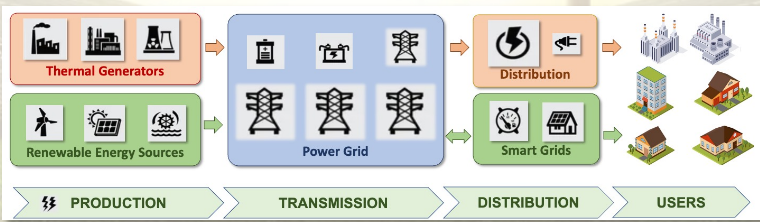
UNIVERSITÀ
DI PAVIA

Statistical Learning algorithms for forecasting wind production

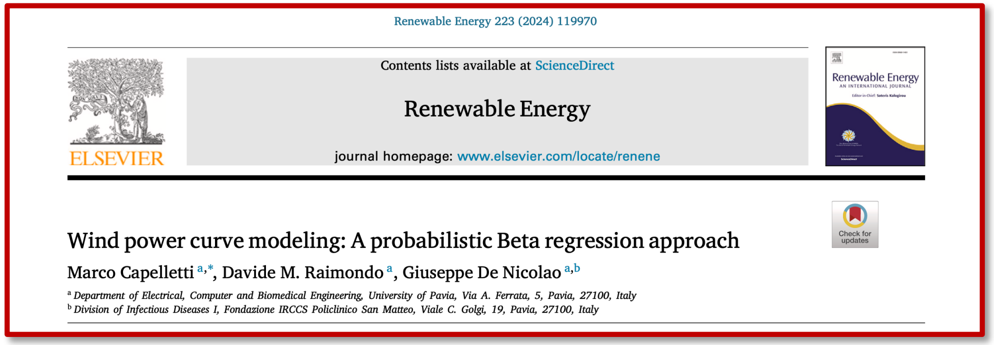
Giuseppe De Nicolao, Marco Capelletti
Department of Electrical, Computer and Biomedical Engineering, University of Pavia

HEXAGON workshop – 18/06/2024



Summary

1. Introduction
2. The challenge of heteroschedasticity and asymmetry
3. Beta Regression Model with preconditioning
4. Results


Renewable Energy 223 (2024) 119970

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

ELSEVIER

Wind power curve modeling: A probabilistic Beta regression approach

Marco Capelletti [a,*], Davide M. Raimondo [a], Giuseppe De Nicolao [a,b]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via A. Ferrata, 5, Pavia, 27100, Italy
[b] Division of Infectious Diseases I, Fondazione IRCCS Policlinico San Matteo, Viale C. Golgi, 19, Pavia, 27100, Italy

UNIVERSITÀ DI PAVIA

Summary

1. Introduction
2. The challenge of heteroschedasticity and asymmetry
3. Beta Regression Model with preconditioning
4. Results

UNIVERSITÀ
DI PAVIA

The team

M. Capelletti
Postdoctoral
researcher

UNIVERSITÀ DI PAVIA
Dipartimento di
Ingegneria Industriale
e dell'Informazione

Prof. G. De Nicolao

UNIVERSITÀ DI PAVIA
Dipartimento di
Ingegneria Industriale
e dell'Informazione

UNIVERSITÀ
DI PAVIA

Introduction

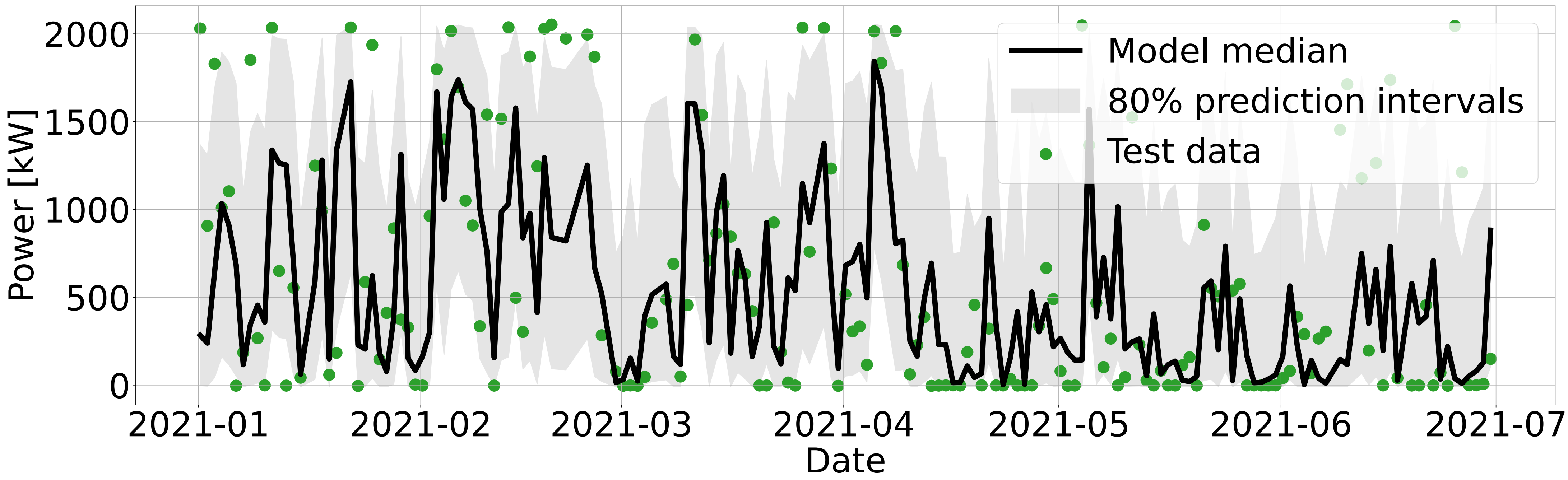
Given a wind farm one of the main goals is to estimate and characterize its power production at least a day ahead

Why is this important?

1. Economic Perspective:
    - Participate in day-ahead electricity markets.
    - Optimize bidding strategies in uncertain scenarios.
2. Grid Stability:
    - Predict and manage grid imbalances.
    - Activate storage or backup power in advance.

UNIVERSITÀ DI PAVIA

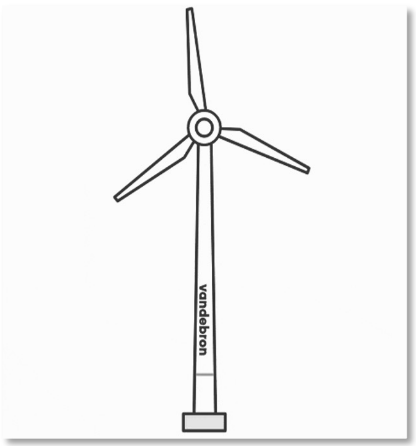
Wind power forecasting:
A probabilistic approach

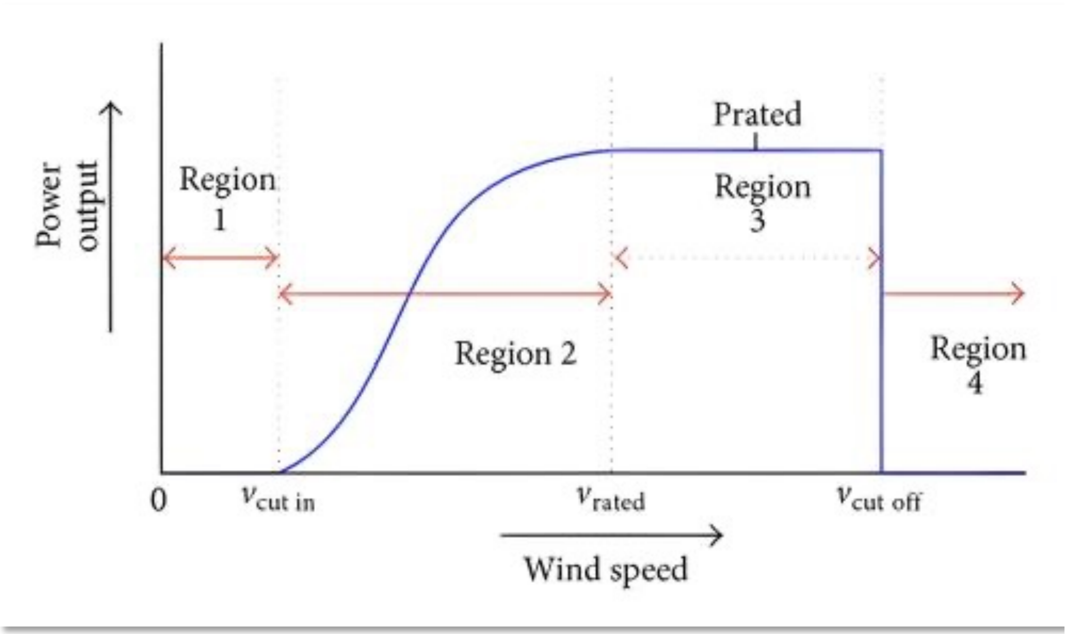


UNIVERSITÀ DI PAVIA

Introduction

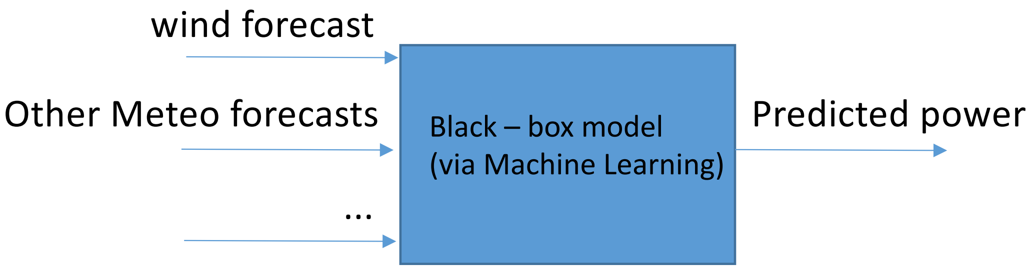
- Power curves are characteristic functions that model and describe both individual wind turbines and virtually the entire wind farm
- Typical uses range from wind power forecasting to wind turbine condition monitoring

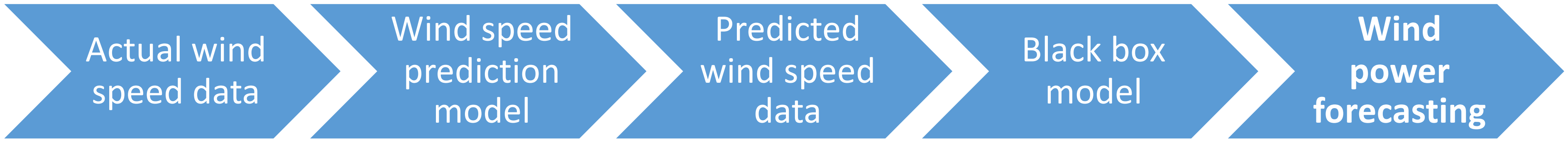




Wind power forecasting: direct vs indirect approach

- Direct approach

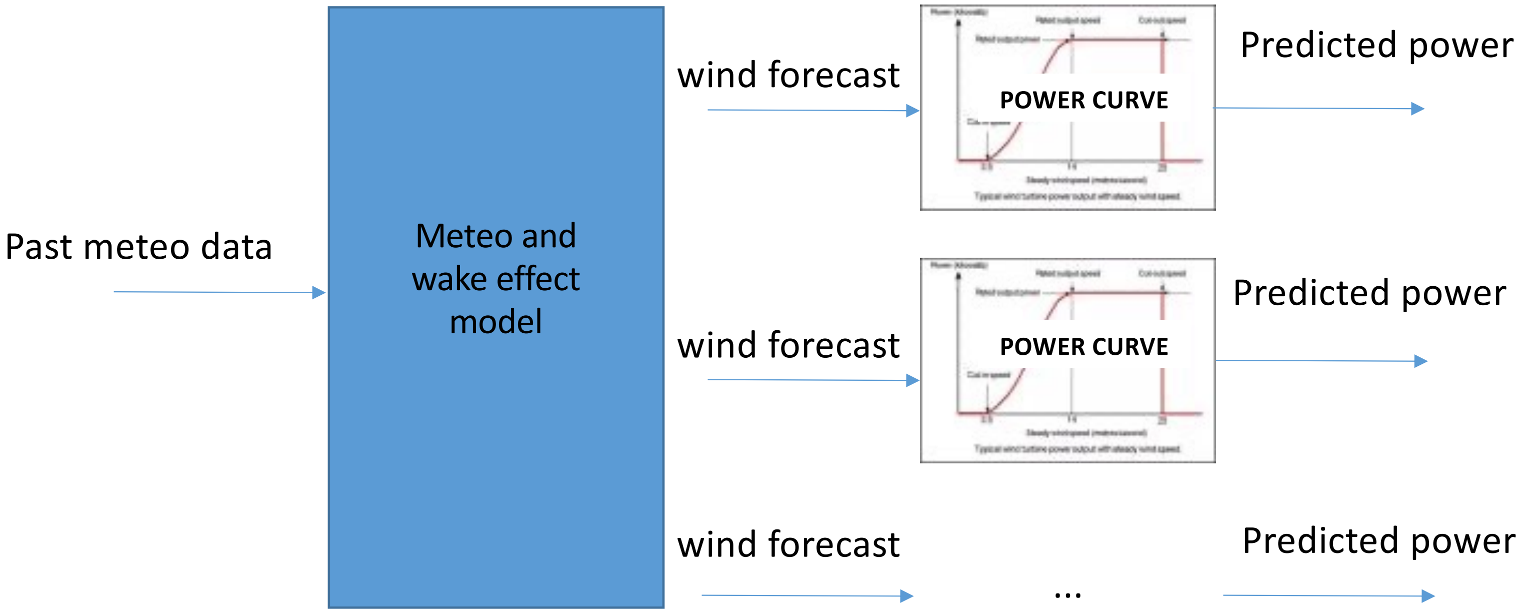




UNIVERSITÀ DI PAVIA

Wind power forecasting: direct vs indirect approach

- Indirect approach



Direct vs indirect approach in renewables forecasting

Direct vs indirect approach: pros

Direct approach

- Faster deployment: given enough data, it is possible to directly identify a model with machine learning techniques

Indirect approach

- Greater interpretability : wake effect & physical phenomena
- Design and Upgrade made easier

UNIVERSITÀ DI PAVIA

Direct vs indirect approach: cons

Direct approach

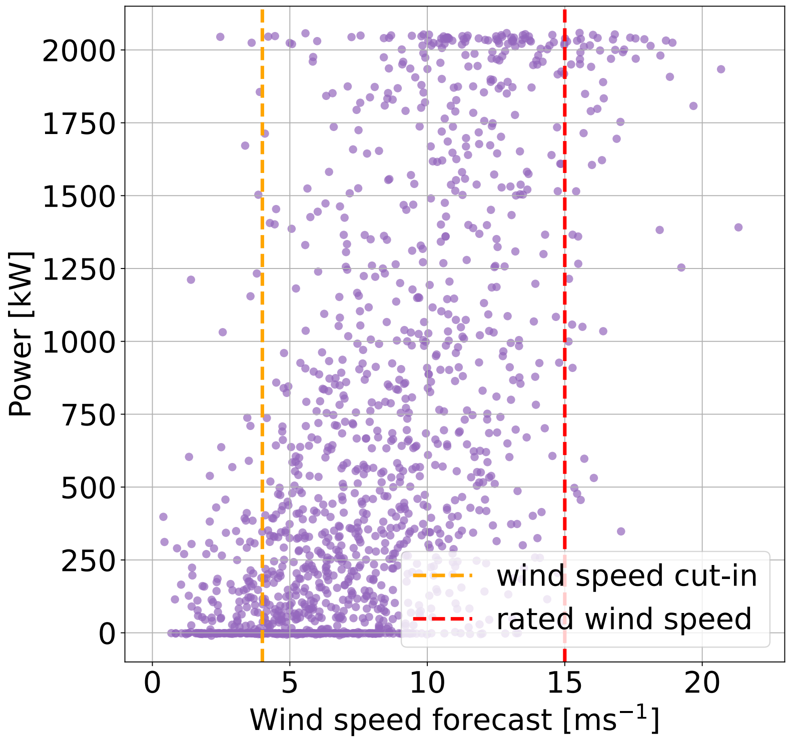
- Poor interpretability
- less flexible and adaptable: it works as long as the operating conditions are maintained

Indirect approach

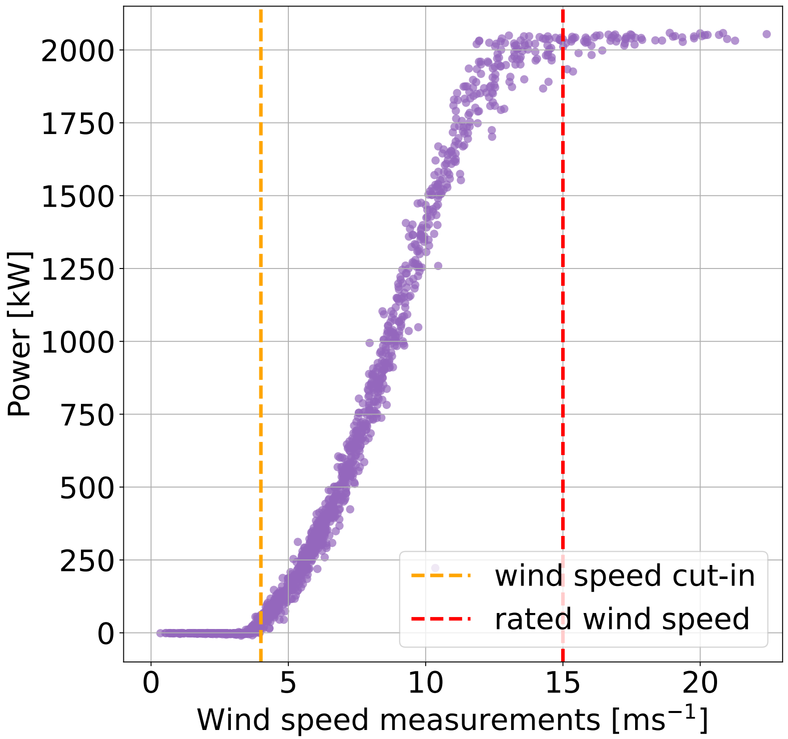
- Complexity: more expertise is required
- Slower deployment

UNIVERSITÀ DI PAVIA

Forecasting problem – in practice

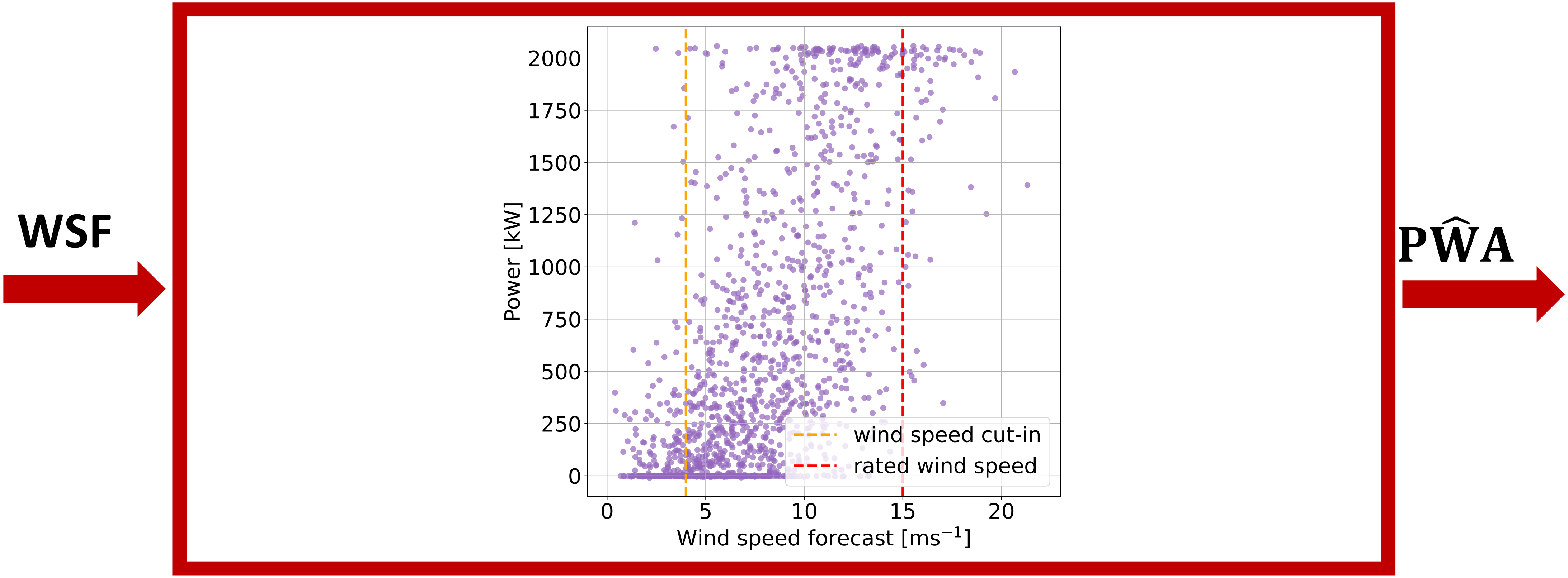


VS



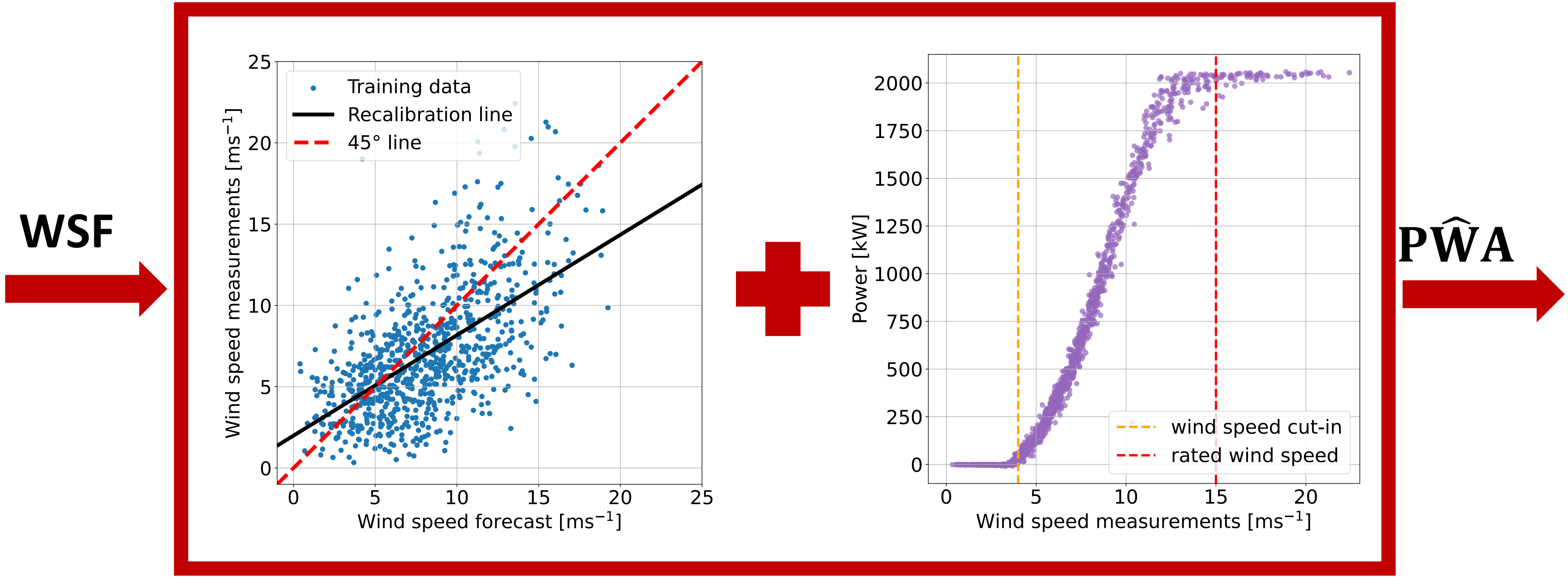
Direct Approach

Legend:
- **WSF:** Wind speed forecast
- **PWA:** Power actual



Indirect Approach

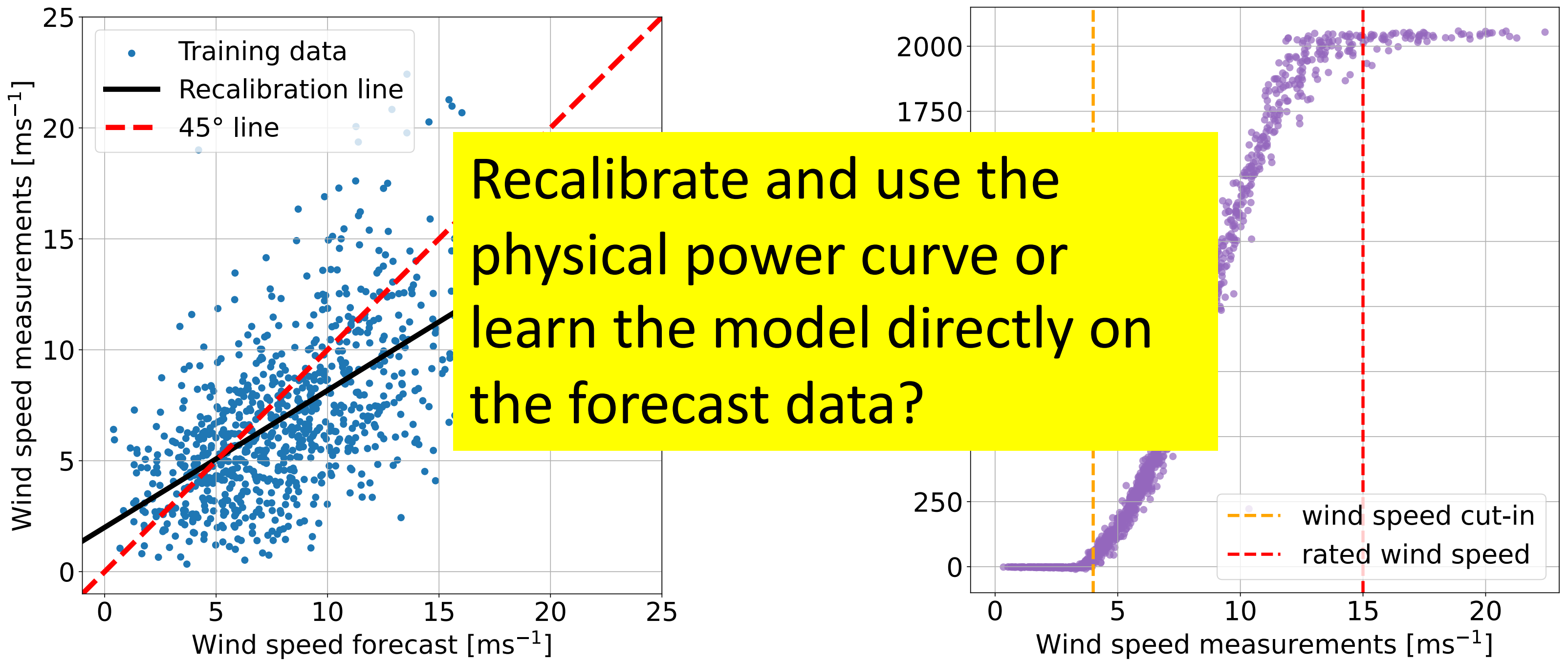
Legend:
- **WSF:** Wind speed forecast
- **PWA:** Power actual



Indirect Approach

Legend:
- **WSF:** Wind speed forecast
- **PWA:** Power actual

WSF



Recalibrate and use the physical power curve or learn the model directly on the forecast data?

$\mathbf{P\widehat{W}A}$

We will now focus on the direct approach

Power curve identification

Power curve identification: → Raw wind speed forecast & power data → Data processing → Choose power curve models → Optimal model parameters → Output optimal power curve

UNIVERSITÀ DI PAVIA

Power curve identification

UNIVERSITÀ DI PAVIA

Dataset

Dataset specifications:

- https://doi.org/10.5281/zenodo.8253010
- https://cds.climate.copernicus.eu/

Dataset Overview:

- **Measurements:** SCADA data from the first Senvion MM82 turbine at Penmanshiel wind farm, UK.
- **Forecasts:** UK Met Office wind speed forecasts at 8 horizons (6, 12, 18, 24, 30, 36, 42, 48 hour ahead) starting from midnight.
- **Time Period:** August 1, 2017 - July 1, 2021.
  - **Training Set:** August 1, 2017 - December 31, 2019.
  - **Test set:** January 1, 2020 - July 1, 2021.

UNIVERSITÀ DI PAVIA

Dataset

Dataset specifications:

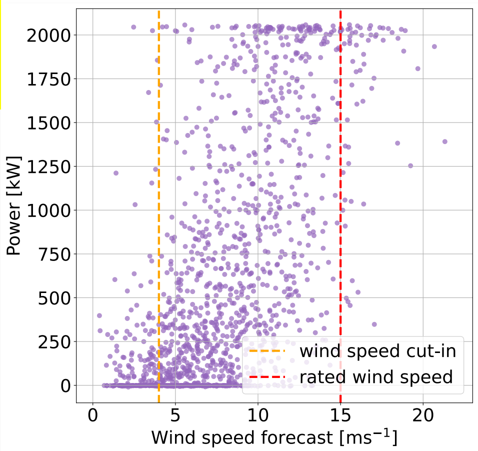
- https://doi.org/10.5281/zenodo.8253010
- https://cds.climate.copernicus.eu/

Dataset Overview:

- **Measurements:** SCADA data from the first Senvion MM82 turbine at Penmanshiel wind farm, UK.
- **Forecasts:** UK [...] rizons (6, 12, 18, 24, 30, 36, [...]

For our analysis, we will initially focus on the first forecasting horizon, which is 6 hour ahead

- **Time Period:** August 1, 2017 - July 1, 2021.
  - **Training Set:** August 1, 2017 - December 31, 2019.
  - **Test set:** January 1, 2020 - July 1, 2021.



UNIVERSITÀ DI PAVIA

Power curve identification

Power curve identification: → Raw wind speed forecast & power data → Data processing → Choose power curve models → Optimal model parameters → Output optimal power curve

UNIVERSITÀ DI PAVIA

Data processing

- Literature Review: Various techniques exist for identifying outliers in wind power vs wind speed data.

- SCADA Data Specifics: Our SCADA data includes a variable called ‘Lost Production to Downtime and Curtailment Total (kWh)’.

- Data Filtering: We retained only the values where 'Lost Production to Downtime and Curtailment Total (kWh)' is equal to zero, ensuring data quality by excluding periods of downtime and curtailment.

Power curve identification

Power curve identification:

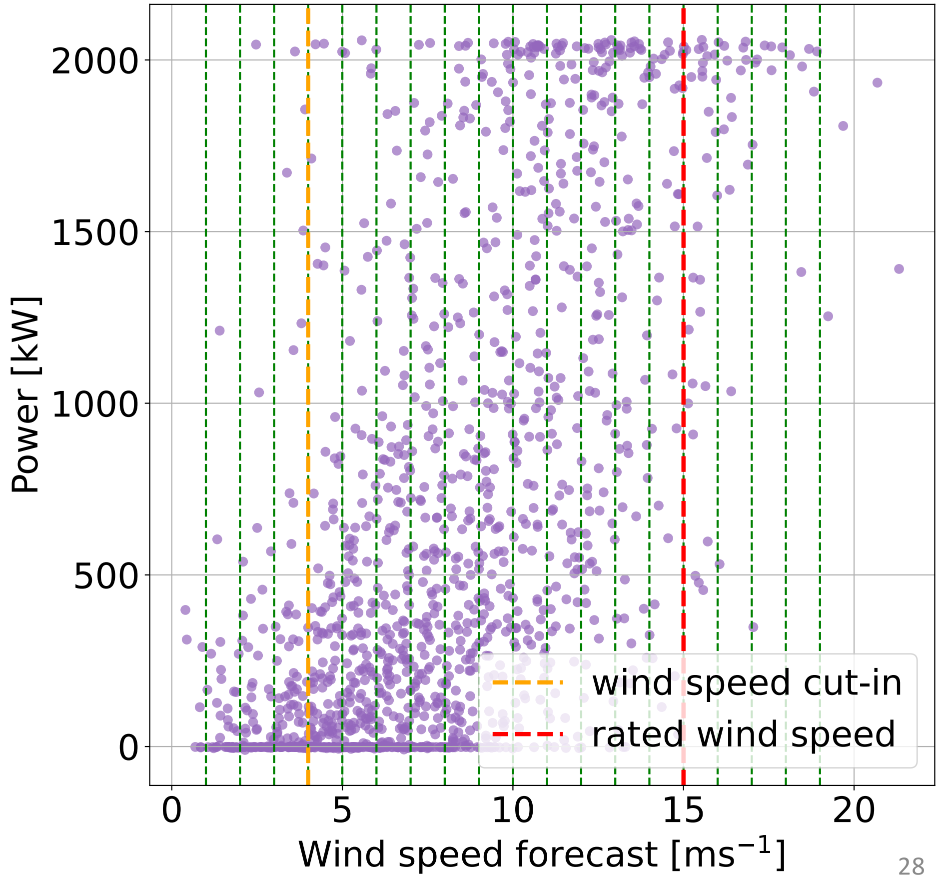
Raw wind speed forecast & power data → Data processing → **Choose power curve models** → Optimal model parameters → Output optimal power curve

UNIVERSITÀ DI PAVIA

Summary

UNIVERSITÀ
DI PAVIA

The challenge of heteroschedasticity and asymmetry



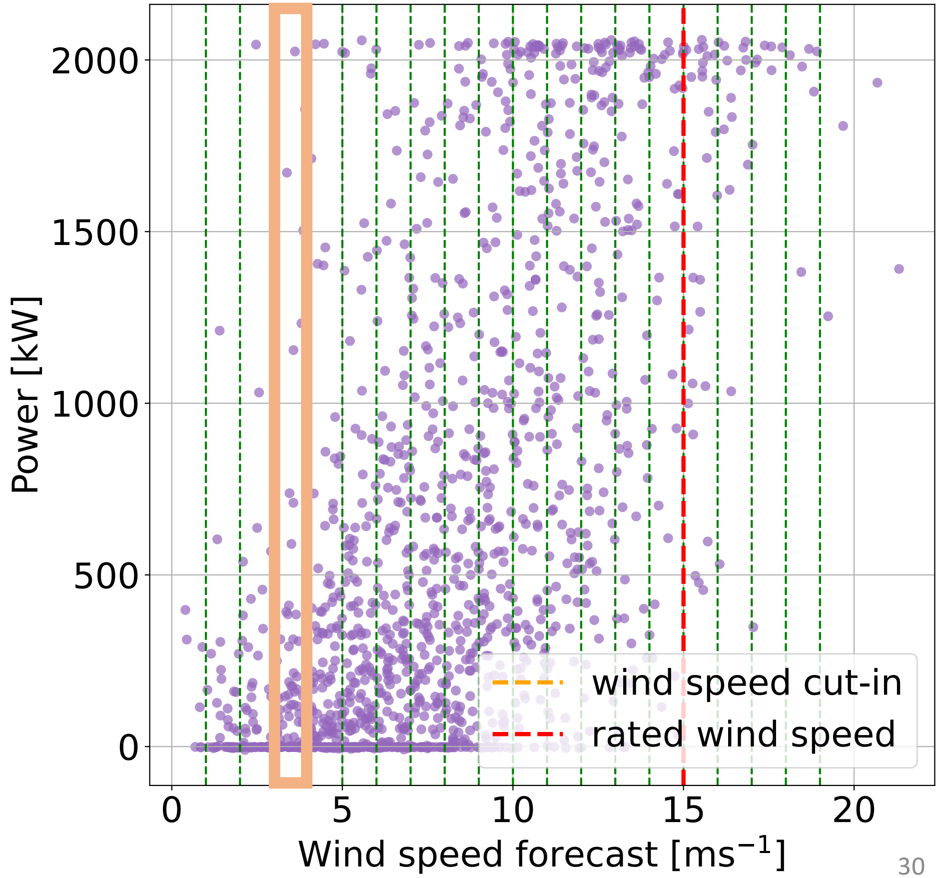
UNIVERSITÀ DI PAVIA

The challenge of heteroschedasticity and asymmetry

To showcase wind power distribution across various wind speeds, we'll analyze histograms of wind power observations within specific speed ranges (distributions of power conditional on wind speed forecast).

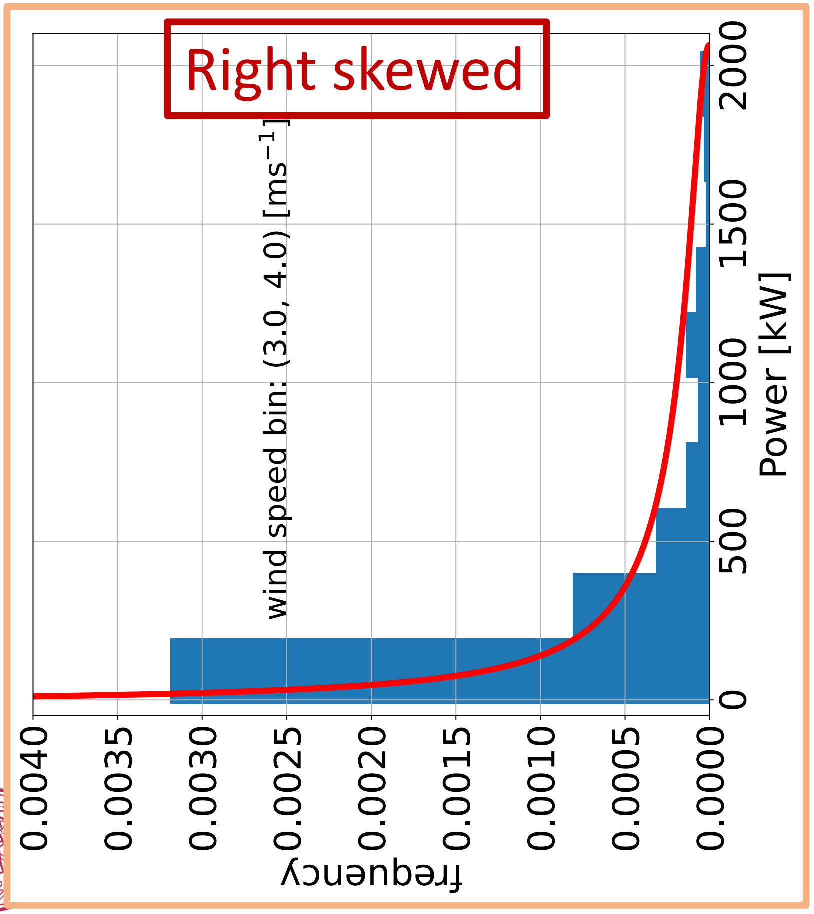
UNIVERSITÀ DI PAVIA

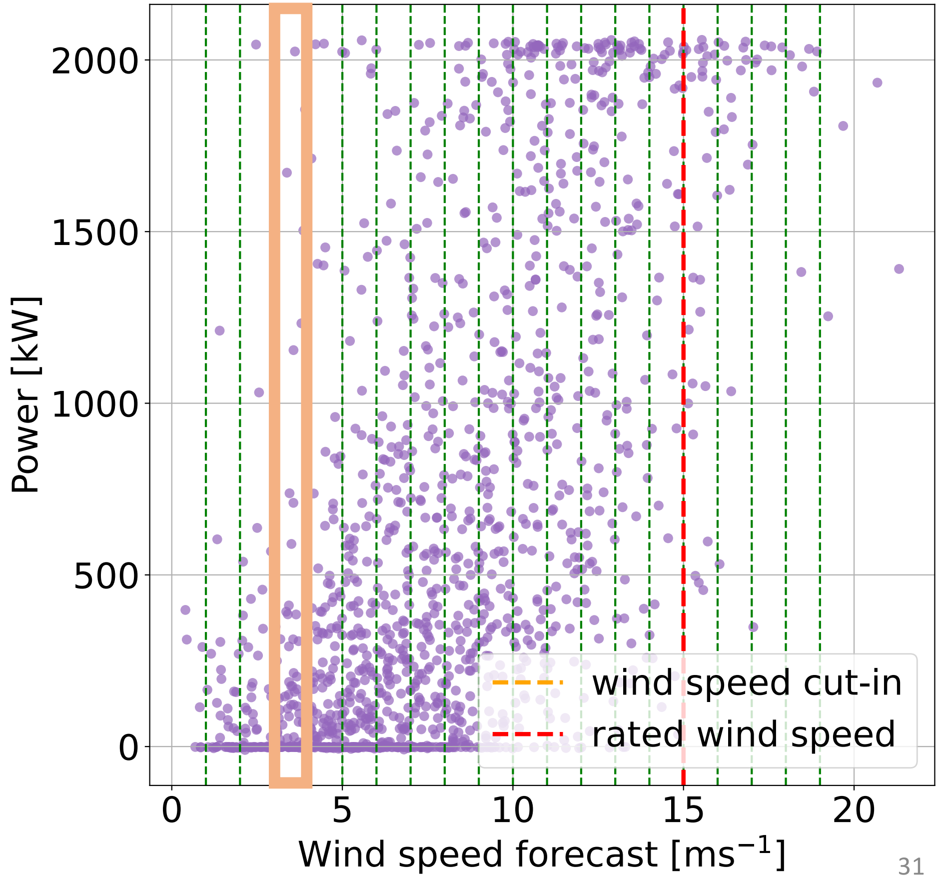
The challenge of heteroschedasticity and asymmetry



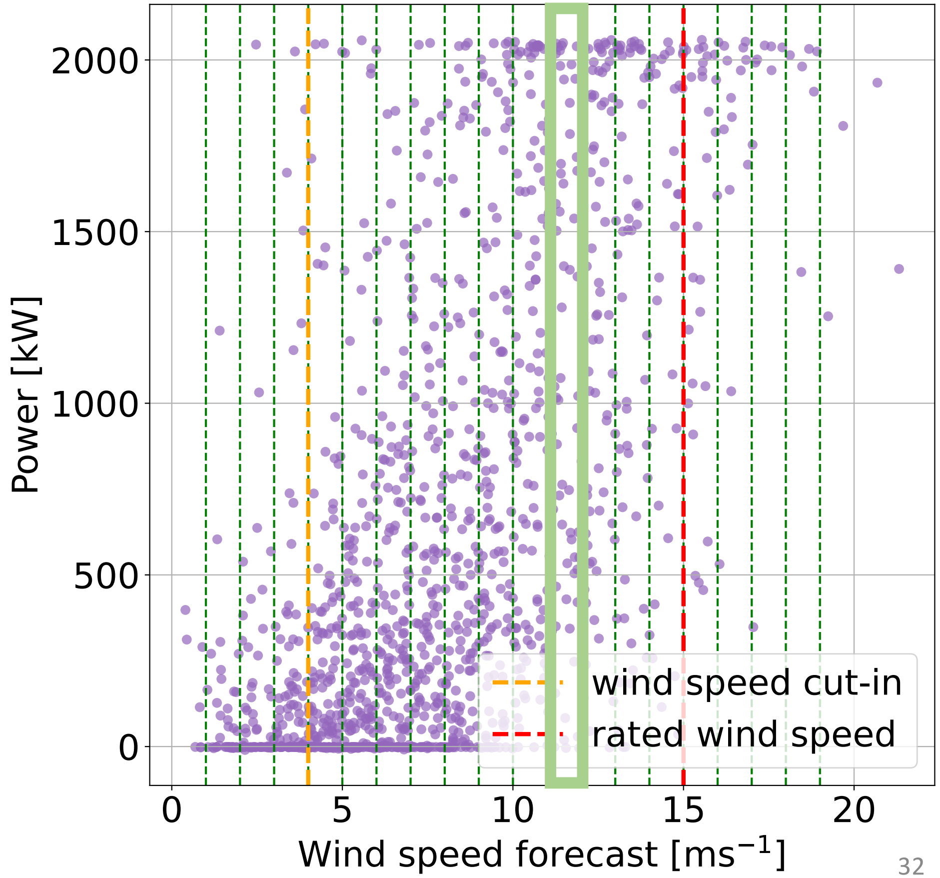
UNIVERSITÀ DI PAVIA

The challenge of heteroschedasticity and asymmetry



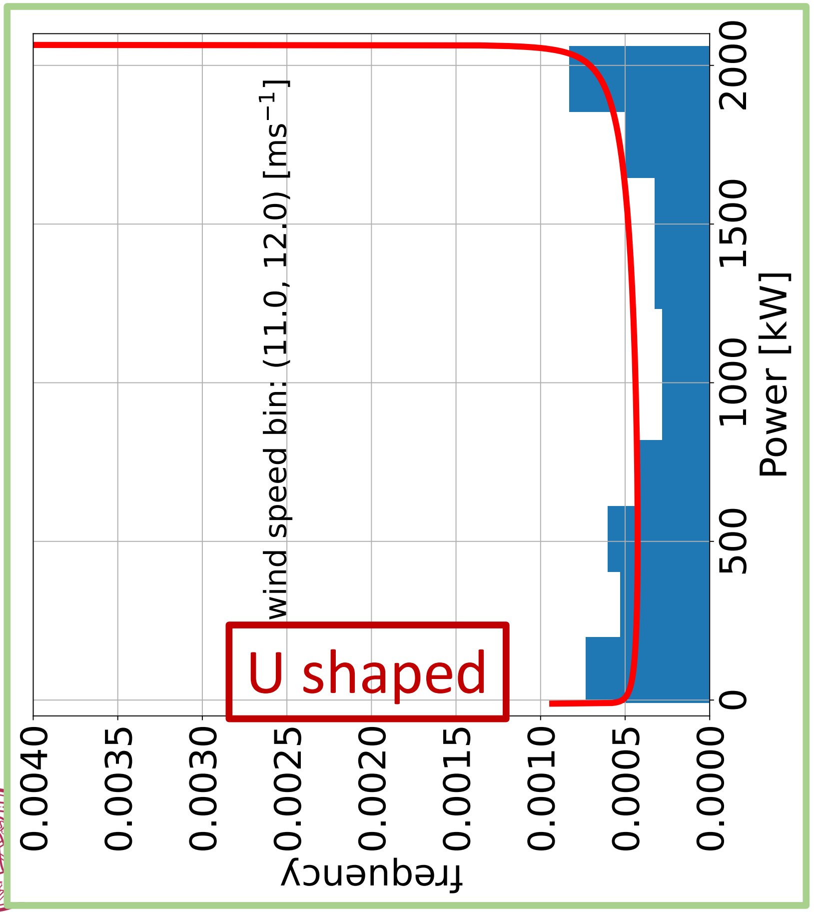


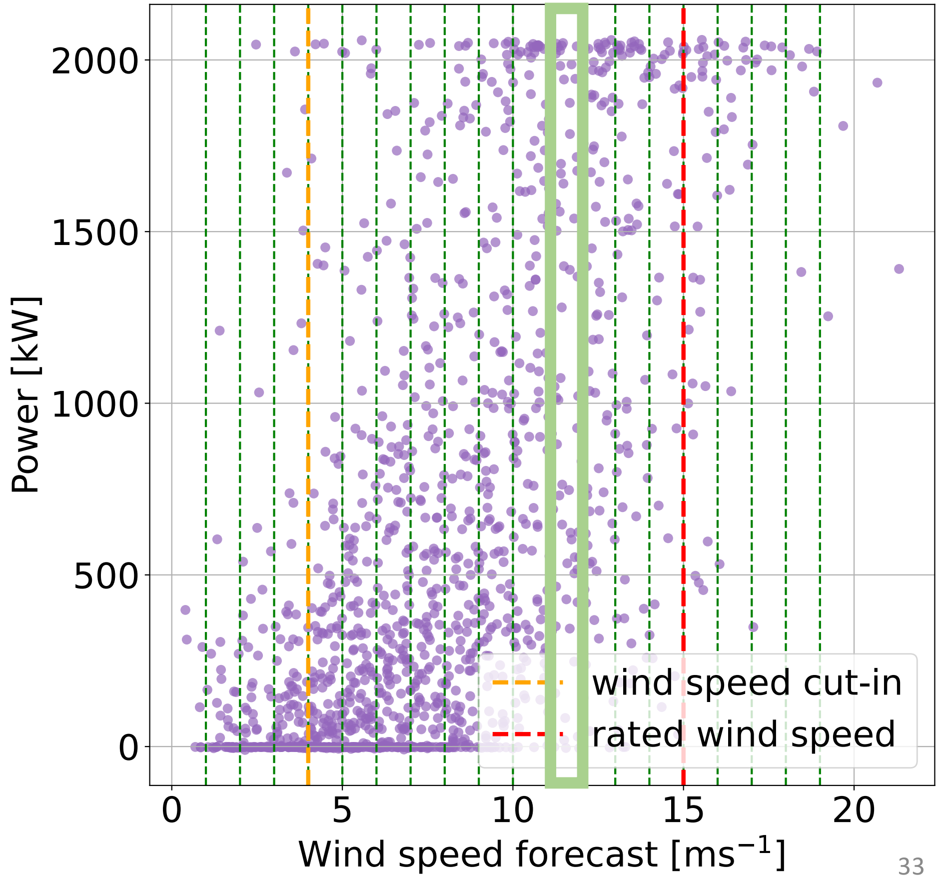
The challenge of heteroschedasticity and asymmetry



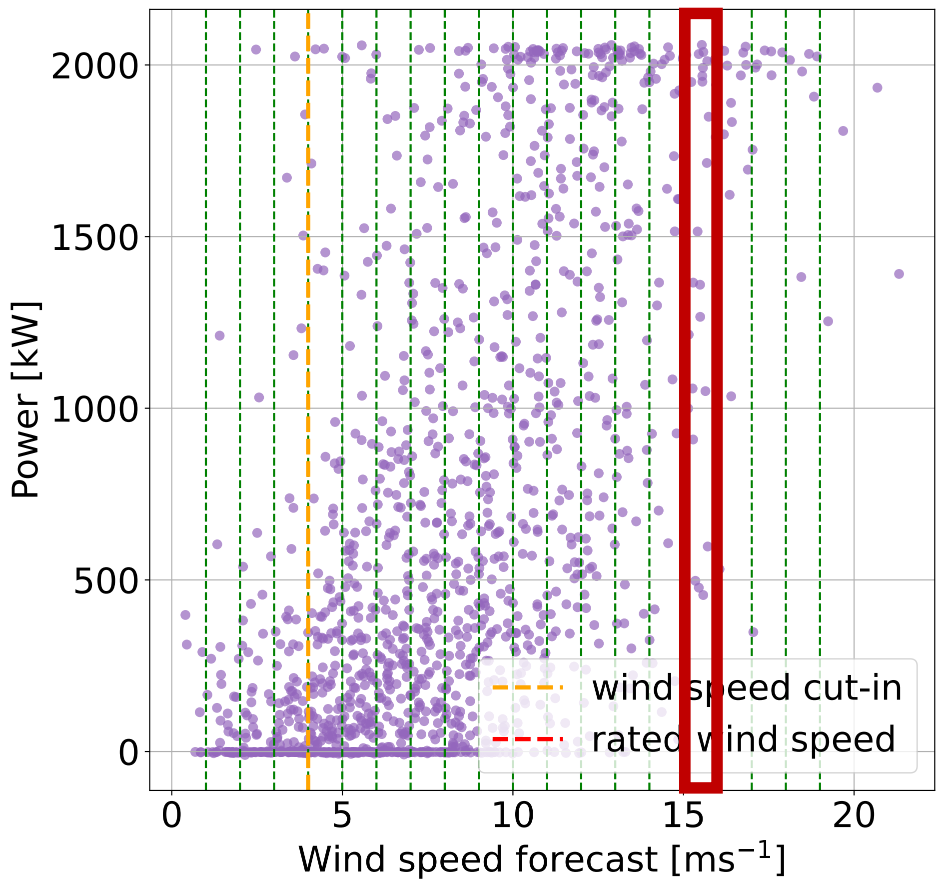
UNIVERSITÀ DI PAVIA

The challenge of heteroschedasticity and asymmetry



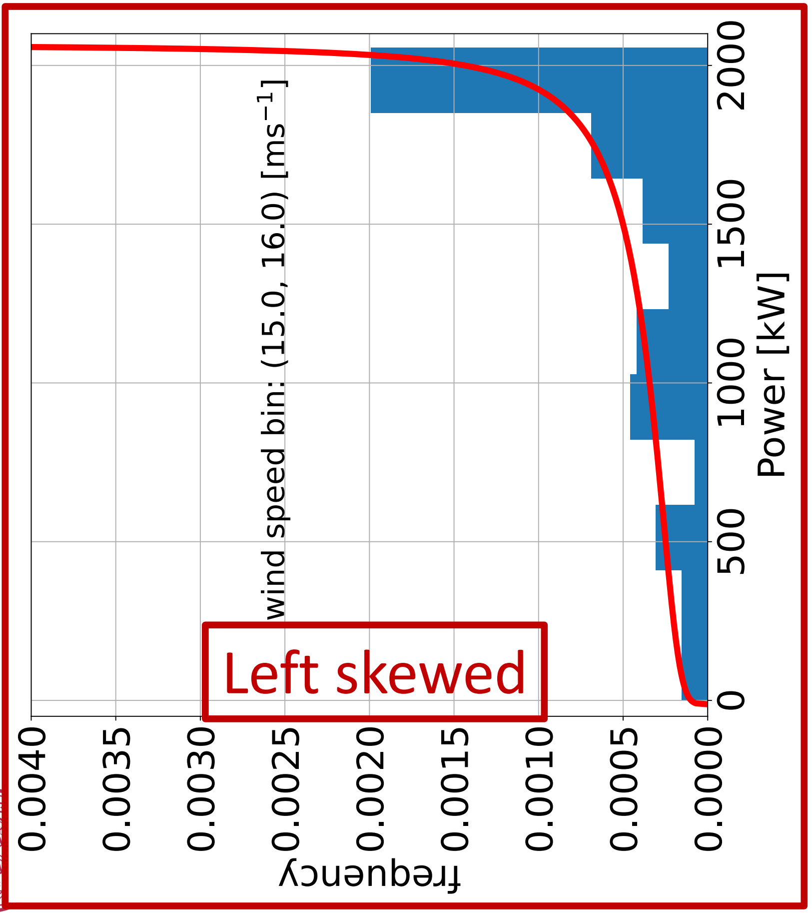


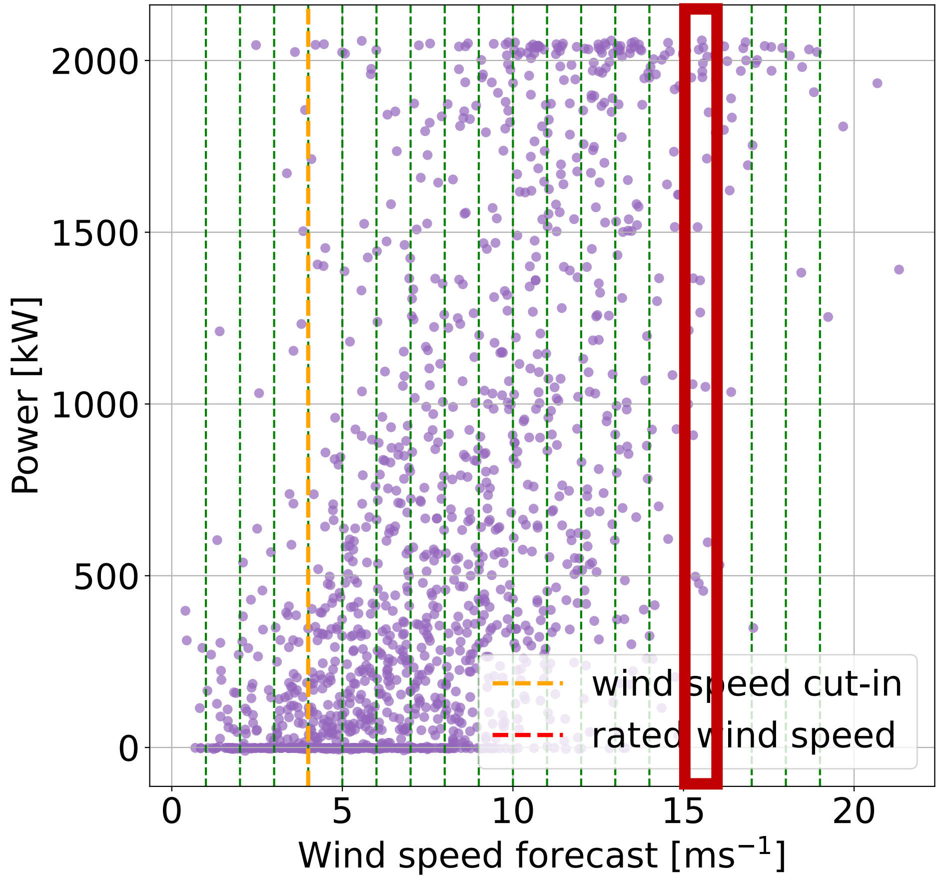
The challenge of heteroschedasticity and asymmetry



UNIVERSITÀ DI PAVIA

The challenge of heteroschedasticity and asymmetry





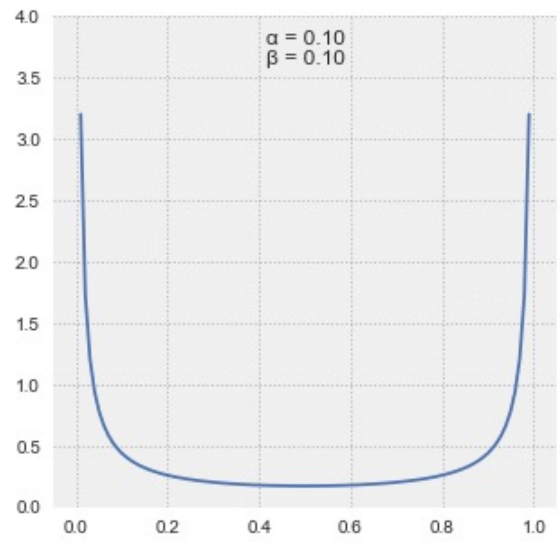
The challenge of heteroschedasticity and asymmetry

We have see that the distribution of wind power conditional on wind speed forecasts is heteroschedastic and asymmetric with skewness changing its sign. How to deal with it? Should we resort to a non parametric approach?

UNIVERSITÀ DI PAVIA

New idea: Beta regression to cope with asymmetrically distributed errors

- Data distribution naturally bounded between zero and the maximum output power of the turbine, further supporting the inadequacy of the Gaussian distribution
- A distribution that best suits this type of data is the Beta distribution
- Parameter optimization is performed via Beta regression (rather than standard LS)



UNIVERSITÀ DI PAVIA

Summary

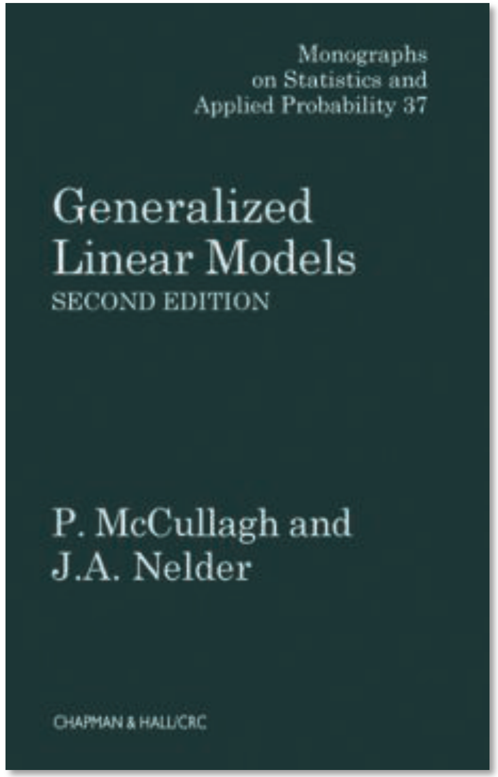
1. Introduction
2. The challenge of heteroschedasticity and asymmetry
3. **Beta Regression Model with preconditioning**
4. Results

UNIVERSITÀ
DI PAVIA

Generalized Linear Models (GLM)

- Generalized Linear Model: allows for nonlinearity while preserving simplicity and interpretability of linear models

- The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function



UNIVERSITÀ
DI PAVIA

Generalized Linear Model in a nutshell

Three components:

1. Linear predictor
2. Non linear link function $g(\cdot)$
3. Probability distribution (*Beta*)

UNIVERSITÀ
DI PAVIA

Generalized Linear Model in a nutshell

Three components:

1. Linear predictor
2. Non linear link function $g(\cdot)$
3. Probability distribution (*Beta*)

θ_0, θ_1: Model parameters

X_i :Wind speed

$$g(\zeta_i) = \theta_0 + \theta_1 X_i$$

$$y_i \sim Beta(\zeta_i)$$

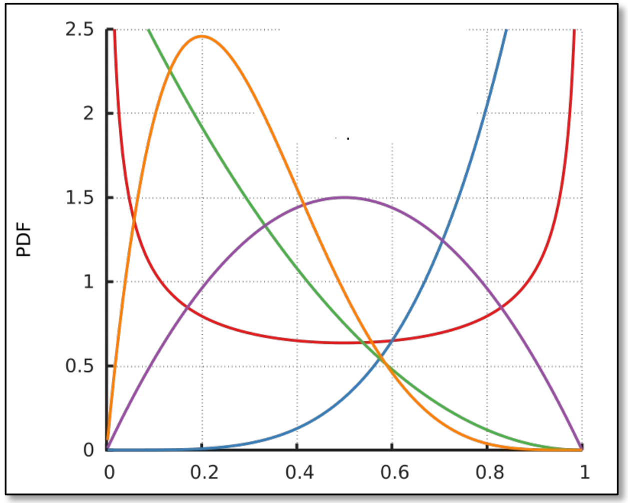
y_i : Actual power

UNIVERSITÀ DI PAVIA

Beta error model – constant dispersion ϕ

From the Beta distribution:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

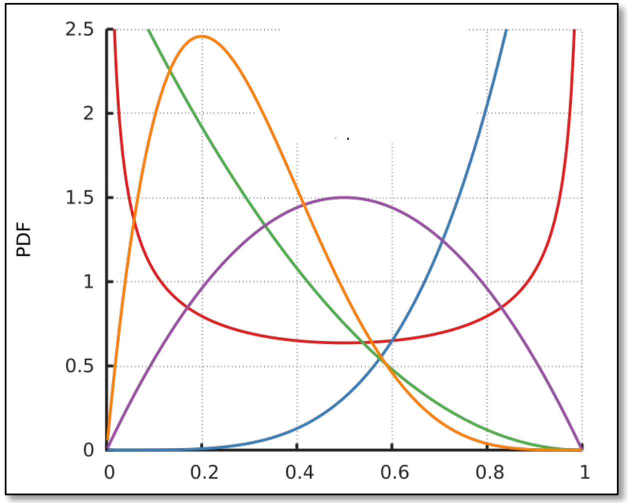


UNIVERSITÀ DI PAVIA

Beta error model – constant dispersion ϕ

Where:

- μ: mean of the Beta distribution
- ϕ: precision parameter of the Beta distribution
- $\Gamma(\cdot)$: the gamma function
- $g(\mu_i) = \theta_0 + \theta_1 x_i$

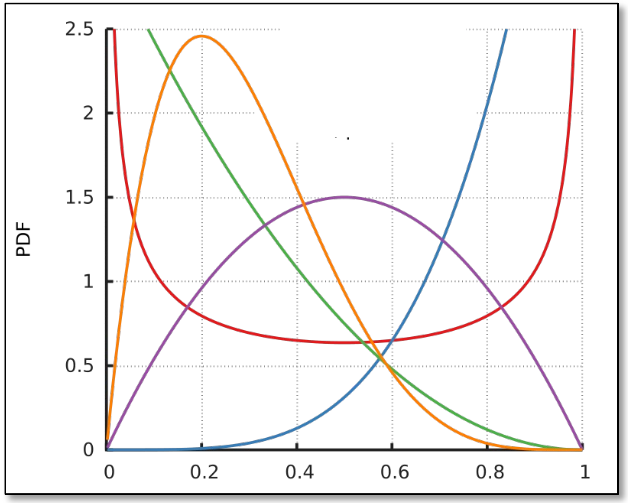


UNIVERSITÀ
DI PAVIA

Beta error model – constant dispersion ϕ

We can compute the expected value $\mathbb{E}(\cdot)$ and the variance $Var(\cdot)$ as:

- $\mathbb{E}(y_i) = \mu_i = g^{-1}(\theta_0 + \theta_1 x_i)$
- $Var(y_i) = \frac{\mu_i(1-\mu_i)}{(1-\phi)}$, *constant* ϕ



UNIVERSITÀ DI PAVIA

Beta Regression Model with preconditioning

- Typical choices of the link function $g(\cdot)$ are the logistic or the double exponential functions

- Neither the logistic nor the double exponential are able to adequately models the wind power curve data

- **New hybrid approach:** use a power curve initially obtained with a parametric or non parametric method as a preconditioner

UNIVERSITÀ
DI PAVIA

Summary

1. Introduction
2. The challenge of heteroschedasticity and asymmetry
3. Beta Regression Model with preconditioning
4. Results

UNIVERSITÀ
DI PAVIA

Results

We compared two Beta regression models with constant and variable dispersion with three "naive" forecasting strategies and a very flexible non parametric approach:

1. Persistence model
2. Enhanced Persistence
3. Open loop model
4. Quantile Regression Forest (QRF)

1. Persistence model

The persistence forecasting method assumes that the future predicted power $\hat{y}(t+k)$ will be the same as the current observed value $y(t)$:

$$\hat{y}(t+k) = y(t)$$

- Advantages: Simple, requires no training data, and often effective for short-term forecasts.
- Limitations: Accuracy decreases with longer forecast horizons and in highly variable conditions.

UNIVERSITÀ
DI PAVIA

2. Enhanced Persistence

Improved Method: Combines persistence forecasting with an autoregressive model of order 1 (AR(1)) on residuals:

Step 1: Apply persistence model: $\hat{y}(t+1) = y(t)$

Step 2: Calculate residuals: $e(t) = \hat{y}(t) - y(t)$

Step 3: Apply AR(1) model on residuals: $\hat{e}(t+1) = \phi e(t)$,

ϕ: AR(1) coefficient

Final Prediction: $\hat{y}_{enhan}(t+1) = y(t) + \phi e(t) = \boldsymbol{(1+\phi)y(t) - \phi y(t-1)}$

UNIVERSITÀ DI PAVIA

2. Enhanced Persistence

Improved Method: Combines persistence forecasting with an autoregressive model of order 1 (AR(1)) on residuals:

Step 1: Apply persistence moc

Step 2: Calculate residuals: $e($

$\hat{y}_{enhan}$ is a weighted mean of $y(t)$ and $y(t-1)$

Step 3: Apply AR(1) model on residuals: $\hat{e}(t+1) = \phi e(t)$,

ϕ: AR(1) coefficient

Final Prediction: $\hat{y}_{enhan}(t+1) = y(t) + \phi e(t) = \boldsymbol{(1+\phi)y(t) - \phi y(t-1)}$

3. Open loop model

- Overview:
  - Utilizes a periodic annual Weibull model on wind speed measurement data.
  - Identifies the model by solving a maximum likelihood problem.
  - Calculates the median of the model for the day of interest.
  - Uses the median as input to the manufacturer's power curve model or to a power curve identified on wind speed and power actual data.

3. Open loop model

Weibull Distribution:

$$f(x; shape, scale) = \frac{shape * \left(\frac{x}{scale}\right)^{shape-1} * e^{-\left(\frac{x}{scale}\right)^{shape}}}{scale}$$

UNIVERSITÀ DI PAVIA

3. Open loop model

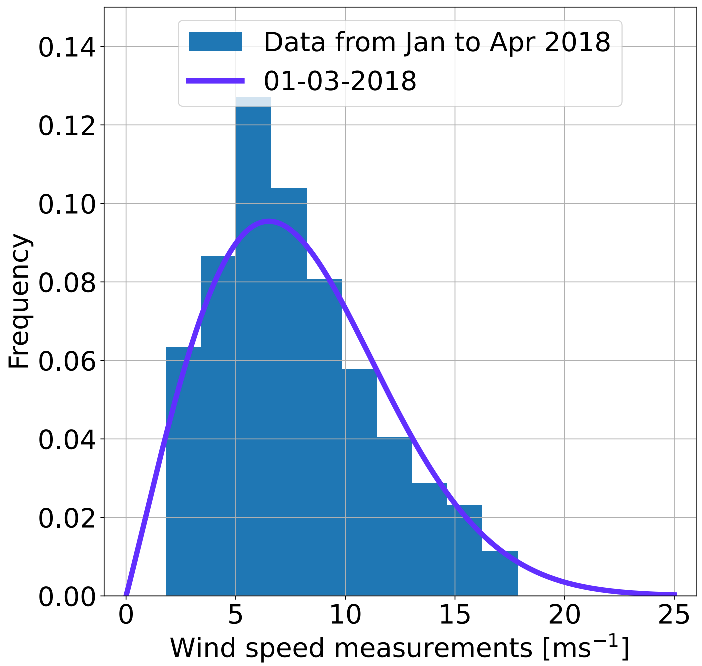
Shape and scale are periodic models depending on the day of the year:

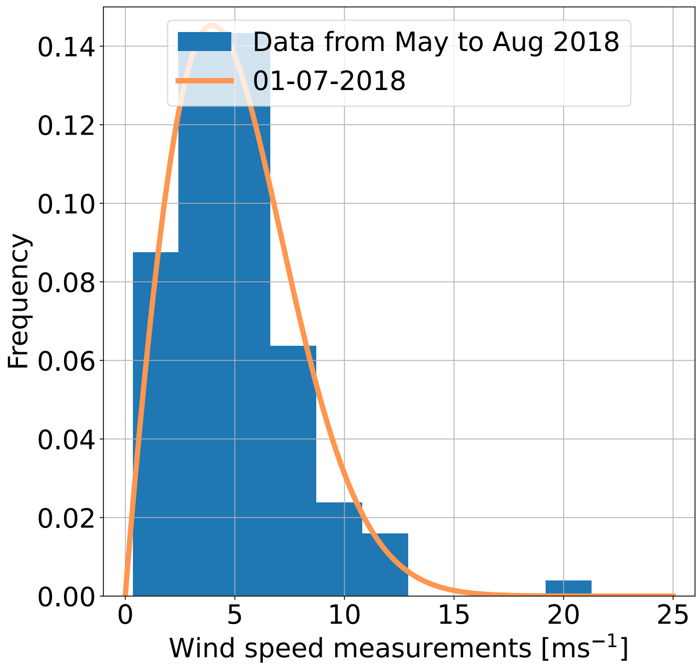
$$\beta_0 + \beta_1 * \sin\left(\frac{\pi * day\_of\_year}{T}\right) + \beta_2 * \cos\left(\frac{\pi * day\_of\_year}{T}\right) +$$

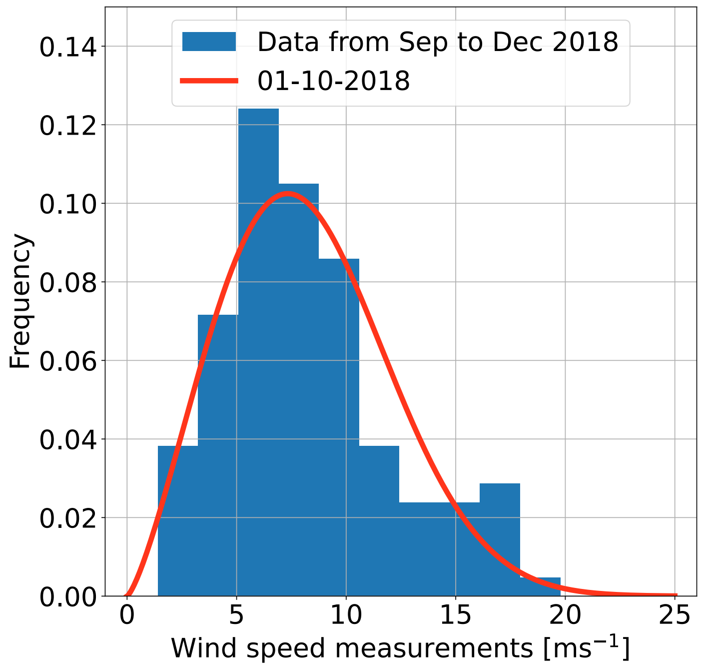
$$+ \beta_3 * \sin\left(\frac{2\pi * day\_of\_year}{T}\right) + \beta_4 * \cos\left(\frac{2\pi * day\_of\_year}{T}\right)$$

UNIVERSITÀ DI PAVIA

3. Open loop model

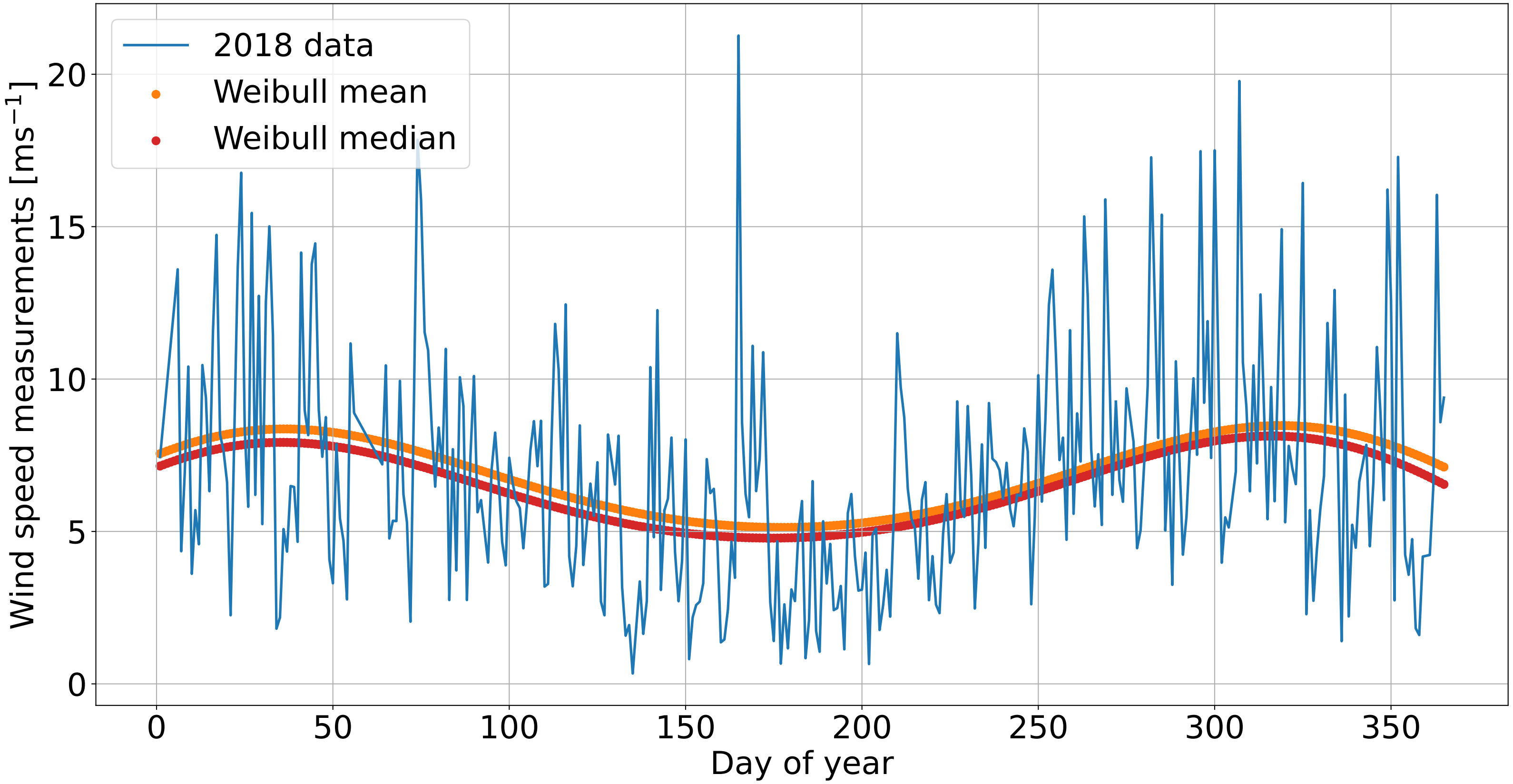






UNIVERSITÀ DI PAVIA

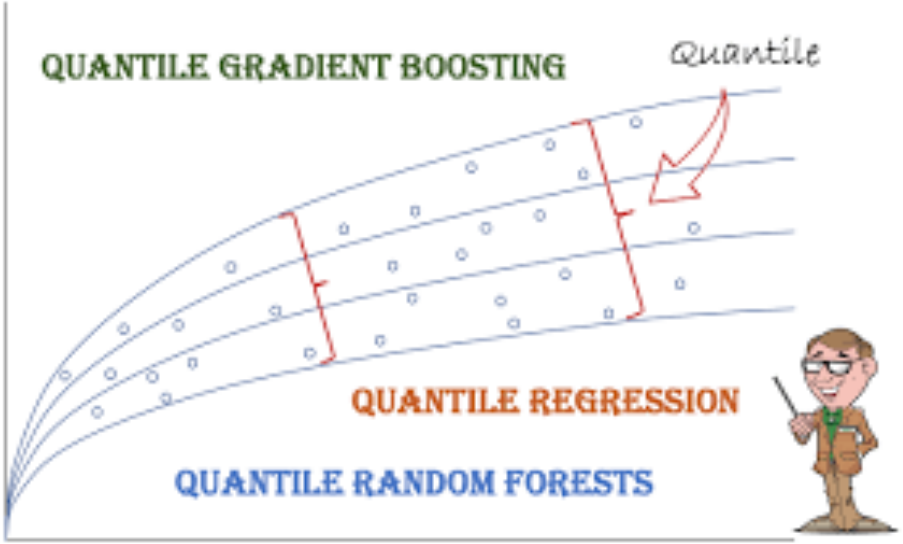
3. Open loop model



UNIVE
DI PAVIA

4. Quantile regression forest

- QRFs, an evolution of random forests, focus on estimating conditional quantiles, offering insights into response variable distributions.
- QRF requires careful hyperparameter tuning to avoid overfitting.
- A cross-validation procedure was conducted using the Python Optuna toolbox to optimize the model's hyperparameters.



UNIVERSITÀ
DI PAVIA

Performance indices

The following performance indices were used:

1. Weighted Mean Absolute Percentage Error (WMAPE):

$$\mathrm{WMAPE\%} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|} \times 100$$

$\bar{y}$: arithmetic mean of y_i
$\overline{\hat{y}_i}$: arithmetic mean of $\hat{y}_i$

2. Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

UNIVERSITÀ
DI PAVIA

Performance indices

3. Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$\bar{y}$: arithmetic mean of y_i
$\bar{\hat{y}}_i$: arithmetic mean of $\hat{y}_i$

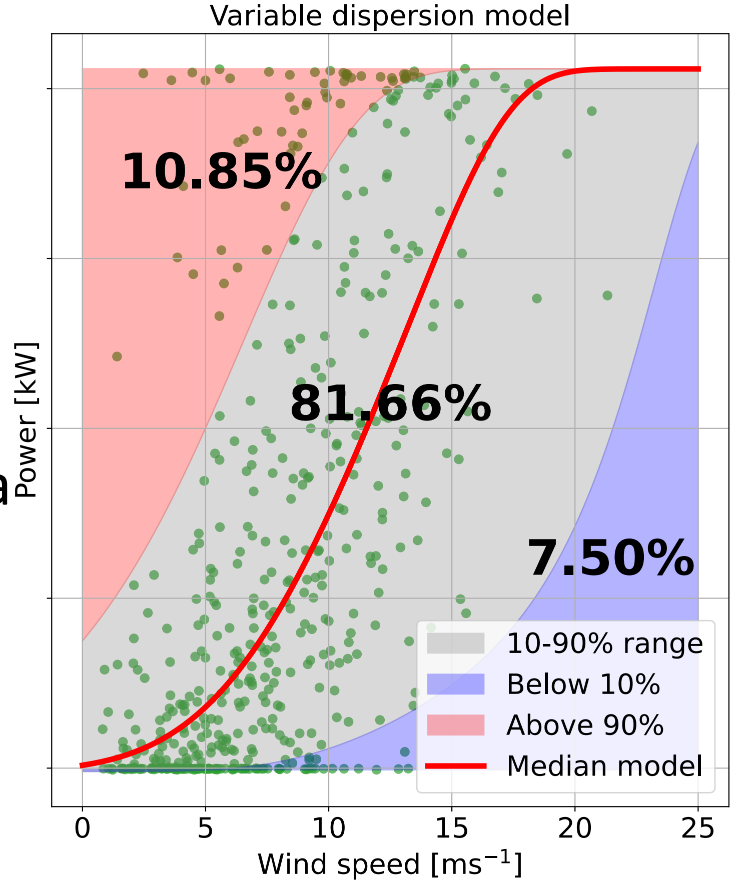
4. Coefficient of Determination (R^2):

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}\right)^2 \times 100$$

UNIVERSITÀ DI PAVIA

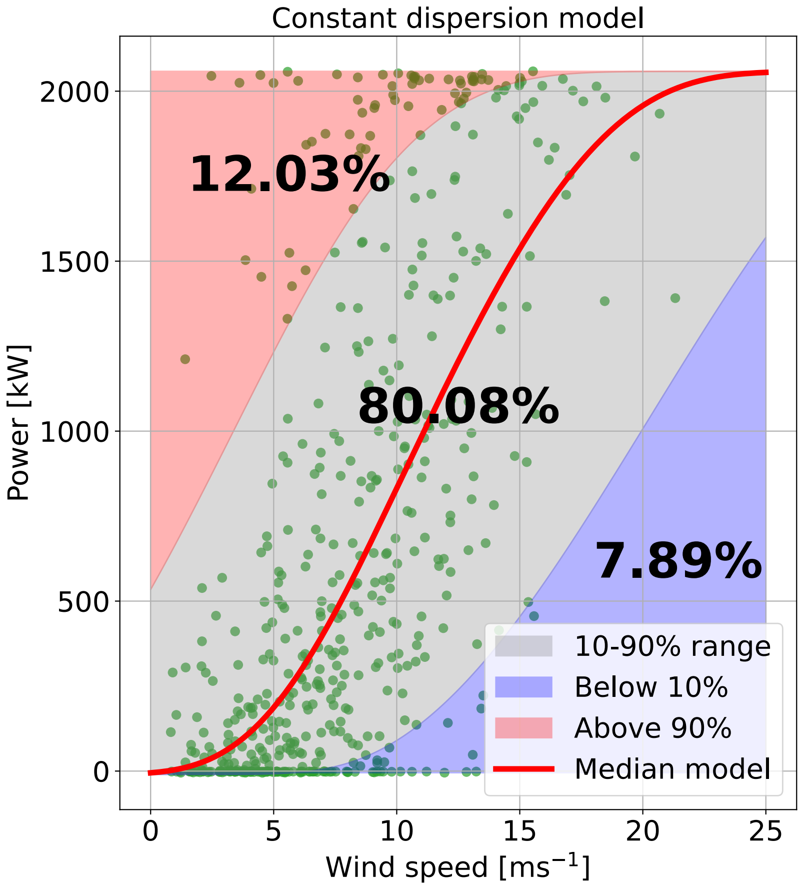
Results on test data

The Variable Dispersion Beta Regression Model performed best in terms of WMAPE and MAE, and effectively characterizes both training and test data distributions



UNIVERSITÀ DI PAVIA

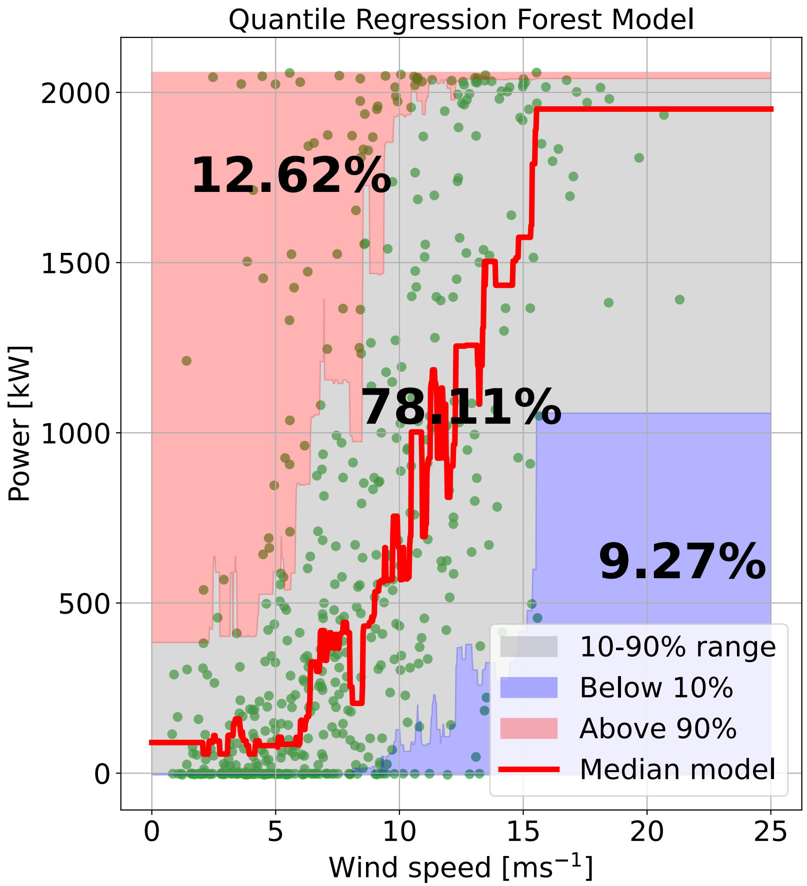
Results on test data



The Constant Dispersion Beta Regression Model performed well, effectively describing data distribution and providing a simpler solution

UNIVERSITÀ DI PAVIA

Results on test data



- The quantile regression forest achieved performance comparable to that of the Beta regression.
- Despite using a cross-validation procedure, the resulting model still exhibits some overfitting.

Results

Model	WMAPE (%)		MAE (kW)		RMSE (kW)		R^2(%)	
	Train	Test	Train	Test	Train	Test	Train	Test
Persistence	**93.72**	**85.17**	**599.50**	**578.34**	**821.19**	**813.53**	**5.61**	12.45
Enhanced P.	85.99	79.96	549.49	543.00	723.57	743.46	7.63	13.09
Open Loop	78.14	77.85	499.26	528.70	656.75	711.25	9.35	**12.08**
QRF	**59.92**	58.32	**382.86**	396.07	**535.14**	584.53	**39.42**	38.70
Const. Beta	63.75	58.72	407.32	398.75	535.64	**559.39**	36.08	**40.36**
Variab. Beta	62.99	**58.31**	402.44	**395.98**	540.55	566.46	35.79	39.86

The weakest performances came from the three naive models, especially the persistence approach

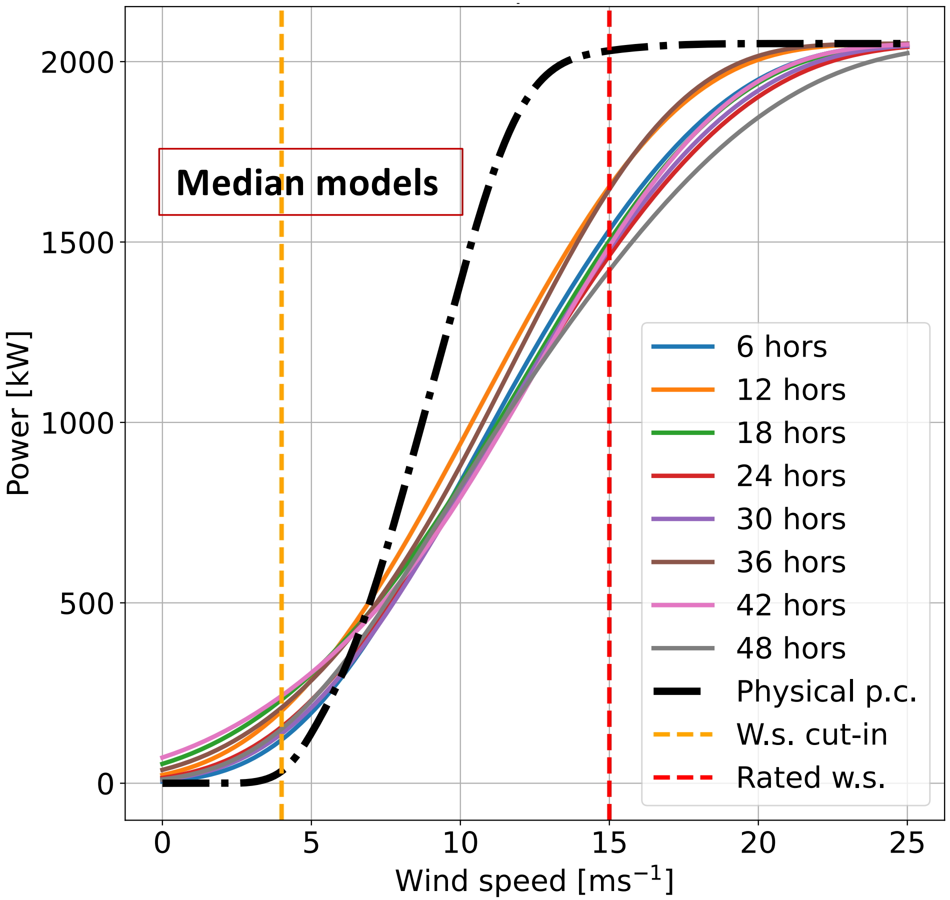
UNIVERSITÀ DI PAVIA

Results

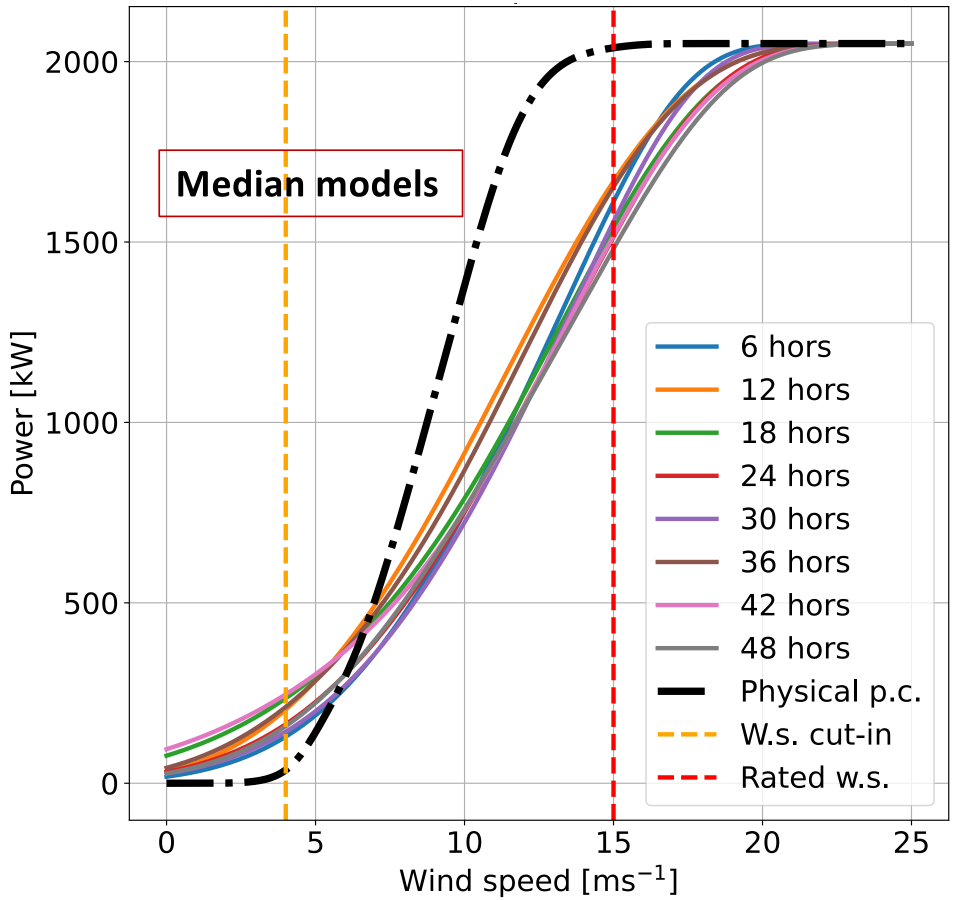
What happens when we change forecasting horizon?

Results

Costant dispersion models



Variable dispersion models



Results

DEPLOYED

Probabilistic models are useful in optimizing bidding strategies aiming to maximize profit in uncertain scenarios.

key takeaways include:

- Despite the complexity of power distribution as wind speed varies, the Beta regression model appears to adequately characterize all the different shapes of the distribution.
- An alternative non-parametric method is the quantile regression forest, though it requires more careful hyperparameter tuning and is less interpretable compared to Beta regression models.
- Both the Beta regression approach and the quantile regression forest outperformed naive approaches.

For any question:

marco.capelletti02@universitadipavia.it

Thanks for your attention