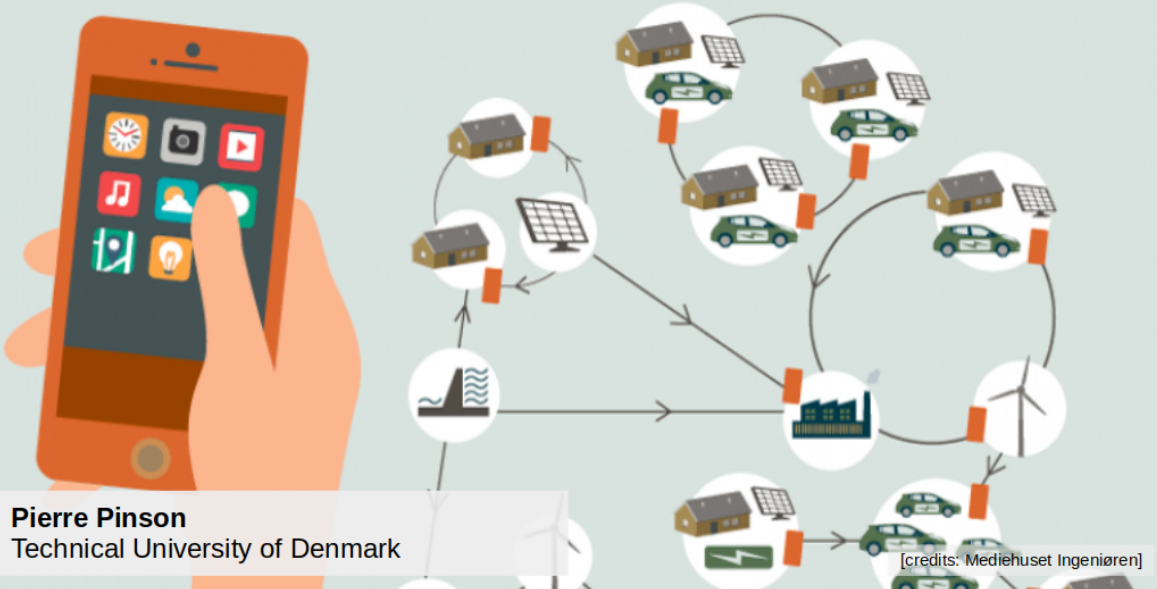
Module 10 – Renewable Energy Forecasting: Advanced Topics

10.3 Data-driven decisions



Pierre Pinson
Technical University of Denmark

[credits: Mediehuset Ingeniøren]

What decisions are we talking about?

- The more advanced the modelling approach, the more likely you need to decide on so-called **meta-parameters**, e.g.
  - order of polynomial regression
  - fitting points and bandwidths for local polynomial regression
  - window size when learning over sliding windows
  - forgetting factor (/learning rate) in online learning
  - etc.

What decisions are we talking about?

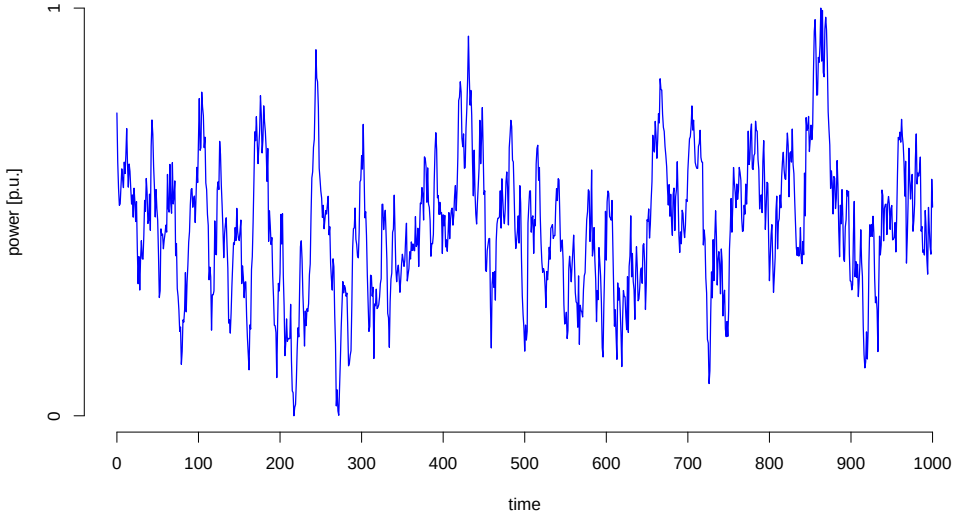
- The more advanced the modelling approach, the more likely you need to decide on so-called **meta-parameters**, e.g.
  - order of polynomial regression
  - fitting points and bandwidths for local polynomial regression
  - window size when learning over sliding windows
  - forgetting factor (/learning rate) in online learning
  - etc.

- And you may also want to see whether a **different model**, or the use of **other explanatory variables**, would yield *higher forecast quality*

What decisions are we talking about?

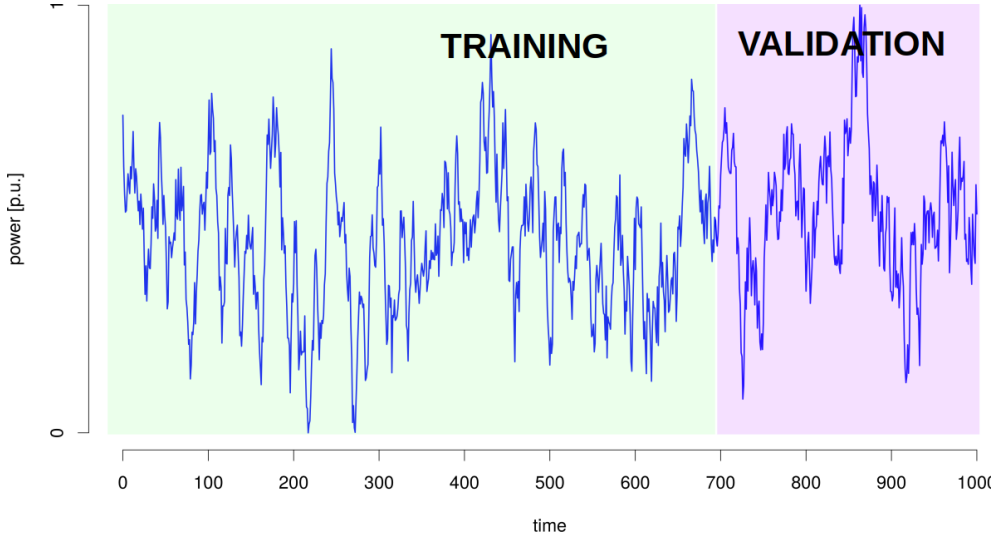
- The more advanced the modelling approach, the more likely you need to decide on so-called **meta-parameters**, e.g.
  - order of polynomial regression
  - fitting points and bandwidths for local polynomial regression
  - window size when learning over sliding windows
  - forgetting factor (/learning rate) in online learning
  - etc.
- And you may also want to see whether a **different model**, or the use of **other explanatory variables**, would yield *higher forecast quality*
- What could be a **well-thought strategy** to make those decisions?

Playing with available data



- We need to find smart ways to use those data to make decisions...

We can organize our own forecast competition



- A part of the dataset is used for any training, given choices about models, meta-parameters, etc.
- The remainder is used for genuine forecast evaluation

k-fold cross validation

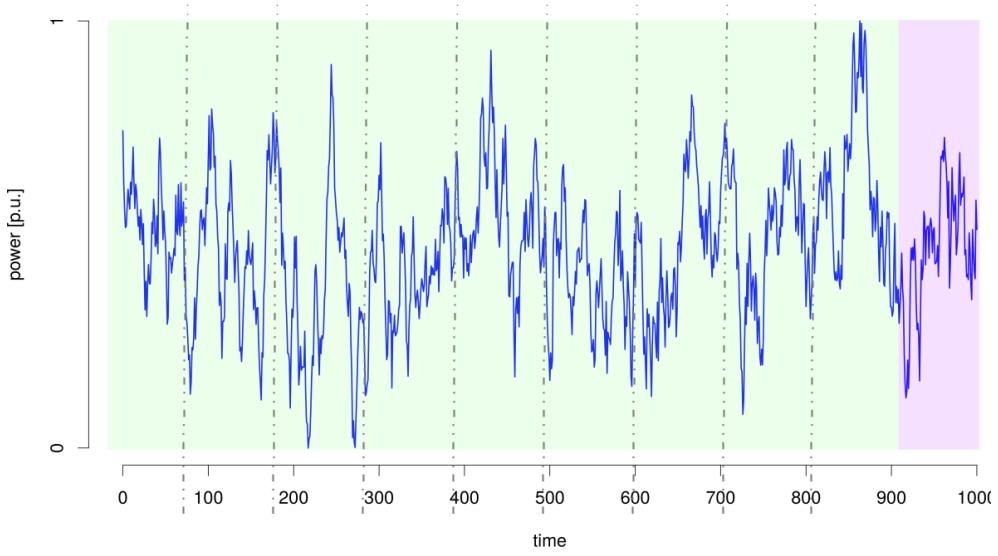
- The process of using part of the data for *training* and the remainder for *validation* is referred to as **cross-validation**

- For the case of k-fold cross validation:
  - Choose a reference score Sc e.g. RMSE
  - Divide the dataset into k parts of (app.) equal length
  - For a given choice of model M and meta-parameters τ,
    - There are k possible validation sets to be considered, $i = 1, \ldots, k$
    - For a given i, train over $k-1$ parts of data...
    - And calculate the score for the last remaining part of data $Sc_i(M, \tau)$
  - Calculate the average score $\overline{Sc}(M, \tau)$ for these k validation data blocks, i.e.

$$\overline{Sc}(M, \tau) = \frac{1}{k} \sum_{i=1}^{k} Sc_i(M, \tau)$$

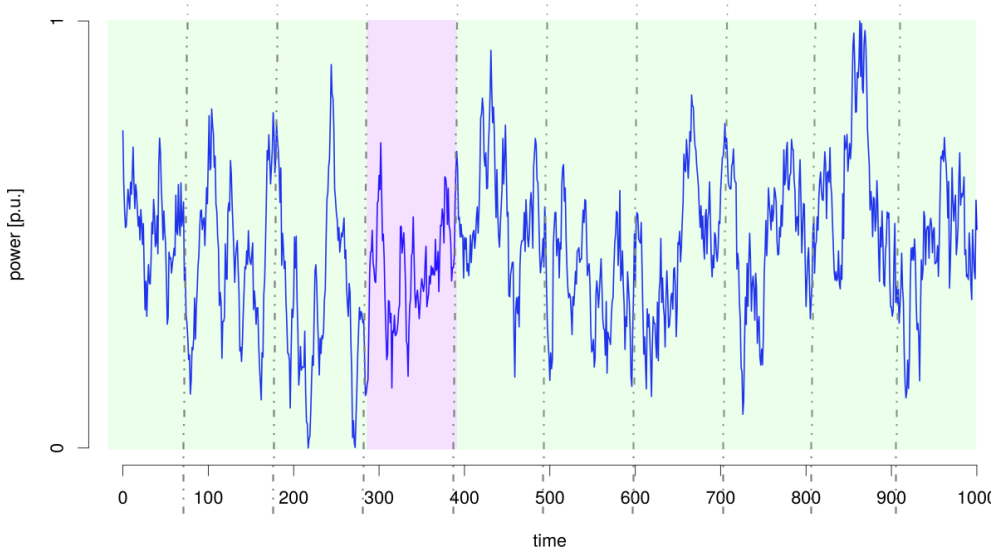
- $k = 10$ is the most common choice

Example: 10-fold cross-validation



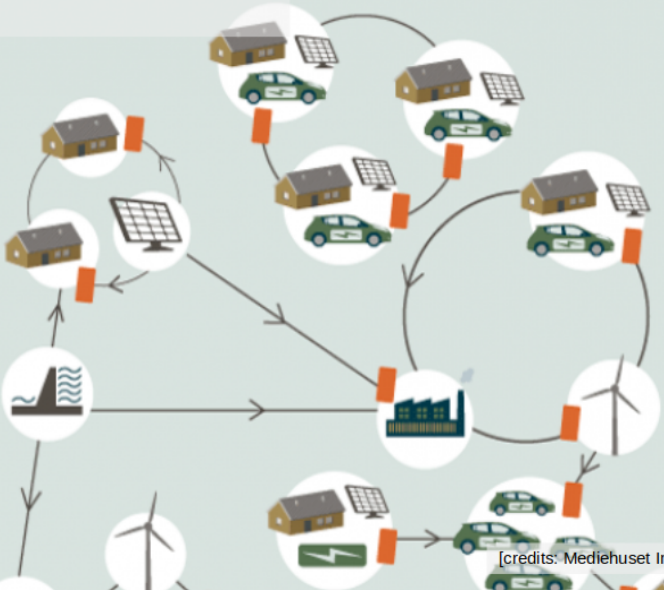
- Use the first 9 blocks of data to train, and the last one to perform genuine forecasting and calculate RMSE score - Obtain $Sc_{10}(M, \tau)$
- Repeat for all other possible combinations...

Example: 10-fold cross-validation



- E.g. when reaching $i = 4$, use the first 3 and last 6 blocks of data to train, and the 4^{th} one to perform genuine forecasting and calculate RMSE score - Obtain $Sc_4(M, \tau)$
- When finished with all $i = 1, \ldots, 10$, calculate the average score value $\overline{Sc}(M, \tau)$

Use the self-assessment quizz to check your understanding!



[credits: Mediehuset Ingeniøren]