Large Synthetic Datasets for Machine Learning Applications in Power Transmission Grids

Marc Gillioz

Hes·so VALAIS WALLIS

April 15, 2025

Working Group – Foundation Models for the Electric Grid

The Team

Institute of Energy and Environment
University of Applied Sciences and Arts of Western Switzerland (Sion)

- Prof. **Philippe Jacquod**
  Head of Energy Efficiency Group

- Dr. **Marc Gillioz**
  Senior scientist

- **Guillaume Dubuis**
  Master's student

Motivation

- Power grids are under pressure → modeling needed
- Why transmission grids?
  - meshed and complex power flow, with universal features → interesting
  - limited losses at ultra-high voltage → DC power flow approx. valid
- armasuisse Cyber-Defence Campus mandate
  - detection of false data injection attacks

Outline

- **The problem**
  realistic time series for large grid models
- **Our solution**
  load modeling and optimal dispatch
- **An application**
  ML algorithms for anomaly detection


www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

A large synthetic dataset for machine learning applications in power transmission grids

Marc Gillioz, Guillaume Dubuis & Philippe Jacquod

With the ongoing energy transition, power grids are evolving fast. They operate more and more often close to their technical limit, under more and more volatile conditions. Fast, essentially real-time computational approaches to evaluate their operational safety, stability and reliability are therefore highly desirable. Machine Learning methods have been advocated to solve this challenge, however they are heavy consumers of training and testing data, while historical operational data for real-world power grids are hard if not impossible to access. This manuscript presents a large synthetic dataset of power injections in an electric transmission grid model of continental Europe, and describes the algorithm developed for its generation. The method allows one to generate arbitrarily large time series from the knowledge of the grid – the admittance of its lines as well as the location, type and capacity of its power generators – and aggregated power consumption data, such as the national load data given by ENTSO-E. The obtained dataset are statistically validated against real-world data.

Background & Summary

Electric power grids are fast-evolving, with the massive integration of new renewable energy resources – either distributed over lower voltage levels, or as large offshore wind farms, multi-GW solar photovoltaic parks or GW battery energy storage parks – and with the accompanying necessary grid upgrades. Incorporating these new resources into the legacy grid poses significant challenges to grid operators. Power grids operate more often closer to their technical limit, while undergoing larger, more frequent fluctuations. Evaluating the safety and reliability of operation needs to be done more often – optimally in real-time – and with better accuracy. This is becoming computationally challenging. Fast, data-based methods are often advocated as solutions to this problem. In particular, Machine Learning (ML) approaches are of interest, because they learn efficiently and provide results on fast timescales[1–3]. ML algorithms are heavy consumers of training and testing data, however, and the need for large, high-quality sets of data has been identified as one of the key challenges to overcome, when constructing ML algorithm ready for real-world application[4].

There is thus a clear need for open-access, large sets of operational data of electric power grids for ML applications. Historical operational data are however notoriously hard, if not impossible, to obtain. For this reason, most studies rely on synthetic datasets that span the entire feasible space of grid configurations and focus on representative operational conditions by appropriate sampling[5–11]. In all such cases, the state of the grid is realistic at any given data point, but the historical structure and correlations between power injections at different times are lost. On the other hand, time series analyses have been proven powerful, but they are generally limited to small systems because of the lack of good-quality data[12]. In fact, real-life time series exist in aggregated form, e.g. with time-resolved power generation and consumption at national level. However, geographical resolution at the level of individual buses is highly desirable and actually necessary for detailed power flow analysis. Existing databases with such resolution are based on synthetic data[13–15]. They are helpful for specific power flow investigations such as operational studies, generation and transmission expansion studies or grid extension planning. They contain however too many correlations between time series of power injections at different buses to be of any help for ML applications. As a matter of fact, training and testing a ML algorithm on artificially correlated ground truth data inevitably leads to biases that will hinder the successful deployment of the algorithm in real-world situations.

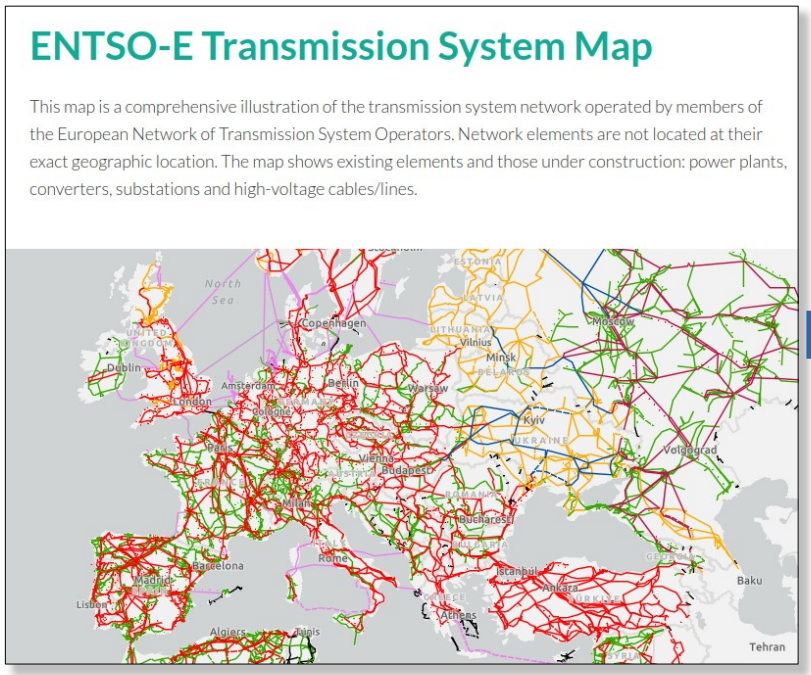
[1] School of Engineering, University of Applied Sciences and Arts of Western Switzerland HES-SO, 1951, Sion, Switzerland. [2] Department of Quantum Matter Physics, University of Geneva, CH-1211, Geneva, Switzerland. e-mail: marc.gillioz@hevs.ch

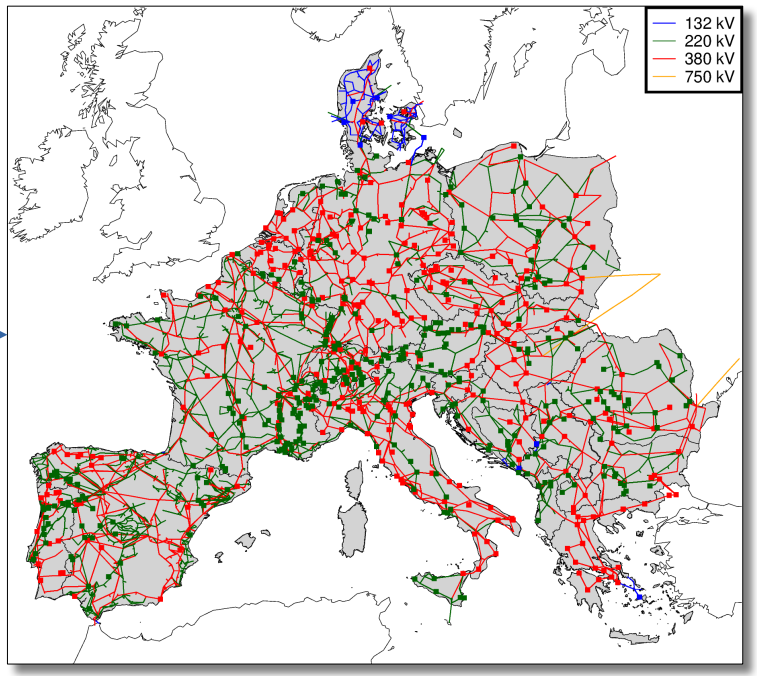
SCIENTIFIC DATA | (2025) 12:168 | https://doi.org/10.1038/s41597-025-04479-x 1

coming soon...

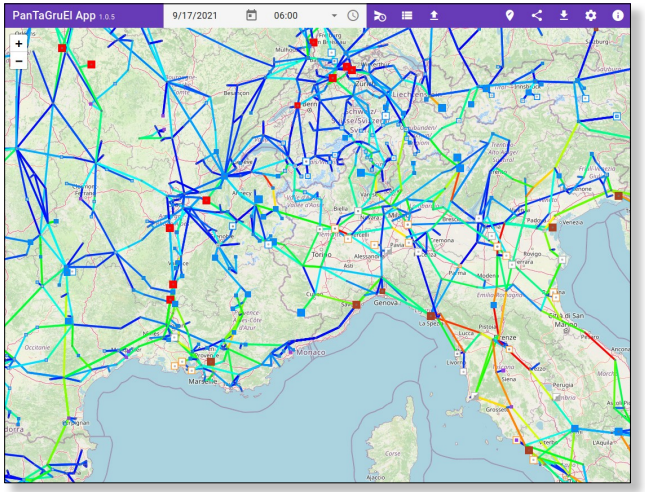
The problem: transmission grid models

- Limited access to real-world data. But…

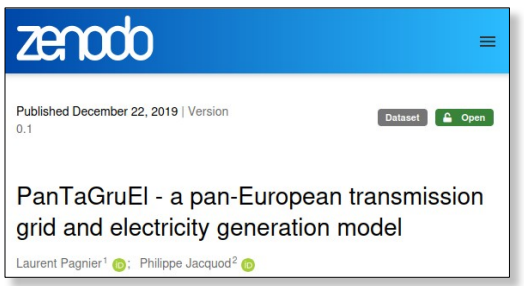




PanTaGruEl



- A Pan-European Transmission Grid and Electricity generation model
  - Interactive version at https://etranselec.ch/pantafrontend/
  - L. Pagnier, P. Jacquod, "*Inertia location and slow network modes determine disturbance propagation in large-scale power grids*"
  - M. Tyloo, L. Pagnier, P. Jacquod, "*The Key Player Problem in Complex Oscillator Networks and Electric Power Grids: Resistance Centralities Identify Local Vulnerabilities*"

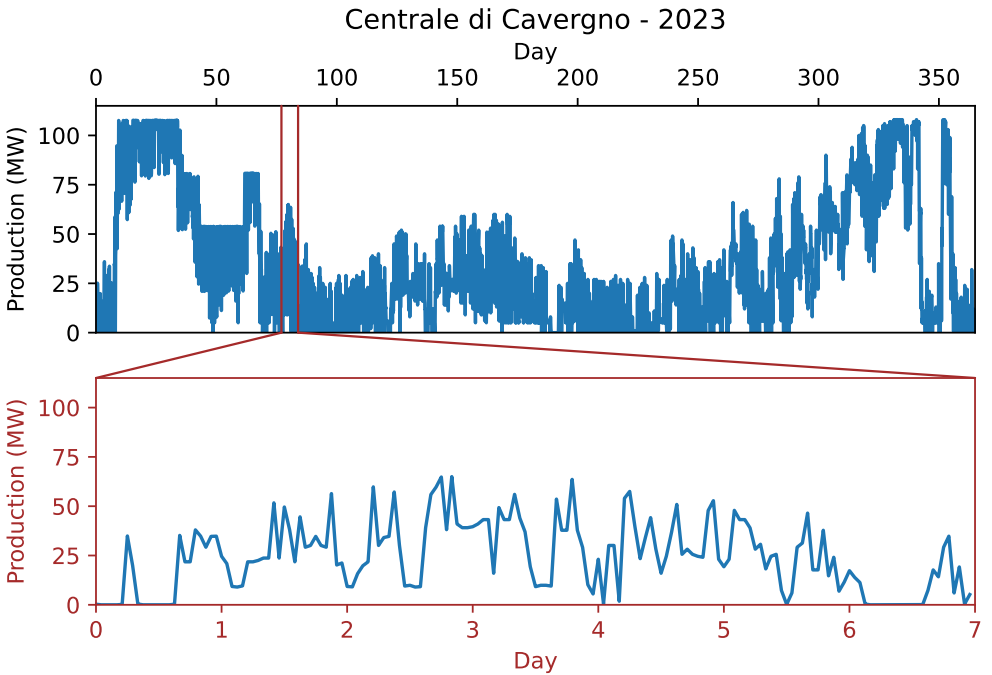


- Our version: PowerModels format
  - 7822 power lines and 553 transformers
  - 4097 buses with load distributed based on population density
  - 815 generators of various types

The other problem: time series

- Steady-state data, one hour time resolution: some data available, but not enough for ML...
- Generating synthetic data is very challenging!
- How do we model *this*?

(data source: https://transparency.entsoe.eu/)



Typical synthetic data approaches

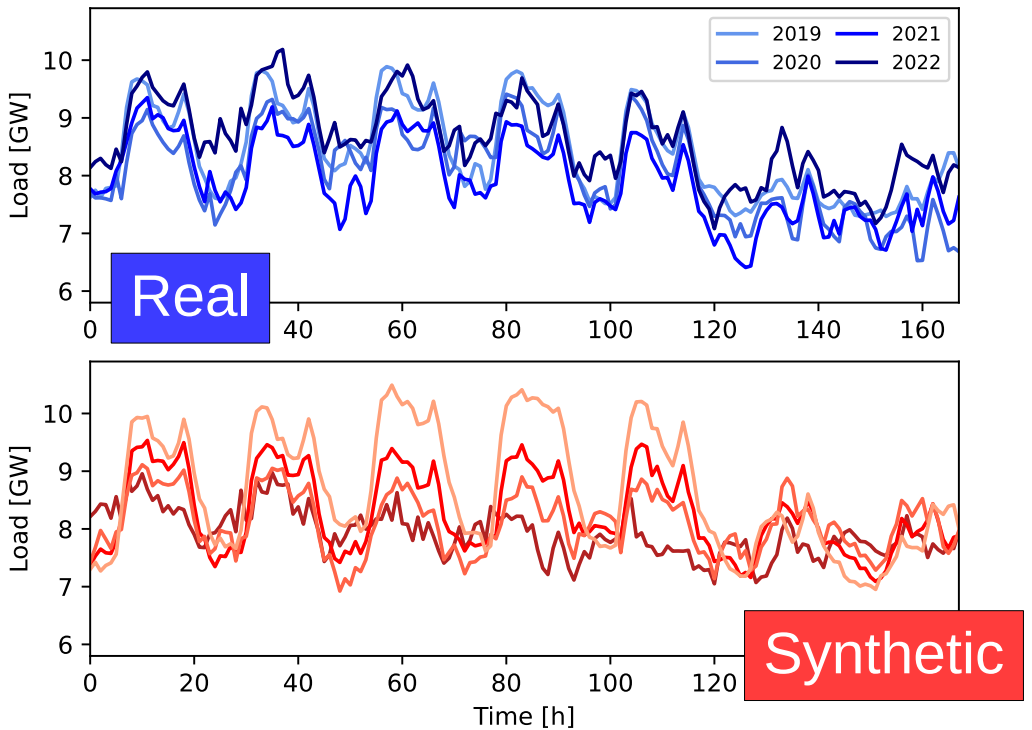
- Often no coherence in time
- When realistic time series exist, typically too few
  → spurious correlations
- Production dispatch from Optimal Power Flow (OPF) leads to unrealistically many saturated lines

Our solution

- Split the problem into load modeling and production dispatch
- **Loads**: build and use a statistical model
- **Production**: optimization problem
  - Thermal limits of the lines as objective instead of hard constraints
  - Constraints on integrated production + ramp constraints
  - Add noise to mimic wild electricity market

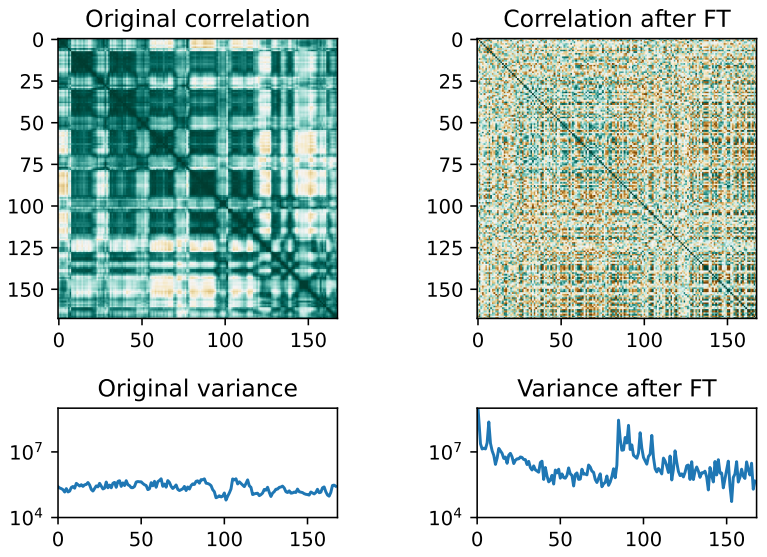
Synthetic load series

- Based on total load by country from public ENTSO-E data
  https://transparency.entsoe.eu/
- Use multiple years to estimate variance
- Build multivariate Gaussian distribution



Technicalities

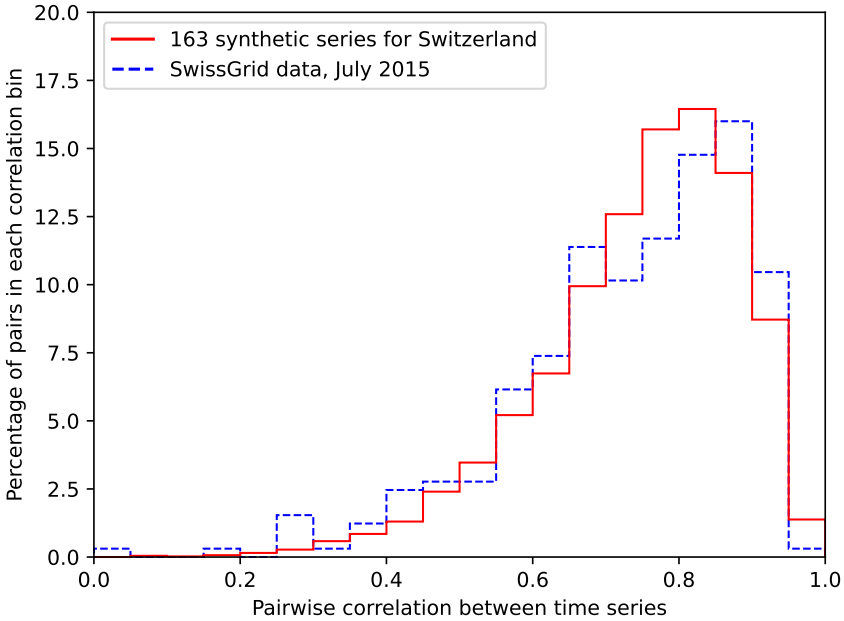
- Long time series → big correlation matrix
- Daily, weekly & yearly patterns
  → large off-diagonal entries
- Pass through Fourier transform
  for efficient modeling
  - Few “signal” frequencies, highly correlated
  - Many “noise” components, uncorrelated
  - FFT algorithm



Example with one-week long series

Realistic load series

- Arbitrarily many series can be generated quickly
- Mean corresponding to historical value
- Adjustable variance → fit correlations



- Room for improvement with more granular data, other scenarios, ...

Production dispatch

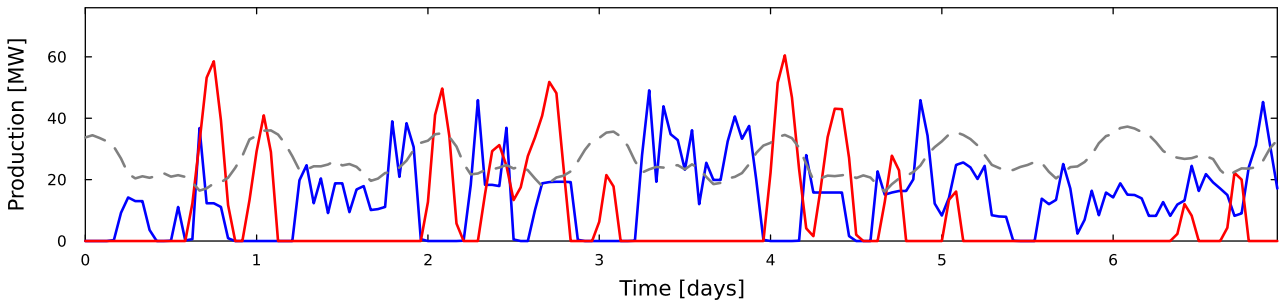
- Optimization problem (~~economic OPF~~)
- Minimize power flows through lines
  - penalize overloaded lines with quadratic cost (instead of hard constraint)
  - achieve relatively local production like TSO
- Annual constraints based on published availability
- Treat non-dispatchable sources separately (nuclear)

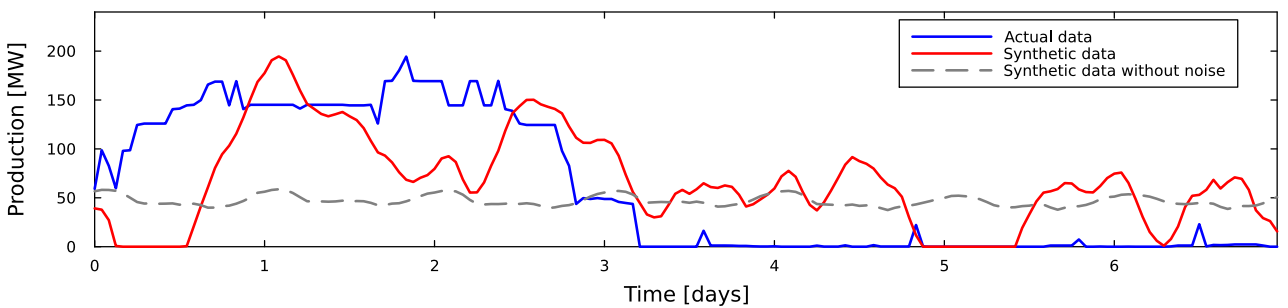
Optimal Power Flow

$$\min_{\vec{P}_t} \left[\sum_t \vec{P}_t^T \mathbf{A} \vec{P}_t + \sum_t \vec{b}_t \vec{P}_t \right] \quad \text{s.t.} \quad 0 \leq \vec{P}_t \leq \vec{P}^{\max}, \quad \sum_t \vec{P}_t = \vec{E}, \quad \vec{P}_t \cdot \vec{1} = L_t$$

- Complex behavior resulting from
  - variable loads
  - linear generation cost (noise with typical freq. + harmonics)
- Optimization problem is convex & feasible
  - runs on laptop with Gurobi: 1 year of data in 2-3 hours

OPF results

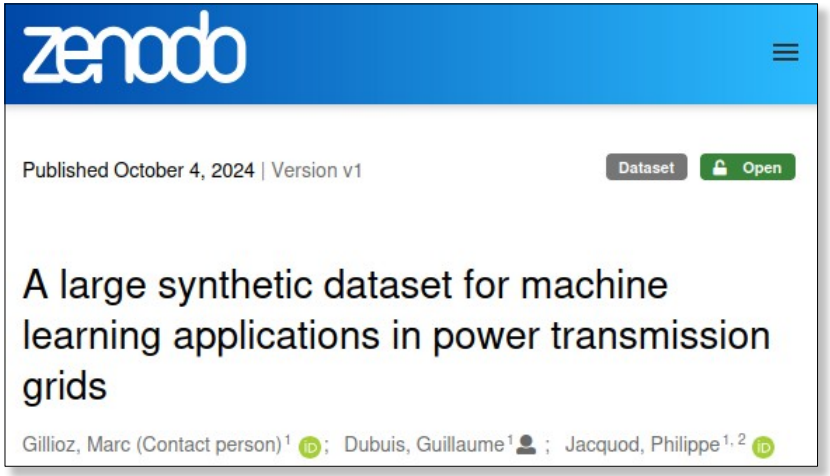




- Noise chosen as small as possible, but sufficient to trigger on/off behavior
- Several operating modes captured
- Realistic aggregated production by country and type

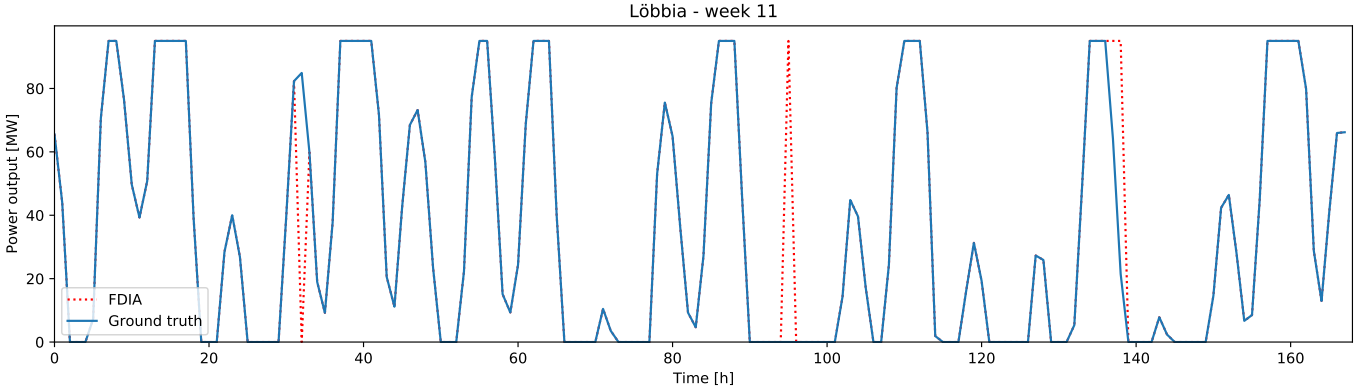
The published dataset

- Data available on Zenodo with data descriptor paper at Nature Scientific Data
- 20 years of time series with one-hour resolution for 815 generators and 4097 loads
- Open-access tools on GitHub repository https://github.com/GeeeHesso/PowerData



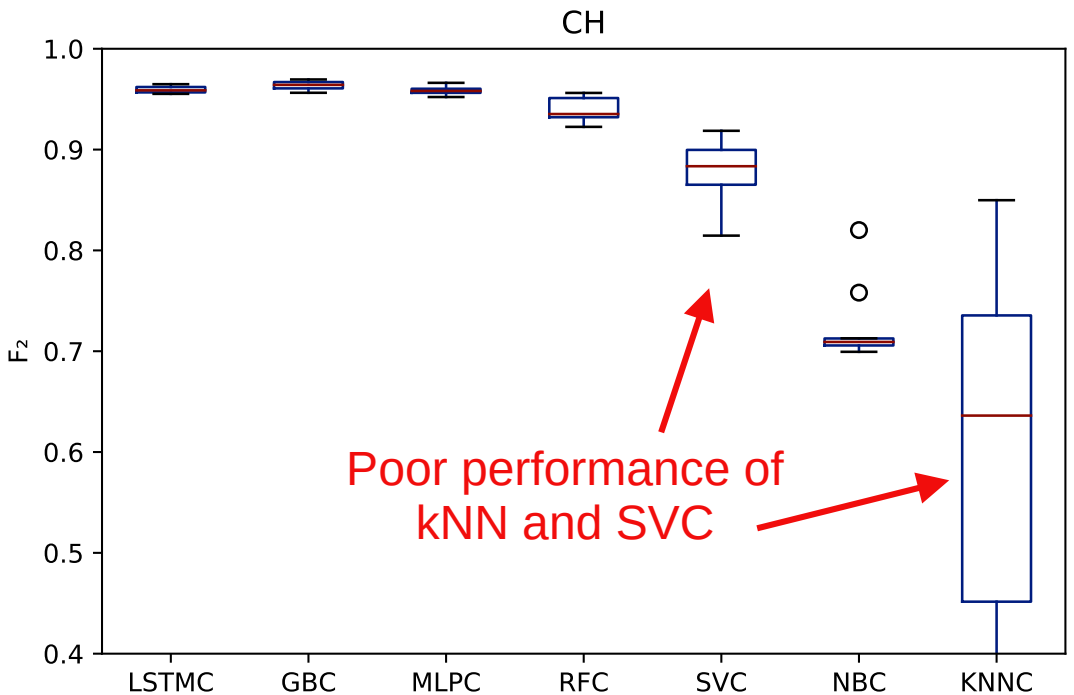
Application: anomaly detection

- Scenario: false data injection attack at one production site

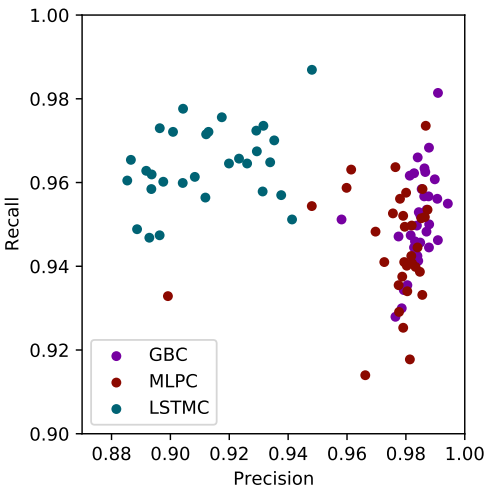


- Impossible to detect in isolation
- Focus on grid of one country (point of view of TSO)
- Use F_2 score (penalize false negatives more than false positives)

Comparing classification algorithms

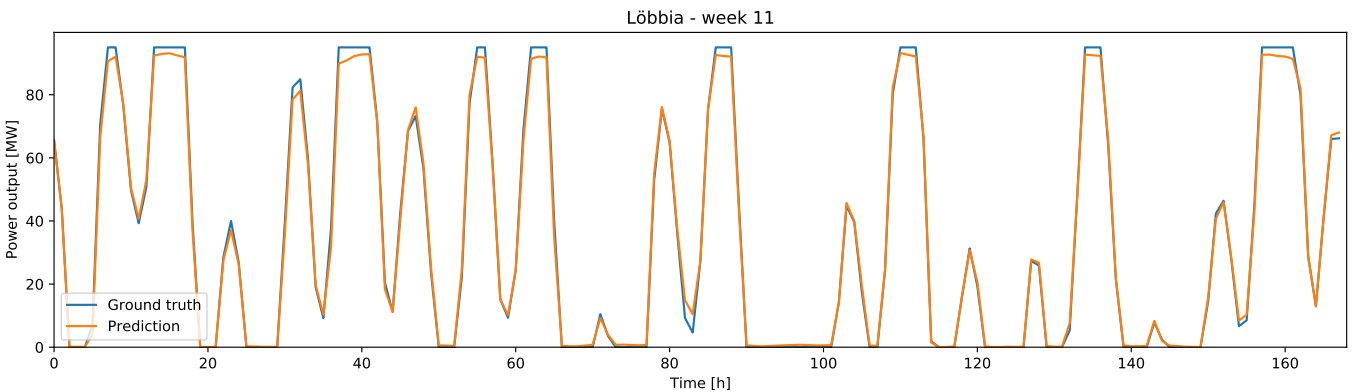


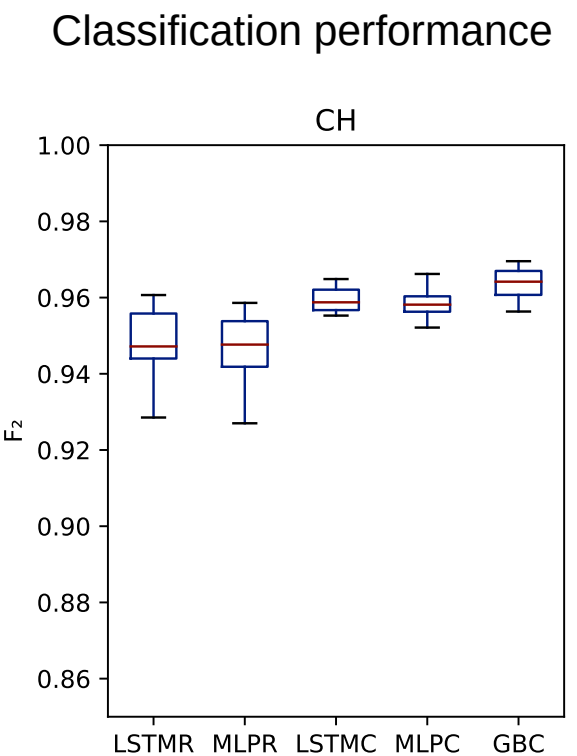
- Good performance of NN (deep and shallow)
- Consistent with purely **contextual** anomalies



Unsupervised algorithms

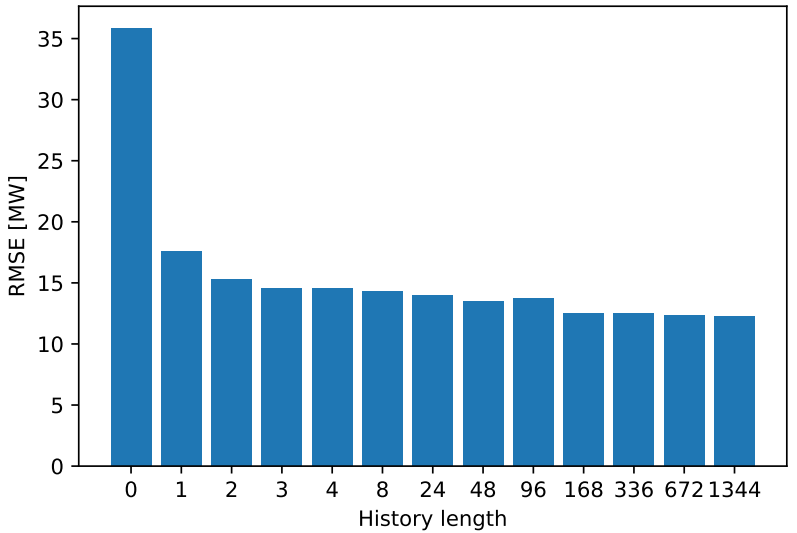
- Similar performances with unsupervised algorithms
- Much more versatile





Some lessons for anomaly detection

- History is crucial, but
  no need to go beyond few hours



- Relevant context:
  other production sources

- Grid distances have little impact

Conclusions

- A large and realistic **dataset** available to you
- New **methods** for data multiplication & modeling
- **Anomaly detection**: can you beat us with FM?

Thank you!