On the importance of the long-term seasonal component in day-ahead electricity price forecasting. Part II - Probabilistic forecasting

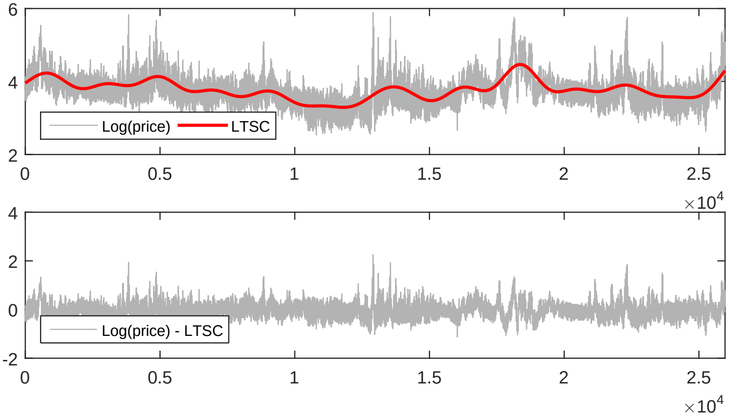
Rafał Weron*

Department of Operations Research
Wrocław University of Science and Technology (PWr), Poland
http://kbo.pwr.edu.pl/en/staff/rafal-weron/

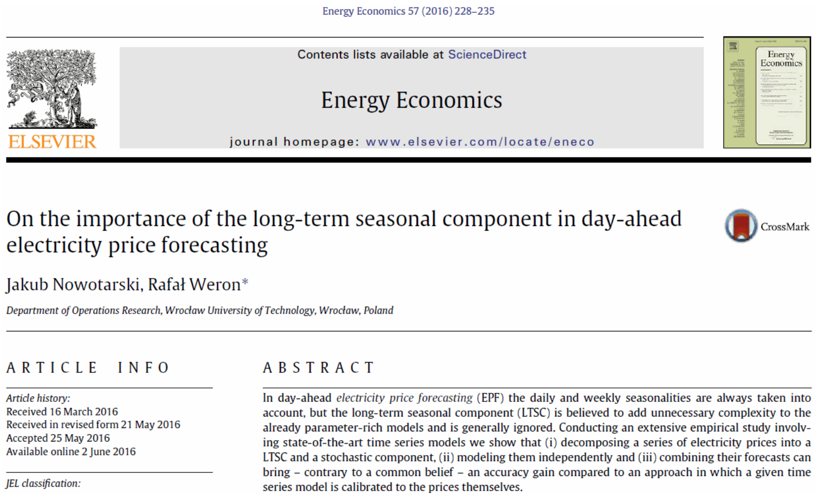
* Based on work with Grzegorz Marcjasz and Bartosz Uniejewski

LTSC and short-term price forecasting

- Can the long-term trend-seasonal component (LTSC) impact short-term (day-ahead) electricity price forecasts?



Point forecasting: Yes, it can!



Energy Economics 57 (2016) 228–235

Contents lists available at ScienceDirect

Energy Economics

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

On the importance of the long-term seasonal component in day-ahead electricity price forecasting

Jakub Nowotarski, Rafał Weron*

Department of Operations Research, Wrocław University of Technology, Wrocław, Poland

ARTICLE INFO

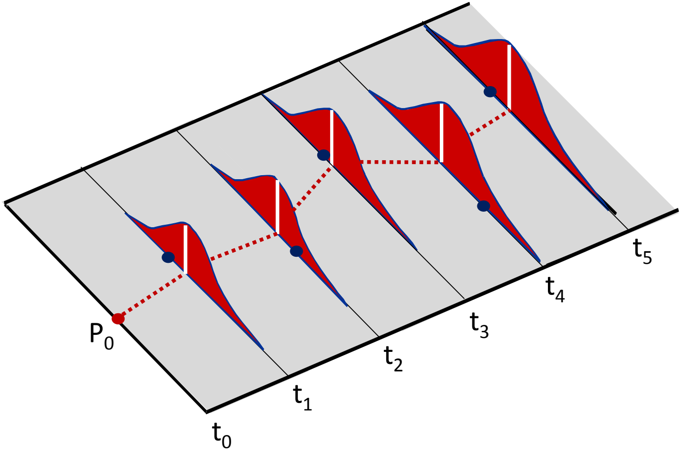
Article history:
Received 16 March 2016
Received in revised form 21 May 2016
Accepted 25 May 2016
Available online 2 June 2016

JEL classification:

ABSTRACT

In day-ahead *electricity price forecasting* (EPF) the daily and weekly seasonalities are always taken into account, but the long-term seasonal component (LTSC) is believed to add unnecessary complexity to the already parameter-rich models and is generally ignored. Conducting an extensive empirical study involving state-of-the-art time series models we show that (i) decomposing a series of electricity prices into a LTSC and a stochastic component, (ii) modeling them independently and (iii) combining their forecasts can bring – contrary to a common belief – an accuracy gain compared to an approach in which a given time series model is calibrated to the prices themselves.

What about probabilistic forecasts?

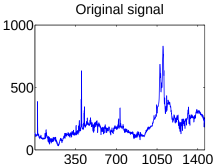


Agenda

1. Introduction
2. The *Seasonal Component* (SC) approach
   - Wavelets and the HP-filter
   - SCAR/SCANN models
3. Study design
   - Datasets
   - Probabilistic forecasts and the pinball score
4. Results
   - Combining SCAR models across LTSCs
   - Combining SCANN models across runs

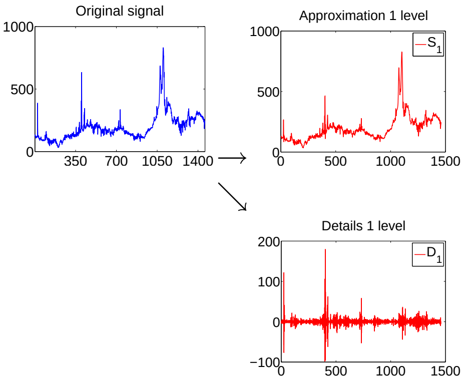
Wavelets

Decomposition of a signal into an approximation (a smoother) and a sequence of detail series: $A_J + D_J + D_{J-1} + \ldots + D_1$):



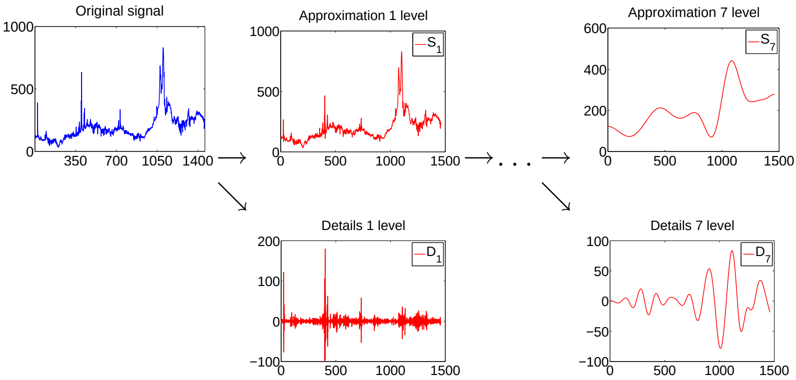
Wavelets

Decomposition of a signal into an approximation (a smoother) and a sequence of detail series: $A_J + D_J + D_{J-1} + \ldots + D_1$):



Wavelets

Decomposition of a signal into an approximation (a smoother) and a sequence of detail series: $A_J + D_J + D_{J-1} + \ldots + D_1$):



The Hodrick-Prescott (1980, 1997) filter

A simple alternative to wavelets

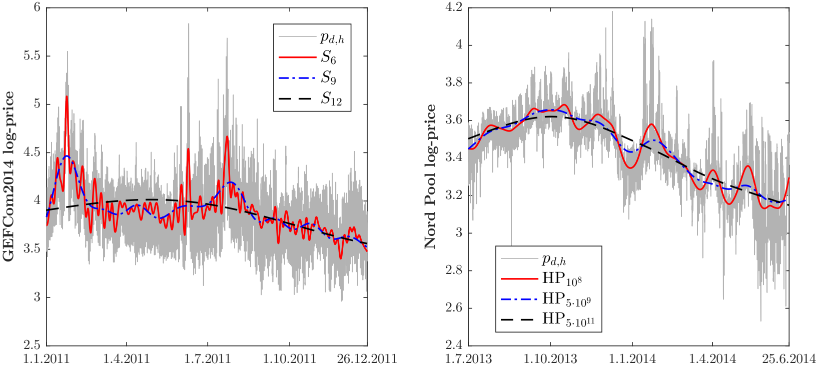
- Originally proposed for decomposing GDP into a long-term growth component and a cyclical component
- Returns a smoothed series τ_t for a noisy input series y_t:

$$\min_{\tau_t} \left\{ \sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \right]^2 \right\},$$

Punish for:

- deviating from the original series
- roughness of the smoothed series

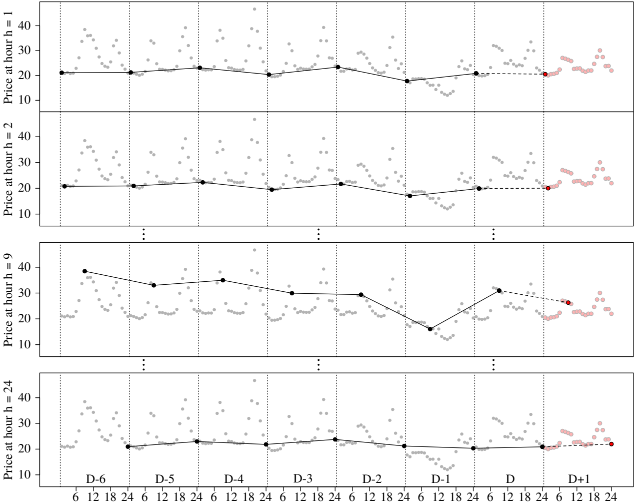
Wavelet and HP-filter based LTSCs



- **Wavelet filters (-S$_J$)**: $S_5, S_6, \ldots, S_{14}$, ranging from 'daily' smoothing ($S_5 \rightarrow 2^5$ hours) up to 'biannual' ($S_{14} \rightarrow 2^{14}$ hours)
- **HP-filters (-HP$_\lambda$)**: with $\lambda = 10^8, 5\cdot10^8, 10^9, \ldots, 5\cdot10^{11}$

A multivariate framework

(Ziel & Weron, 2018, ENEECO)



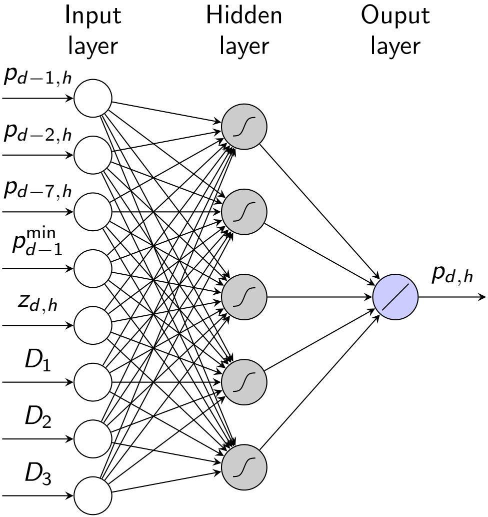
The **ARX** model

For the log-price, i.e., $p_{d,h} = log(P_{d,h})$, the model is given by:

$$p_{d,h} = \underbrace{\beta_{h,1} p_{d-1,h} + \beta_{h,2} p_{d-2,h} + \beta_{h,3} p_{d-7,h}}_{\text{autoregressive effects}} + \underbrace{\beta_{h,4} p_{d-1,\min}}_{\text{non-linear effect}} + \underbrace{\beta_{h,5} z_{d,h}}_{\text{load forecast}} + \underbrace{\sum_{i=1}^{3} \beta_{h,i+5} D_i}_{\text{Mon, Sat, Sun dummies}} + \varepsilon_{d,h} \quad (1)$$

- $p_{d-1,min}$ is yesterday's minimum hourly price
- z_t is the logarithm of system load/consumption
- Dummy variables D_1, D_2 and D_3 refer to Monday, Saturday and Sunday, respectively

A nonlinear alternative: The **ANN** model



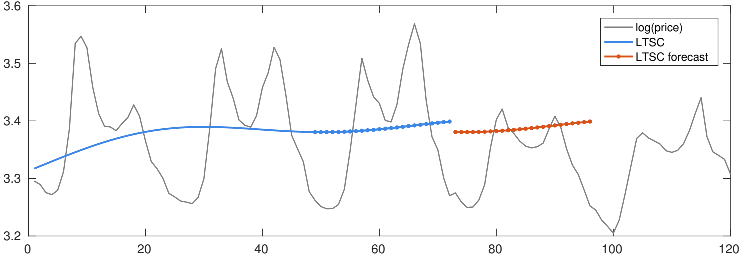
The SCAR/SCANN modeling framework

Nowotarski-Weron, 2016, ENEECO; Marcjasz et al., 2018, IJF; Uniejewski et al., 2018, ENEECO

The *Seasonal Component AutoRegressive* (SCAR) / *Artificial Neural Network* (SCANN) framework consists of the following steps:

1. (a) Decompose the (log-)price in the calibration window into the LTSC $T_{d,h}$ and the stochastic component $q_{d,h}$
   (b) Decompose the exogenous series in the calibration window using the same type of LTSC as for prices
2. Calibrate the **ARX** or **ANN** model to q_t and compute forecasts for the 24 hours of the next day (24 separate series)

The SCAR/SCANN modeling framework cont.



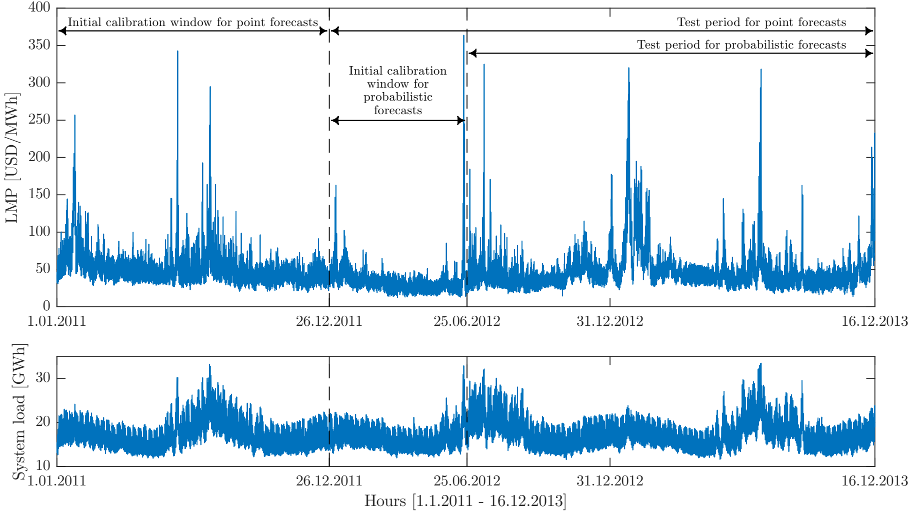
3. Add stochastic component forecasts $\hat{q}_{d+1,h}$ to persistent forecasts $\hat{T}_{d+1,h}$ of the LTSC to yield log-price forecasts $\hat{p}_{d+1,h}$
4. Convert them into price forecasts of the **SCAR** or **SCANN** model, i.e., $\hat{P}_{d+1,h} = \exp\left(\hat{p}_{d+1,h}\right)$

Agenda

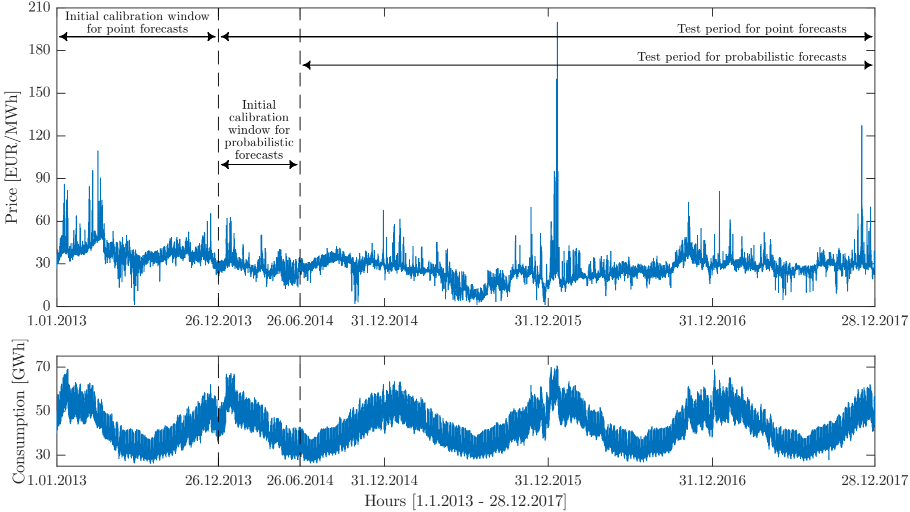
1. Introduction
2. The *Seasonal Component* (SC) approach
   - Wavelets and the HP-filter
   - SCAR/SCANN models
3. Study design
   - Datasets
   - Probabilistic forecasts and the pinball score
4. Results
   - Combining SCAR models across LTSCs
   - Combining SCANN models across runs

GEFCom2014 (2011-2013)

From the Global Energy Forecasting Competition



Nord Pool (2013-2017)



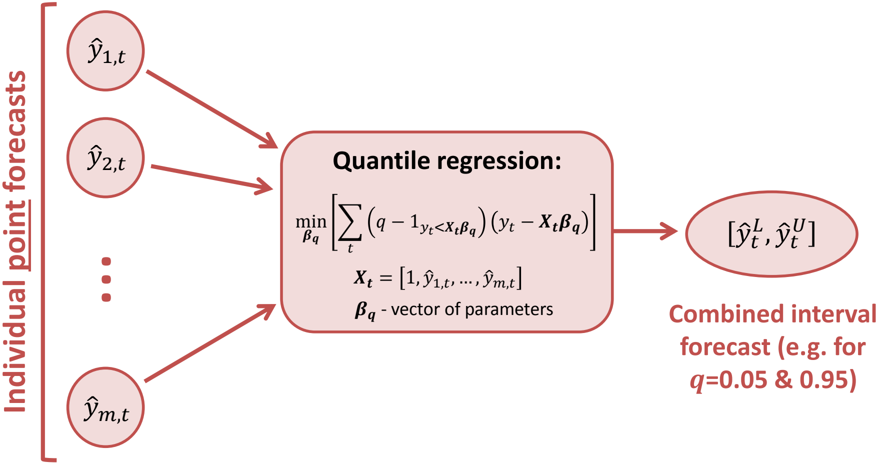
Three methods of constructing PIs

1. Historical simulation (**H**), which consists of computing sample quantiles of the empirical distribution of $\varepsilon_{d,h}$'s
2. Bootstrapping (**B**), which first generates pseudo-prices recursively using sampled normalized residuals, then computes desired quantiles of the bootstrapped prices
    - Takes into account not only historical forecast errors but also parameter uncertainty
3. Quantile Regression Averaging (**Q**)

Note: All require that one-day ahead point prediction errors are available in the calibration window for probabilistic forecasts

Quantile Regression Averaging (QRA): The idea

(Nowotarski & Weron, 2015, COST)



Individual point forecasts

$\hat{y}_{1,t}$

$\hat{y}_{2,t}$

$\vdots$

$\hat{y}_{m,t}$

Quantile regression:

$$\min_{\boldsymbol{\beta}_q} \left[\sum_t \left(q - 1_{y_t < \boldsymbol{X}_t \boldsymbol{\beta}_q} \right) \left(y_t - \boldsymbol{X}_t \boldsymbol{\beta}_q \right) \right]$$

$$\boldsymbol{X}_t = \left[1, \hat{y}_{1,t}, \dots, \hat{y}_{m,t} \right]$$

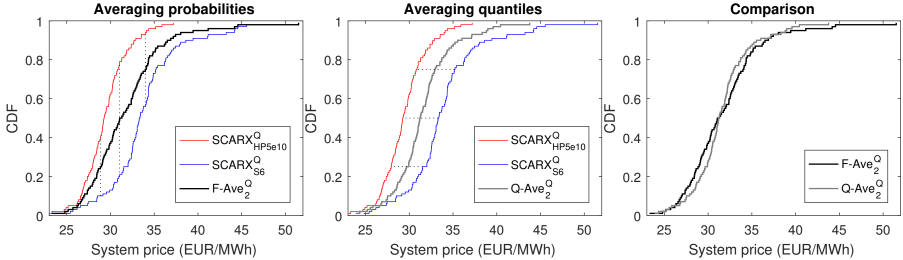
$\boldsymbol{\beta}_q$ - vector of parameters

$[\hat{y}_t^L, \hat{y}_t^U]$

Combined interval forecast (e.g. for *q*=0.05 & 0.95)

Combining probabilistic forecasts

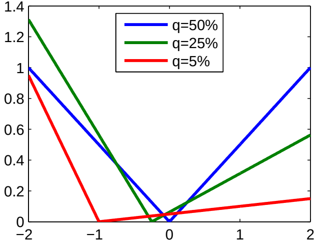
- Average probability forecast: **F-Ave**$^*_n \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{F}_i(x)$
  $\Rightarrow$ a vertical average of predictive distributions
- Average quantile forecast: **Q-Ave**$^*_n \equiv \hat{Q}^{-1}(x)$
  with $\hat{Q}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}_i(x)$ and quantile forecast $\hat{Q}_i(x) = \hat{F}_i^{-1}(x)$
  $\Rightarrow$ a horizontal average
- $* =$ **H**, **B** or **Q** denotes the method of constructing PIs



Sharpness and the pinball loss

$$\textbf{Pinball}\left(\hat{Q}_{P_t}(q), P_t, q\right) = \begin{cases} (1-q)\left(\hat{Q}_{P_t}(q) - P_t\right), & \text{for } P_t < \hat{Q}_{P_t}(q), \\ q\left(P_t - \hat{Q}_{P_t}(q)\right), & \text{for } P_t \geqslant \hat{Q}_{P_t}(q), \end{cases}$$

- $\hat{Q}_{P_t}(q)$ is the price forecast at the q-th quantile
- P_t is the actually observed price
- To provide an aggregate score we average:
  - across all hours in the test period
  - across 99 percentiles



Diebold-Mariano (DM) tests

Define the 'multivariate' loss differential series in the $\| \cdot \|_1$-norm as:

$$\Delta_{X,Y,d} = \|\pi_{X,d}\|_1 - \|\pi_{Y,d}\|_1$$

where

- $\pi_{X,d} = (\pi_{X,d,1}, \ldots, \pi_{X,d,24})'$ is the vector of pinball scores for model X and day d
- $\|\pi_{X,d}\|_1 = \sum_{h=1}^{24} |\pi_{X,d,h}|$ is the average across the 24 hours
- As in the standard DM test, we assume that the loss differential series is covariance stationary

For each model pair we compute two one-sided DM tests:

1. $H_0 : E(\Delta_{X,Y,d}) \leqslant 0 \Rightarrow$ X yields better forecasts
2. $H_0^R : E(\Delta_{X,Y,d}) \geqslant 0 \Rightarrow$ Y yields better forecasts

Agenda

1. Introduction
2. The *Seasonal Component* (SC) approach
   - Wavelets and the HP-filter
   - SCAR/SCANN models
3. Study design
   - Datasets
   - Probabilistic forecasts and the pinball score
4. Results
   - Combining SCAR models across LTSCs
   - Combining SCANN models across runs

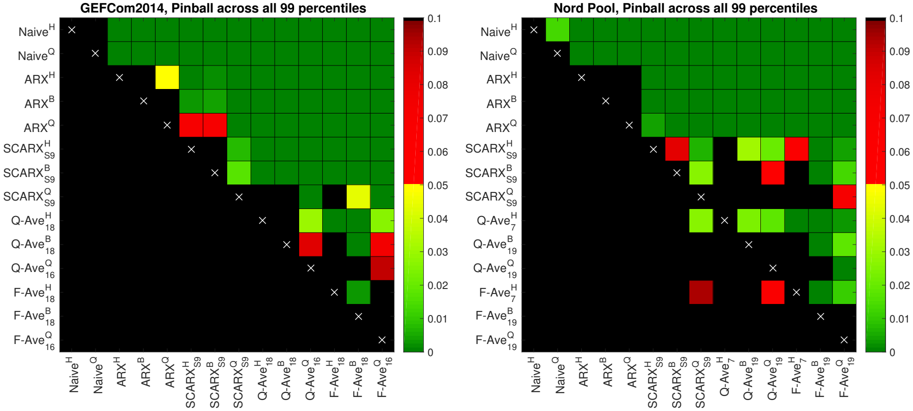
Combining SCAR models across LTSCs

(Uniejewski et al., 2018, ENEECO)

We present results for 14 selected ARX-type models:

- Both naive benchmarks – $\textbf{Naive}^{\textbf{H}}$, $\textbf{Naive}^{\textbf{Q}}$
- All three ARX benchmarks – $\textbf{ARX}^{\textbf{H}}$, $\textbf{ARX}^{\textbf{B}}$, $\textbf{ARX}^{\textbf{Q}}$
- The best *ex-post*
  - $\textbf{SCARX}^{\textbf{H}}_{*}$, $\textbf{SCARX}^{\textbf{B}}_{*}$ and $\textbf{SCARX}^{\textbf{Q}}_{*}$ models
  - $\textbf{Q-Ave}^{\textbf{H}}_{n}$, $\textbf{Q-Ave}^{\textbf{B}}_{n}$ and $\textbf{Q-Ave}^{\textbf{Q}}_{n}$ average quantile forecasts
  - $\textbf{F-Ave}^{\textbf{H}}_{n}$, $\textbf{F-Ave}^{\textbf{B}}_{n}$ and $\textbf{F-Ave}^{\textbf{Q}}_{n}$ average probability forecasts

p-values of the DM test across 99 percentiles



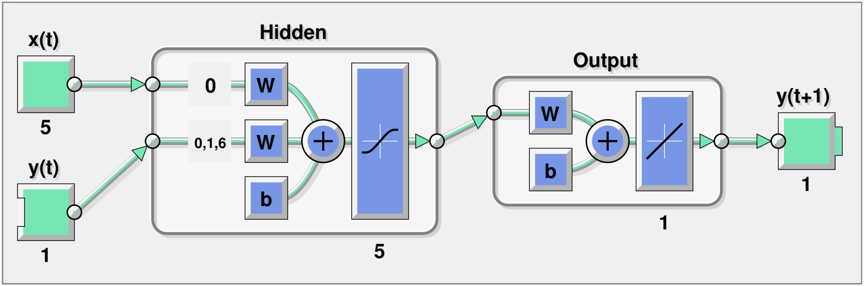
We use a heat map to indicate the range of the *p*-values – the closer they are to zero (→ **dark green**) the more significant is the difference between the forecasts of a model on the X-axis (better) and the forecasts of a model on the Y-axis (worse)

Main findings

- ‘Probabilistic’ **SCARX** models (nearly always) significantly outperform the **Naive** and **ARX** benchmarks
  - $\mathbf{SCARX^{Q}}$ models (nearly always) significantly outperform $\mathbf{SCARX^{H}}$ and $\mathbf{SCARX^{B}}$
- Both averaging schemes generally significantly outperform the benchmarks and the non-combined **SCARX** models
- Averaging over probabilities ($\mathbf{F\text{-}Ave}_n^*$) generally yields better probabilistic EPFs than averaging over quantiles ($\mathbf{Q\text{-}Ave}_n^*$)
  - In contrast to typically encountered economic forecasting problems (Lichtendahl et al., 2013, Mgnt Sci.)

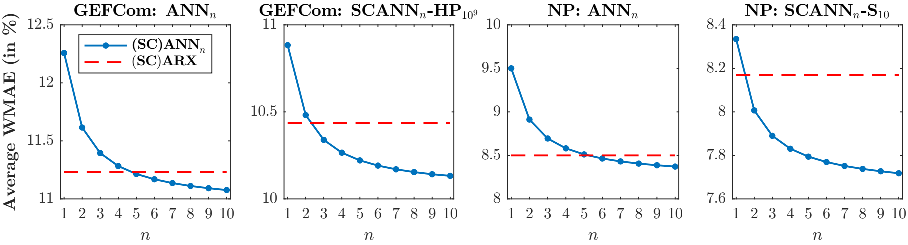
The SCANN modeling framework

(Marcjasz et al., 2018, IJF; Marcjasz et al., 2018, WP)



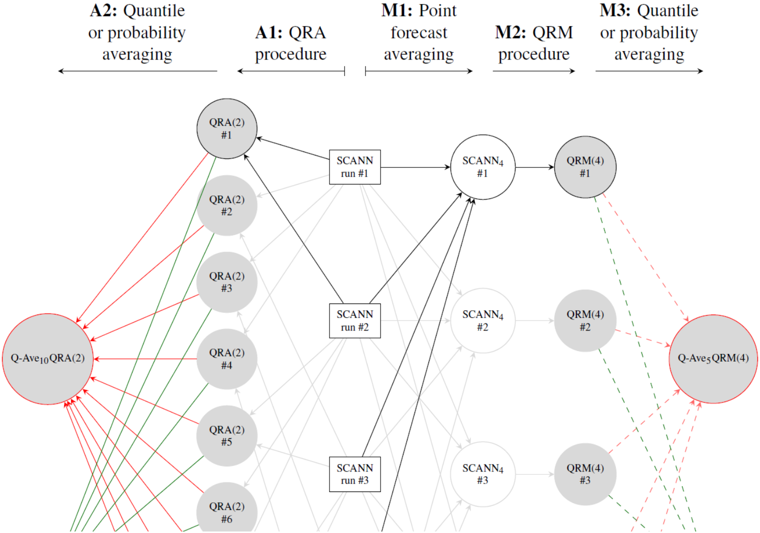
- One hidden layer with 5 neurons and sigmoid activation functions
- NARX inputs identical as in the **ARX** model
- Trained using Matlab's `trainlm` function, utilizing the Levenberg-Marquardt algorithm for supervised learning

Sample gains from using committee machines



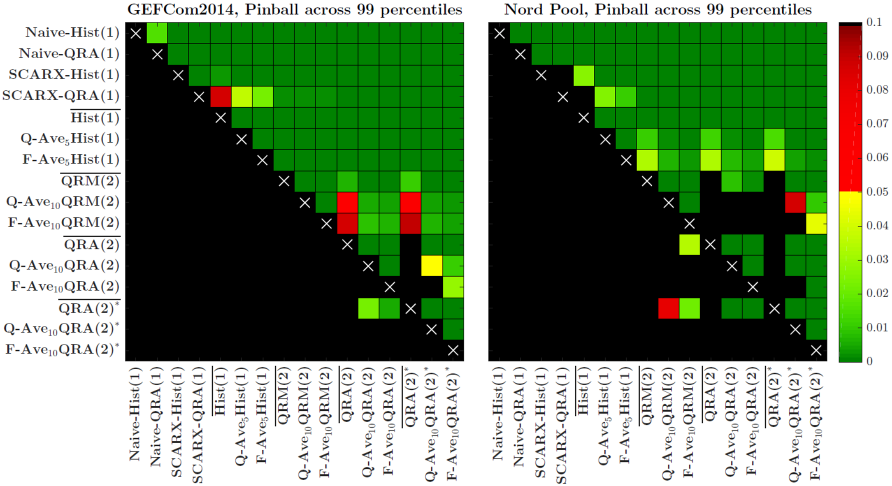
- Forecast errors roughly scale as a power-law function of the number of networks (runs) in a committee machine
- We should use as large committee machines as we can ...
- ... or use a more efficient training algorithm → FANN library, see Marcjasz, Uniejewski & Weron (2018, WP)

Combine point or probabilistic forecasts?



p-values of the DM test across 99 percentiles

(Marcjasz, Uniejewski & Weron, 2018, WP)



Main findings

- Point forecasts
  - **(SC)ANN** models outperform **(SC)ARX** for every LTSC

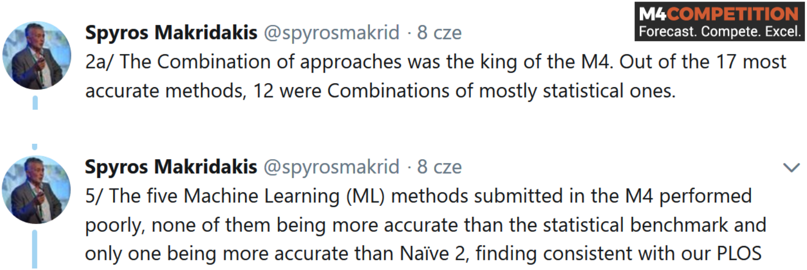
- Probabilistic forecasts
  - The **QRA** and **QRM** models are much better than **Hist**
  - **QRA(2)** is the best performer, but much slower than **QRM(N)**
  - It is always beneficial to average forecasts (vertical is better)

- Times needed to produce one week of hourly forecasts:

QRM(N)	**QRA(2)**	**QRA(3)**	**QRA(4)**	**QRA(5)**
25 s	68 s	111.5 s	171.5 s	248 s

Wrap up

- Point forecasts
  - **(SC)ANN** models outperform **(SC)ARX** ... if properly designed
- Probabilistic forecasts
  - **(SC)ANN** models offer opportunities that **(SC)ARX** do not
- However:



Spyros Makridakis @spyrosmakrid · 8 cze

2a/ The Combination of approaches was the king of the M4. Out of the 17 most accurate methods, 12 were Combinations of mostly statistical ones.

Spyros Makridakis @spyrosmakrid · 8 cze

5/ The five Machine Learning (ML) methods submitted in the M4 performed poorly, none of them being more accurate than the statistical benchmark and only one being more accurate than Naïve 2, finding consistent with our PLOS