# A data augmentation methodology for machine learning modelling of distribution power grid: Application on optimal storage sizing and control.

*Amine Weibel[1*], Prof. Nicolas Jordan[1*], Prof David Wannier[*]*

*[1]Institute of Sustainable Energy (IEE), HES-SO Valais Wallis, Sion, Switzerland.*
**amine.weibel@hevs.ch, nicolas.jordan@hevs.ch, david.wannier@hevs.ch*

**Keywords**: MODELLING, SIMULATION, MACHINE LEARNING, VOLTAGE PREDICTION, OPTIMIZATION

## Abstract

The growth of distributed energy generations and electric vehicle charging stations in the low voltage grid brings out new challenges for distribution system operators. Actual methods of distribution network modelling are computationally expensive and omit aging of network physical components. This paper proposes a method of data augmentation for data-driven modelling. The methodology is divided into four parts: data generation, data-driven modelling, power flow boundaries estimation, and finally the application on optimal energy storage sizing and control.

## 1. Introduction

The growth of distributed energy generations (DEG) and electric vehicle charging stations (EVCS) in the low voltage (LV) grid brings out new challenges for distribution system operators (DSOs). Large photovoltaic (PV installations of DEG can cause located over-voltage; Whereas the high peak consumption introduced by EVCS can lead to low-voltage problem for customers. PV and EVCS do not just lead to voltage problems but as well to power problems (overcurrent lead to thermal overheating of power grid equipment resp. tripping of protection relays). One solution would involve energy storage integration to absorb the peak PV feed-in or compensate the EVCS consumption. Therefore, the energy storage must be optimally sized, placed and controlled. Many of the proposed methods require good knowledge of the network physical components. However, the previous distribution network modelling methods are computationally expensive. Besides, these methods omit environmental impacts (e.g temperature dynamics) and the aging of network physical components. Consequently, the resulting model are less accurate. Nevertheless, the authors of [2] attempt to predict voltage distribution using deep learning techniques. However, their methodology is based on a synthetic imbalanced dataset from uncalibrated model simulation, resulting in a fragile model. The author in this paper [3] developed a control strategy for energy storage system based on power prediction, but they omitted network constraint (voltage limits). Similarly, the authors of this paper [4] implemented an energy management strategy by designing an energy storage algorithm based on load and generation forecasting but also did not consider voltage constraints.

In this paper, a novel method of distribution network voltage prediction based on machine learning (ML) modelling is proposed. The goal is to build an accurate model of the distribution network and exploit this model for power flow boundaries estimation. Then, the estimated power flow boundaries are used to optimally size and control energy storage system.

## 2. Methodology

### *2.1 General concept*

The global concept of the proposed method is represented in the figure 1. First, data augmentation and machine learning modelling are described, and secondly the model application for sizing and control of energy storage.

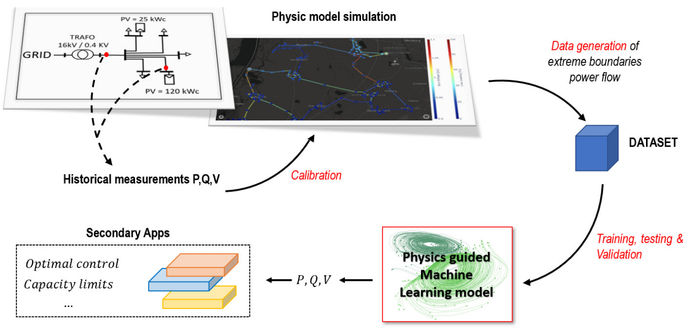


Fig.1 Data-augmented machine learning modelling concept.

The methodology requires some prior data:

- Network line diagram and components characteristic.
- Historical measurements of active power and voltage.

The historical measurements are used to calibrate the physic-based simulation of power flows. To be more restrictive, only active power (omit reactive power) are required. For each set points, data from extreme boundaries power flow simulation are generated. Iteration is made for a few set points and therefore build a dataset describing the network physics. Then, a machine learning model is trained on the dataset aiming to accurately predict voltage for given active power.

### *2.2 Power grid modelling and simulation framework*

Pandapower [5] which is an open-source python library has been used for power grid physic-based modelling and simulation. This library allows to conveniently automate the simulation process of solving the power flow equations and collect the results.

*2.3 Data generation*

The data generation part consists of collecting the calibrated power flow simulations results for each set point and build a dataset that describe the physics of the studied network. First, the physic model of the network is built. Hopefully, all required information of the components is available, otherwise similar characteristics are estimated for same voltage grid level.

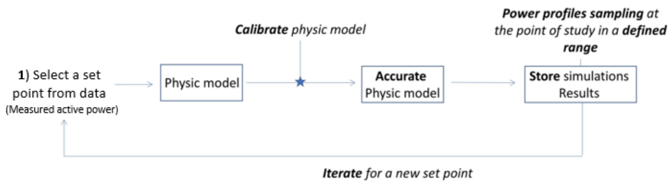


Fig.2 Data generation pipeline

When the physic model is ready, the iteration process is started. In the first step measured active power is inputted at the different location of measurement (minimum of two point, with one at the point of study). Then the power flow simulations are calculated. The simulated voltage result is compared to the actual measured one. The goal is to achieve lowest error between simulated and real measured voltage. To calibrate the network model, the higher voltage side of the main transformer is optimized. When the error passes a defined threshold, the optimal voltage magnitude is kept and the power profiles are sampled from a defined range at the study point and generate many simulations (e.g., inputted active power from -50kW to 50kW at the point of study, with a step of 1kW). Finally, those simulations results are stored and concatenated to constitute the so called "generated dataset".

*2.4 Data-driven modelling*

In this part of the procedure, the data-driven model is train and test. First, the generated dataset is used to train a machine learning model. In this paper the work is focused on the concept application, therefore state-of-the-art machine learning explanation and detailed will not be delivered. To train the models, a distributed gradient boosting open-source library called CatBoost [6] is used. This framework implements parallel tree boosting algorithms also known as gradient boosting decision trees (GBDT).

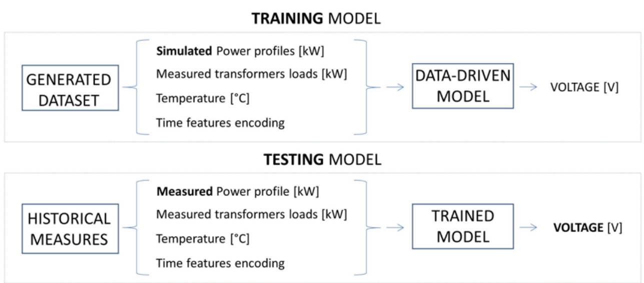


Fig.3 Data-driven modelling input/output

First, the model is trained with the following inputs: simulated power profiles, measured active power at transformer location, temperature, and encoded time features. The output is the voltage at the point of study. Then model is test with the historical measured active power excluding the few samples used for data generation.

The work on the modelling side is now achieved. The last two part consists in exploiting the accurate network model.

Now, the power flow boundaries as to be estimated because it will allow to relax the optimization problem related to energy storage sizing and control, by avoiding non-linearity in the optimization problem due to voltage constraints.

*2.5 Power flow boundaries estimation*

Firstly, the obtained data-driven model is exploited to predict voltages under the bounded power range. Note that voltage constraints are defined with (Umax, Umin). Then, the voltage prediction errors are computed. Finally, the power profiles with the lowest errors are maintained, being in range of the defined voltage limits.

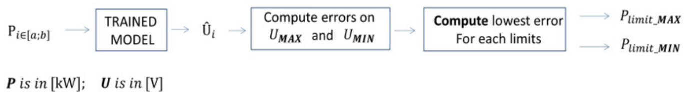


Fig.4 Power flow boundaries estimation

*2.6 Optimal energy storage sizing and control*

This section is dedicated to the optimization problem related to energy storage sizing and control including voltage constraints. The definition of this optimization problem can refer the author's previous thesis work entitled "Development of a pre-sizing tool for renewable energy systems" [1].

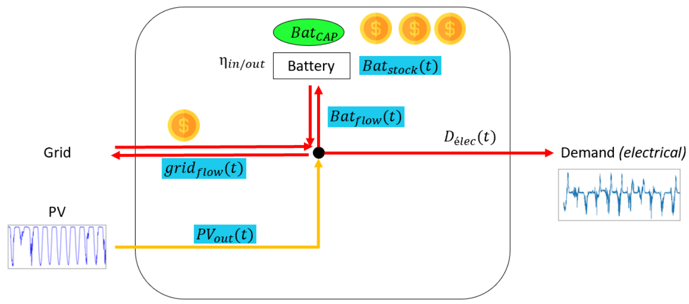


Fig.5 Linear power flow energy modelling diagram [3]

Figure 5 illustrates the linear diagram of the power flow modelling. The PV production and the grid must compensate the electrical demand. A battery is added to store excess of PV energy. The constraint equations of balance node (black dot in fig. 5) are expressed as follow:

$$grid_{flow}(t) = D_{elec}(t) - PV_{out}(t) + Batt_{flow}(t) \quad (1)$$

$$grid_{flow}(t) \geq p_{\boldsymbol{limit\_min}}(t) \quad (2)$$

$$grid_{flow}(t) \leq p_{\boldsymbol{limit\_max}}(t) \quad (3)$$

$$cost(t) \geq objective \cdot c_{buy} \cdot \Delta t \cdot grid_{flow}(t) \qquad (4)$$

$$cost(t) \geq objective \cdot c_{sell} \cdot \Delta t \cdot grid_{flow}(t) \qquad (5)$$

With $p_{limit\_min}$ and $p_{limit\_max}$ the power flow boundaries estimated using the data-augmented ML model. The operational cost of the grid energy consumption or feed-in are considered with the constant $c_{buy}$ and $c_{sell}$. The objective constant is binary, depending on whether the goal is to minimize consumer cost or maximize DSO gains.

$$Batt_{stock}(t) = \tau_{1-loss} \cdot Batt_{stock}(t-1) + \Delta t \cdot \tau_{batt} \cdot Batt_{flow}(t) \qquad (6)$$

$$Batt_{stock}(t) \leq Batt_{\max\_cap} \qquad (7)$$

$Batt_{stock}(t)$ is the state of charge of the battery, who is limited arbitrary by a maximum capacity in kWh ($Batt_{\max\_cap}$).

Finally, the objective function (eq.8) aims to minimize operational cost and capital expenditure. It means, find the smallest (less expensive) battery that satisfies the constraints of the established power flow:

$$\min[(c_{battery} \cdot \max(Batt_{stock})) + \sum \cdot cost] \qquad (8)$$

To implement the optimization problem, the python open-source library Pulp [7] is used. This library is suited for linear optimization problem with timeseries dimension.

## 3. Results

The concept has been applied and validated during the research project DIMS (Distributed Intelligent Micro Storage) from the Institute of Energy and Environment of the engineering school of Valais (HES-SO Valais Wallis). The work package aims to optimally size and control energy storage system. For this paper, one of the use cases is presented. The use case concern a DSO project partner FMA SA (Force mortice de l'Avançon).

### *3.1 Case under study*

The study case provided by the partner is about a PV plant on a farm, located on one of their distribution network edge. On this location, the power demand in summer is low, however, the energy production is high, which cause over-voltage problems. The wish of the DSO is to solve this voltage problem by adding an energy storage to absorb the production surplus.

### *3.2 Available data*

The DSO gave the network physical components characteristic and the data they collected. The figure 9 shows the network topology, with the point of study (Farm) and two red dots representing the location of measurement points. The graphs of the figure 7 represent the measurements of voltage and active power.

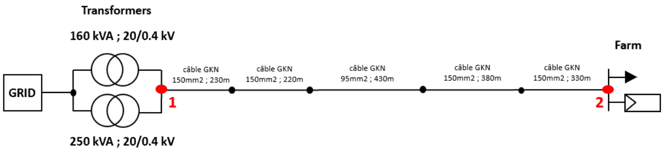


Fig.6 Network topology and characteristic with two measurement points (red dots).

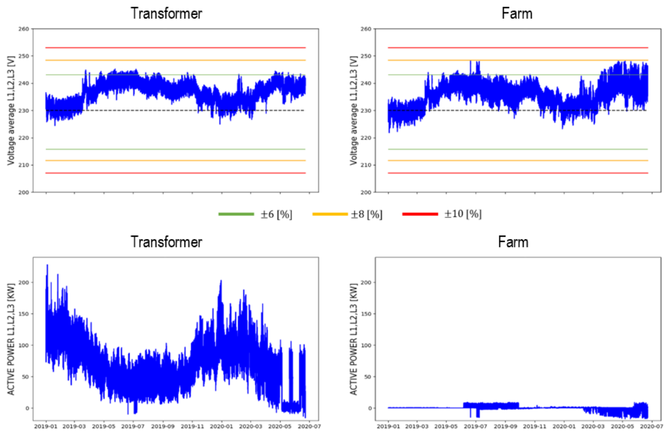


Fig.7 Transformer (left) and farm (right), voltage (top) and active power (bottom) 1-year measurement.

### *3.3 Data exploration*

The figure 8 helps analyse that most of the time the farm has a very small demand. 3 clusters can be identified: the yellow cloud is a demand near zero, the red cloud corresponds to measurements over 3kW and the blue cloud is the PV production. Therefore, the measured data are highly unbalanced and will not be usable for training a machine learning model.

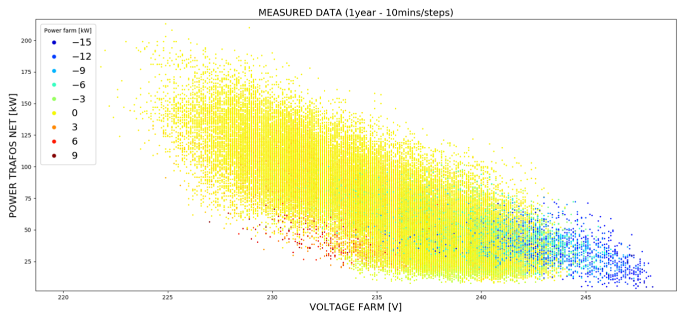


Fig.8 Active power at transformer against voltage at farm and active power at farm.

### *3.4 Data generation*

Multiple datasets are generated, characterized by different samples (number of days variation). The power profile interval is defined between -60kW and 50kW, with a step of 1kW.

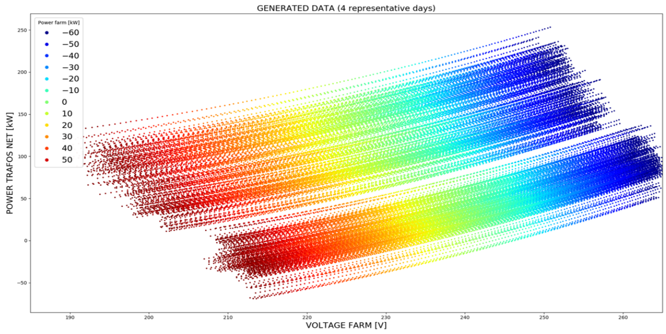


Fig.9 Active power at transformer against voltage at farm.

### *3.5 Data-driven modelling*

Finally, with the multiple generated datasets composed of different number and type of days sample the machine learning model is trained to predict voltage at the farm.

| Model | Dataset *(days)* | RMSE* [V] | MAE* [V] | R2* [-] |
|---|---|---|---|---|
| Physic | None | 2.67 | 2.12 | 0.600 |
| Model 1 | 4 (min/max) | 0.93 | 0.59 | 0.950 |
| Model 2 | 4 (season) | 0.69 | 0.40 | 0.974 |
| **Model 3** | 8 (min/max & season) | **0.68** | **0.33** | **0.977** |

Table.1 Model's performance metrics on voltage prediction (RMSE: root mean squared error, MAE: mean absolute error, R2: coefficient of determination)

In the table 1, the data driven model based on generated dataset achieves the best performance. The model 3 trained on a dataset with 8 days as sample reaches a mean squared error of 0.68 [V], which is an improvement of 75% compared to the physic only model. Those 8 days selected are from minimum and maximum value of the measured data and some representative days for every season of the year. The figure 10 illustrated the voltage prediction of the model 3 on a 2-week sample (timestep of 10 minutes).

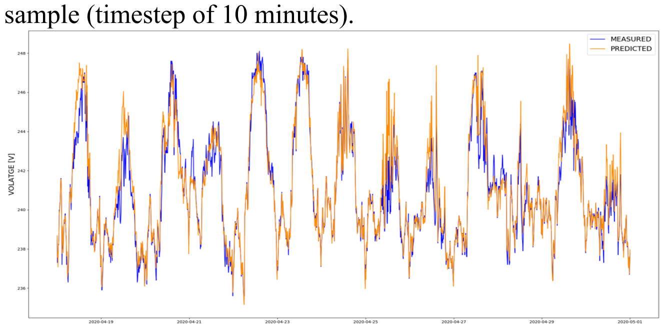


Figure.10 Example of 2 weeks voltage prediction compared to real measurement.

### *3.6 Power flow boundaries estimation*

Now that an accurate model of the distribution network of the farm is obtained, the latter is exploited to estimate the power flow boundaries corresponding to a defined voltage limit. For example, for every timestep the maximal and minimal active power is estimated to respect the voltage limits.

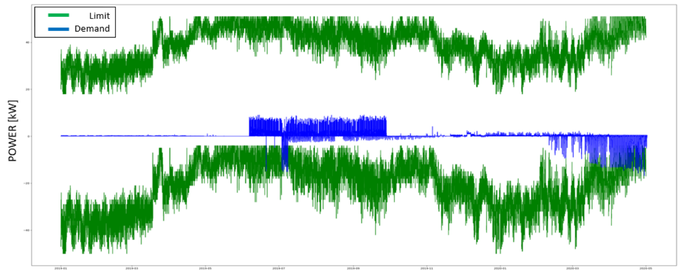


Figure.11 In green the estimated higher and lower power flow limits to respect 230[V] $\pm 0.06pu$ at the farm. In blue the measured net active power at the farm. The data are from 01/01/2019 to 01/05/2020.

### *3.7 Optimal energy storage sizing and control*

Finally, the power flow boundaries estimations are collected to constrain the optimization problem of the storage sizing. "*Bridage*" constant is then iterated, which signify limiting the maximum power allowed to be injected in the grid by the PV. In this manner, the battery capacity is not overestimated, hence removing the peak production of the PV plant.

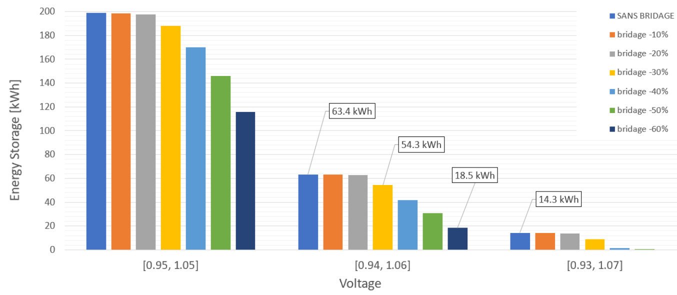


Fig.12 Optimal capacities of energy storage against voltage constraints.

The figure 12 resumes the optimal solutions for the energy storage sizing. First, analysis show that the more restricted the voltage is, the higher the energy capacity is required. Secondly, the optimal "bridage" ratio is between -30% and -20% of the maximum power injected by the PV. This means cutting 20% to 30% of the peak production should be optimal.

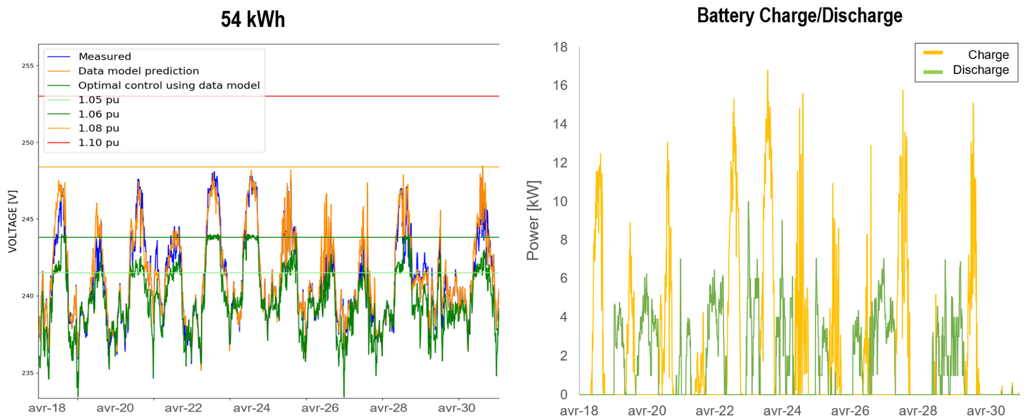


Fig.13 Optimal energy storage control for voltage regulations

Finally, the figure 13 represents the optimal control using the data-driven model (green curve). The maximum voltage was set to 1.06pu (243.8V) and the injection maximum voltage to 240V. As shown in the figure the control effectively follows the constraints. The graphs illustrated the effectiveness of the optimal control algorithm. On the right-side of the graph, the charge and discharge power profile describe the control that generate the voltage on the left side graphs.

## 4 Conclusion

In this paper, a method for power grid modelling and simulation based on data-augmented machine learning is developed. The process of data generation for the purpose of model training, highlighted the importance of physic model calibration. The ML model is exploited to estimate power flow boundaries and used as balance equations constraints for the optimal energy storage sizing and control optimisation problem. Finally, the results obtained demonstrate on a real case the accuracy of such methodology.

## 5 Acknowledgements

I would first like to thank my professor Prof. N. Jordan and Prof D. Wannier for the confidence they have in my ideas and their unwavering support. I also thank Dr.Baoling Guo for her help and advice in writing this paper. In addition, I thank Dr.Iordan Doytchinov, a close friend with whom we give birth to innovative ideas. To conclude, I dedicate this research paper to my dear father who has always supported me in my projects and taught me to never give up.

## 6 References

**Thesis**

[1] Weibel, A.: 'Outil de prédimensionnement des systèmes énergétiques renouvelable’. BSc thesis, HES-SO, 2019.

**Conference Paper**

[2] Ashish B. G., Tanzu A., Anthony B. M. : 'Predicting Voltage Variations in Low Voltage Networks with Prosumers’. APPEEC, 2018.

[3] Joe-Air, J.: 'An intelligent control strategy for energy storage systems in solar power generation based on long short-term power prediction’. iEECON, 2020.

[4] Md Morshed A., Md. Habibur R.. : 'Energy management by scheduling ESS with active demand response in low voltage grid ’. ICTC, IEEE, 2020.

**Websites**

[5]‘Pandapower’,https://pandapower.readthedocs.io/en/v2.5.0

[6]‘CatBoost’, https://catboost.ai/

[7]‘Pulp’, https://coin-or.github.io/pulp/