-year periods. For each location, we are given values for a normalized metric, called soiling derate, which compares the daily performance of a solar panel to an identical solar panel that is cleaned during daily maintenance. We compare soiling derate with the corresponding performance index induced by our methods and a performance index that is based on an analytical formula, as employed by the stochastic rate and recovery (SRR) method, which is a state-of-the-art method for quantifying soiling loss. Our experiments show that our method achieves higher accuracy in predicting daily losses due to soiling than the other alternative.

Improving energy production of wind farms through data analysis

Julien Masson, Nicolas Girard
ENGIE Laborelec

Large-scale renewable power generation is gaining momentum with megawatt-size renewable assets at industrial sites. However, power generation market dynamics put pressure on wind farm operators to raise the bar. They need to improve asset performance and maximize power output while keeping operating costs as low as possible and ensuring the grid's stability.



One of the issues in operating wind farms is yaw misalignment issues. Wind turbines are designed to face the wind to harvest as much energy as possible. When the wind changes direction, the wind turbine nacelle changes direction to face the wind. The turbine does not immediately follow the wind direction when it changes. A temporary yaw misalignment is created during a brief moment that could last some minutes. When a yaw misalignment is constant (static), the wind turbine (WT) continuously sub performs. It is advised to correct WT when the static yaw misalignment is higher than about 4 degrees.

Energy losses due to yaw misalignment can be significant, but the effective correction will boost output. A laser device, LIDAR, can measure wind speed and direction between 80m and 400m upstream of the wind turbine and determine how accurately the turbine is aligning itself to the incoming wind. The wind vane can be adjusted with this information, and turbine settings can be modified to increase performance. Gains in annual energy production (AEP) can be as high as 5%. However, a Lidar installation is expensive, and it is recommended to remotely detect suspicious misaligned wind turbines to target LIDAR installation in interesting cases. That is where data sciences and engineering merge to better design smart algorithms and data treatments to detect yaw misaligned wind turbines.

ENGIE Laborelec is a leading expertise and research center in electrical power technology. Drawing on the skills of about 335 specialized engineers and technicians, the company is active on the whole electricity value chain, focusing on the energy transition and net-zero carbon owning about 10GW of wind installed capacity. ENGIE is providing MORE partners with high-frequency data of hundreds of wind turbines related to use cases to be investigated. ENGIE already analyzed data pertaining to these use cases but not using high-frequency data and with limited variables taken into account. MORE's partners can analyze the high-frequency data with the idea to apply state-of-the-art machine learning algorithms to find innovative solutions addressing these use cases. Some very interesting outputs were already discussed, and further accurate results are expected.

Webinar: IT, big data, and machine learning challenges in time-series data

IBM - Athena Research Center

A dedicated MORE webinar was organized on March 14, 2022, called “IT, big data and machine learning challenges in time series data - Renewable Energy Sources (RES) sector | Management of Real-time Energy Data”. The bi-annual IBM-Athena colloquium series addresses the industry challenges due to the exponential growth of time-series data. The colloquium series strives to cover the challenges arising from high frequency and/or high-volume time series data in various sectors like RES, water, inventory management, etc. The first colloquium aims to understand the challenges in handling big data and machine learning algorithms in the renewable energy sources sector. Traditionally, data in RES has been aggregated over 5–10-minute intervals and business use cases were built on this aggregated data for the RES sector. However, with the lowering the cost of sensors and communications, and increasing demand for high-frequency updates and use cases, there is an exponential growth in the data generated by the devices. There are natural challenges of persistence and analytics on this data. Privacy and security add an additional layer of complexity. The colloquium will cover the use cases that arise from high-frequency data and the technical challenges to handle this data and provide analytics on top. The online event takes took place between 1 PM - 3:30 PM GMT on March 14, 2022. Speakers and panelists are international industry and academic experts that would provide a holistic viewpoint on the future of computing in the RES sector.”. The list of talks included: The Software- and AI-Driven Future of Renewables (Shivkumar Kalyanaraman), Easy, Accurate, and Fast Complex Analytics on Big Data Series Collections: Renewable Energy Sources and Beyond (Themis Palpanas), Optimizing operation and maintenance of renewable energy assets: some machine learning challenges (Paul Poncet).

Newsletter editing by Danae Pla - Karidi

MORE

Management of Real-time Energy Data

Thanks for subscribing to our newsletter.
Our newsletter is intended to inform our subscribers about our overall performance trajectory.

We'll keep you posted every six months on the latest news on MORE's achievements, results, trends, and interesting news.

For more information, visit:
www.more2020.eu

MORE receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 957345.