DRZEWA REGRESYJNE I LASY LOSOWE JAKO NARZĘDZIA PREDYKCJI SZEREGÓW CZASOWYCH Z WAHANIAMI SEZONOWYMI

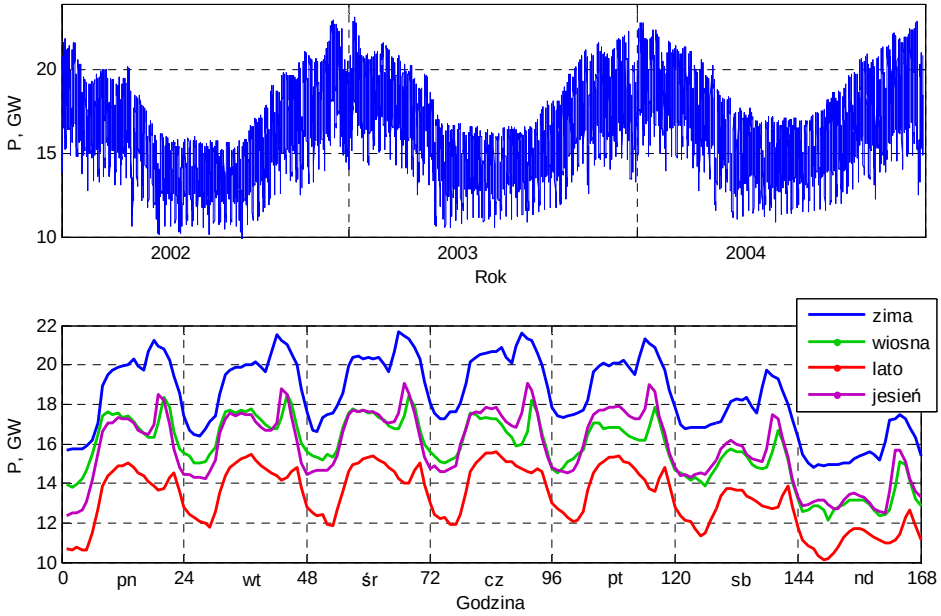
Grzegorz Dudek

Instytut Informatyki
Wydział Elektryczny
Politechnika Częstochowska
www.gdudek.el.pcz.pl

VI spotkanie Polskiej Grupy Badawczej Systemów Uczących się
Częstochowa, 14.04.2016 r.

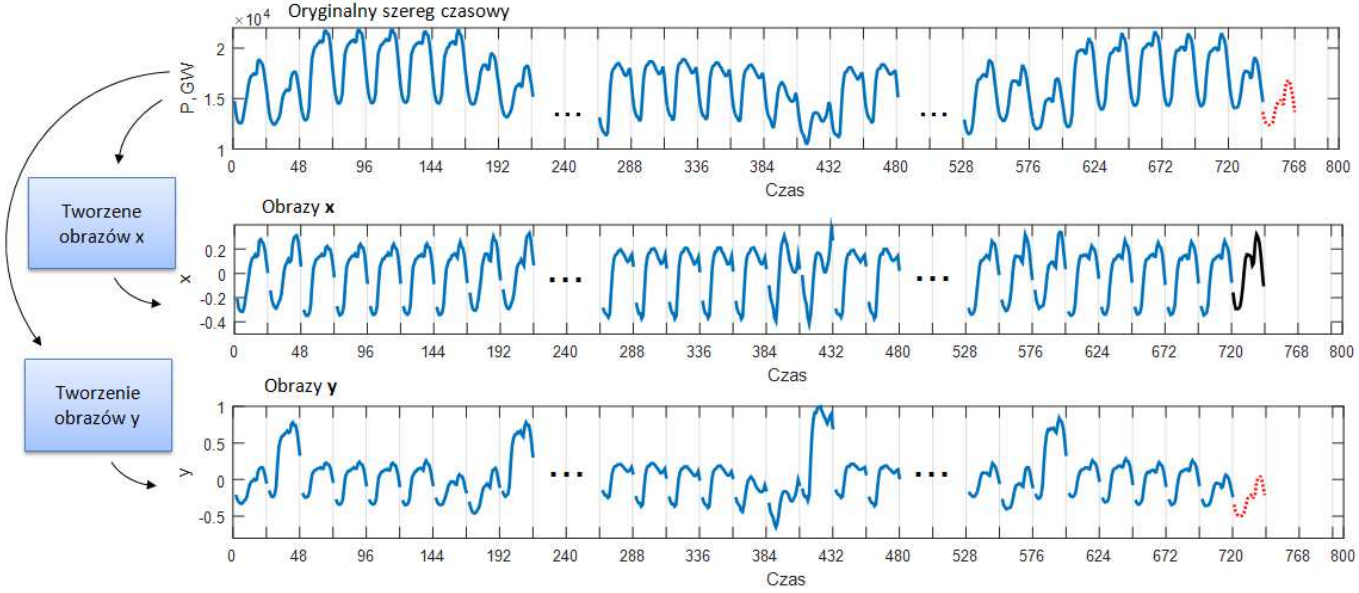
PROBLEM PREDYKCJI SZEREGU CZASOWEGO

Predykcja szeregu czasowego z wieloma cyklami wahań sezonowych w horyzoncie τ na podstawie przebiegu historycznego.

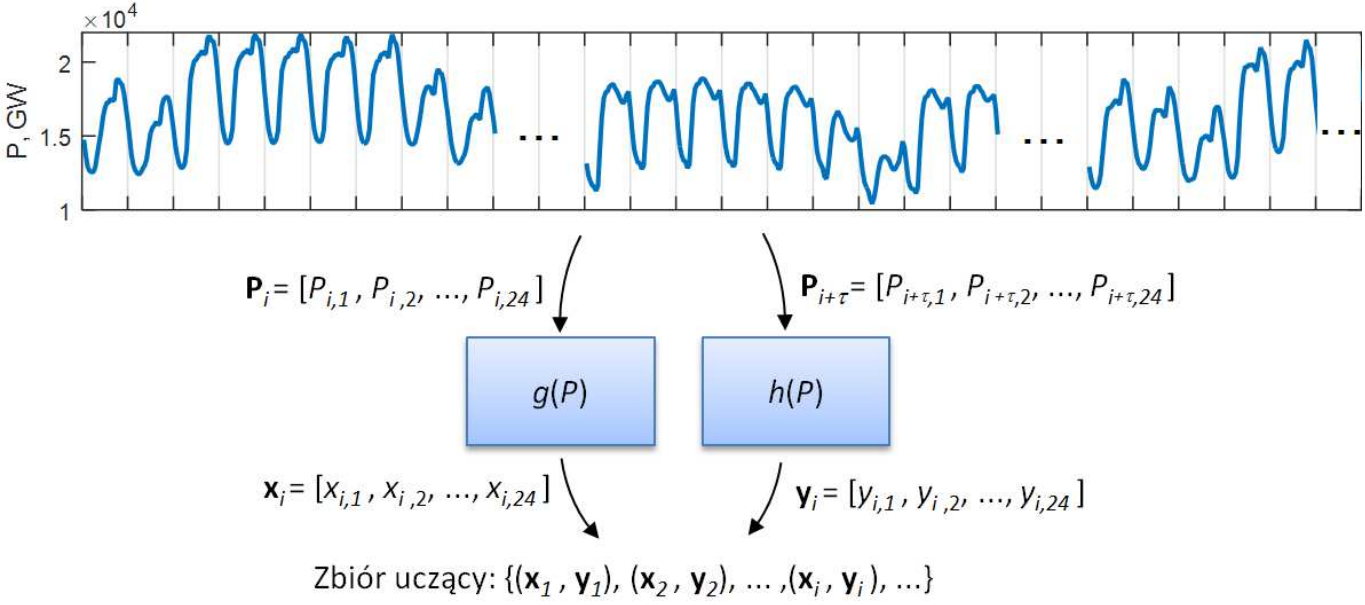


Obciążenie system elektroenergetycznego z cyklami rocznymi, tygodniowymi i dobowymi

IDEA


Oryginalny szereg czasowy
×10^4
P, GW
Czas
Tworzene obrazów x
Obrazy x
x
Czas
Tworzenie obrazów y
Obrazy y
y
Czas

IDEA


×10^4
2
1.5
1
P, GW
...
...
...
P_i = [P_i,1 , P_i ,2, ..., P_i,24]
P_i+τ = [P_i+τ,1 , P_i +τ,2, ..., P_i+τ,24]
g(P)
h(P)
x_i = [x_i,1 , x_i ,2, ..., x_i,24]
y_i = [y_i,1 , y_i ,2, ..., y_i,24]
Zbiór uczący: {(x_1 , y_1), (x_2 , y_2), ... ,(x_i , y_i), ...}

Definicja obrazów cykli dobowych

Obrazy wejściowe $\mathbf{x}_i = [x_{i,1}\ x_{i,2}\ \ldots\ x_{i,n}]$ odwzorowują wyrazy poprzedzające moment prognozy – obciążenia doby i: $\mathbf{P}_i = [P_{i,1}\ P_{i,2}\ \ldots\ P_{i,n}]$

$$x_{i,t} = g(P_{i,t}) = \frac{P_{i,t} - \overline{P}_i}{\sqrt{\sum_{j=1}^{n} (P_{i,j} - \overline{P}_i)^2}}$$

Obrazy $\mathbf{x}_i$ są unormowanymi wersjami wektorów $\mathbf{P}_i$
Ich długość jest jednostkowa, średnia zerowa, a wariancja jednakowa

Obrazy wyjściowe $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \ldots\ y_{i,n}]$ odwzorowują wyrazy w okresie prognozowanym – w kolejnych chwilach doby prognozy $i+\tau$: $\mathbf{P}_{i+\tau} = [P_{i+\tau,1}\ P_{i+\tau,2}\ \ldots\ P_{i+\tau,n}]$

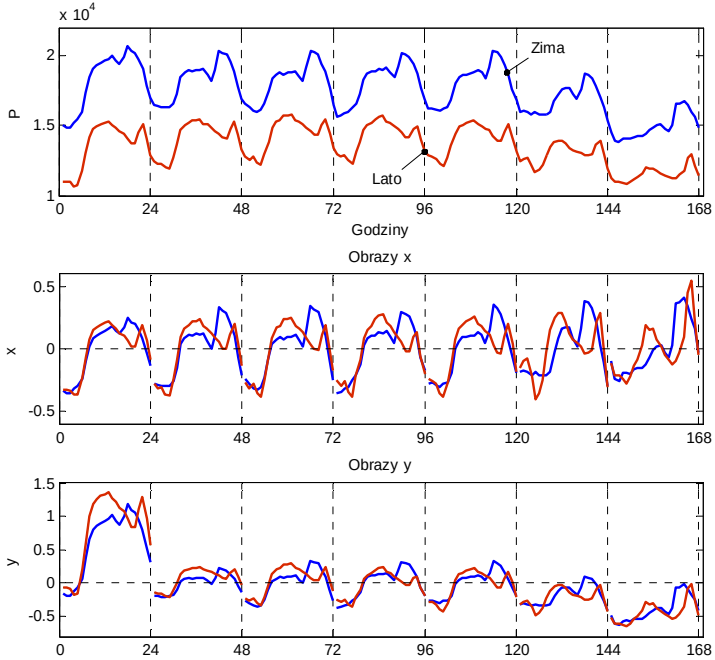
$$y_{i,t} = h(P_{i+\tau,t}) = \frac{P_{i+\tau,t} - \overline{P}_i}{\sqrt{\sum_{j=1}^{n}(P_{i,j} - \overline{P}_i)^2}}$$

Inne definicje obrazów:

- ⇒ Dudek G.: *Systemy uczące się oparte na podobieństwie obrazów do prognozowania szeregów czasowych obciążeń elektroeneregtycznych*. EXIT, Warszawa 2012
- ⇒ Dudek G.: *Pattern Similarity-based Methods for Short-term Load Forecasting – Part 1: Principles*. Applied Soft Computing, vol. 37, pp. 277-287, 2015

Cel

Odfiltrowanie trendu i cykli dłuższych niż podstawowy (dobowy), sprowadzenie szeregu do stacjonarności



Model prognostyczny

$$f : X \rightarrow Y$$

Wyjściem modelu jest prognoza obrazu **y** (lub jego składowej)

Prognoza wyrazów szeregu czasowego

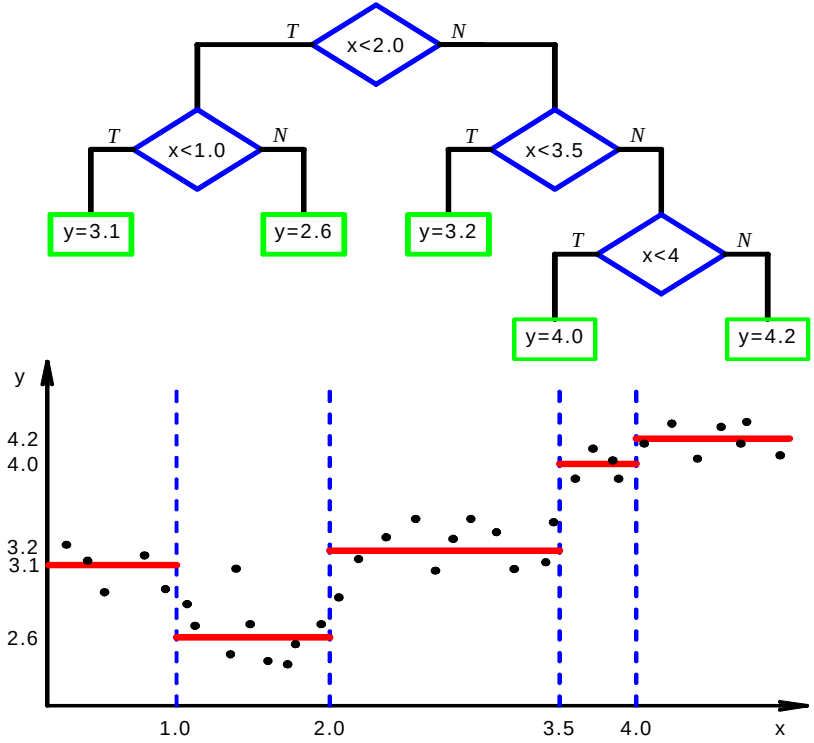
$$\widehat{P}_{i+\tau,t} = h^{-1}(P_{i+\tau,t}) = \widehat{y}_{i,t}\sqrt{\sum_{j=1}^{n}(P_{i,j} - \overline{P}_i)^2} + \overline{P}_i$$

DRZEWO REGRESYJNE (CART)

Cechy

- Reprezentacja drzewiasta lub zbiór reguł decyzyjnych „jeśli – to”
- Działanie na zmiennych ilościowych i jakościowych
- Podział przestrzeni cech na hiper-prostopadłościany
- Lokalna aproksymacja funkcji stałą wewnątrz hiperprostopadłościanu (aproksymacja dyskretna)
- Zależnie od funkcji docelowej drzewo decyzyjne może pełnić rolę klasyfikatora lub modelu regresyjnego

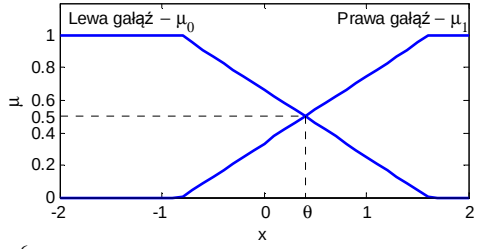
DRZEWO REGRESYJNE (CART)

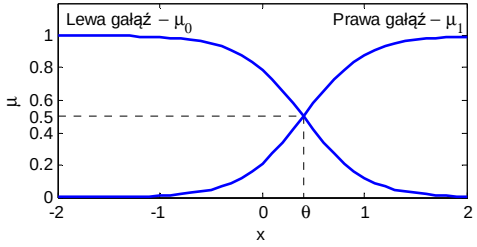


DRZEWO REGRESYJNE Z ROZMYTYMI WĘZŁAMI

- Sposób konstrukcji drzewa regresyjnego z rozmytymi węzłami jest taki sam jak drzewa w wariancie podstawowym
- Testy przeprowadzanie w węzłach pośrednich zmieniają postać:

$$T(\mathbf{x}) = \begin{cases} 1, & \text{jeśli } x_i > \theta_i \\ 0, & \text{jeśli } x_i \le \theta_i \end{cases} \quad \rightarrow \quad T(\mathbf{x}) = (\mu_1(\mathbf{x}), \mu_0(\mathbf{x}))$$





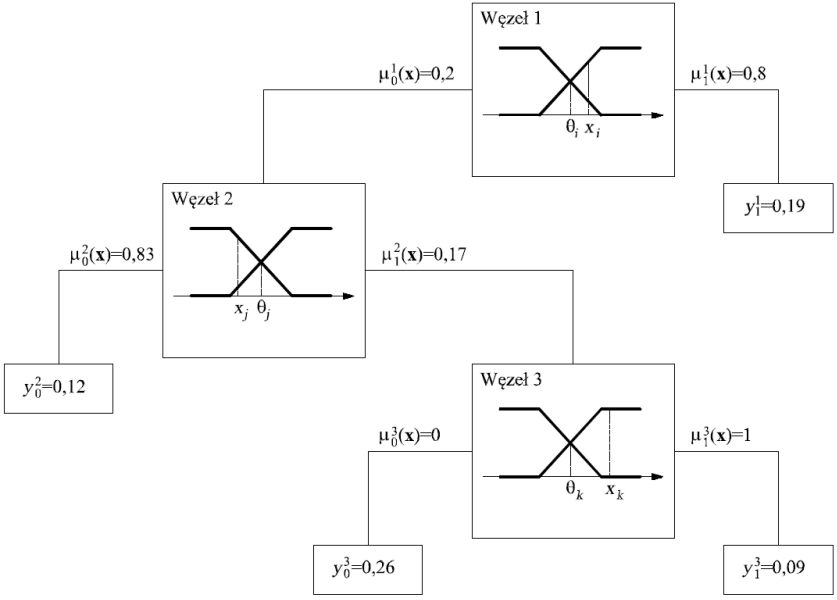
$$\mu_1(\mathbf{x}) = \begin{cases} 0, & \text{jeśli } x_i \le \theta_i - \dfrac{0{,}5}{a} \\ 1, & \text{jeśli } x_i \ge \theta_i + \dfrac{0{,}5}{a} \\ a(x_i - \theta) + 0{,}5, & \text{jeśli } \theta_i - \dfrac{0{,}5}{a} > x_i > \theta_i + \dfrac{0{,}5}{a} \end{cases}$$

$$\mu_1(\mathbf{x}) = \frac{1}{1 + \exp(-a(x_i - \theta_i))}$$

$$\mu_0(\mathbf{x}) = 1 - \mu_1(\mathbf{x})$$

⇒ Dudek G.: **Prognozowanie krótkoterminowe obciążeń systemów elektroenergetycznych z wykorzystaniem rozmytych drzew regresyjnych.** Przegląd Elektrotechniczny, r. 90, nr 4, s. 108-111, 2014.

DRZEWO REGRESYJNE Z ROZMYTYMI WĘZŁAMI



$$y(\mathbf{x}) = y_1^1\mu_1^1(\mathbf{x}) + y_0^2\mu_0^2(\mathbf{x})\mu_0^1(\mathbf{x}) + y_0^3\mu_0^3(\mathbf{x})\mu_1^2(\mathbf{x})\mu_0^1(\mathbf{x})$$
$$+ y_1^3\mu_1^3(\mathbf{x})\mu_1^2(\mathbf{x})\mu_0^1(\mathbf{x}) = 0,19\cdot 0,8 + 0,12\cdot 0,83\cdot 0,2$$
$$+ 0,26\cdot 0\cdot 0,17\cdot 0,2 + 0,09\cdot 1\cdot 0,17\cdot 0,2 = 0,175$$

LAS LOSOWY

1. Powtarzaj dla każdego drzewa (dla k = 1 do K)

    1.1. Wylosuj ze zbioru uczącego próbę bootstrapową o rozmiarze N

    1.2. Zbuduj drzewo T_k na próbie bootstrapowej, powtarzając dla każdego węzła, jeśli jego rozmiar jest większy od m

        1.2.1. Wylosuj $F \leq n$ składowych obrazu **x**

        1.2.2. Znajdź składową x_i i wartość progową θ_i (przegląd zupełny)

        1.2.3. Rozdziel węzeł na dwa węzły potomne

2. Zwróć drzewa $\{T_k\}_{k= 1, 2, \ldots, K}$

Wyznaczenie prognozy dla obrazu **x**:

$$f(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} T_k(\mathbf{x})$$

⇒ Hastie T., Tibshirani R., Friedman J.: **The Elements of Statistical Learning. Data Mining, Inference, and Prediction**. Springer 2009

Dane

Szereg czasowy obciążeń krajowego systemu elektroenergetycznego w okresie 2002-2004

Problem prognostyczny

Prognoza obciążeń godzinowych w kolejnych dniach stycznia i lipca 2004, $\tau = 1$

Zbiór uczący

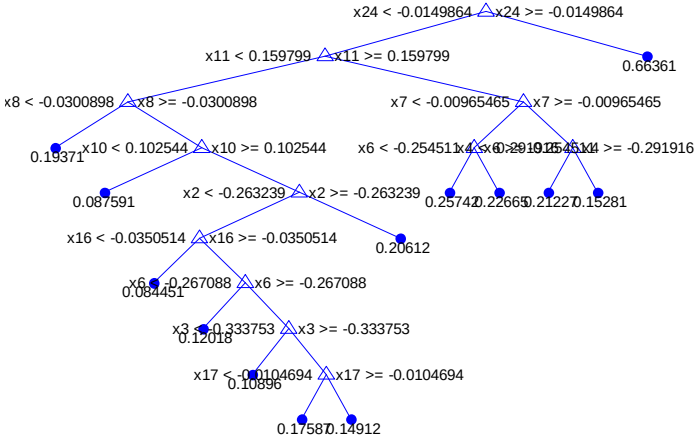
Zbiór uczący zawierał przykłady reprezentujące te same typy dni tygodnia, co przykład testowy

Błąd prognozy

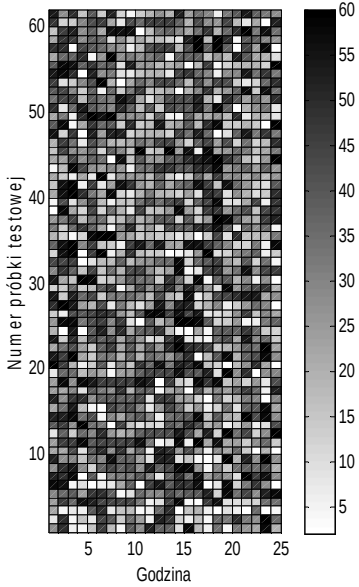
$$MAPE = \frac{100}{M} \sum_{j=1}^{M} \left| \frac{P_j - \hat{P}_j}{P_j} \right|$$

Drzewo regresyjne (CART)

Parametr - *m* (przegląd zupełny, local leave-one-out)



Drzewo regresyjne utworzone w zadaniu prognozy obciążenia dn. 01.07.2004 r. o godz. 12, (*m* = 18)

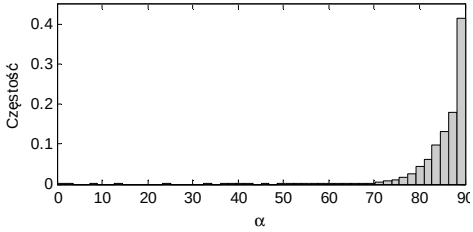


Optymalne wartości *m*

Drzewo regresyjne z rozmytymi węzłami (Fuzzy CART)

Parametr - kąt nachylenia funkcji przynależności α (przegląd zupełny, local leave-one-out)

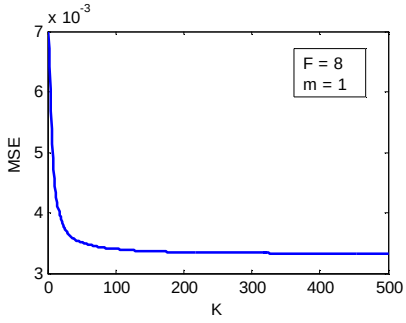
Wariant drzewa	Parametry	$MAPE_{wal}$	$MAPE_{tst}$
CART	m = var	1,27	1,42
Fuzzy CART	$m = m_{CART}$, α = var	1,12	1,33
Fuzzy CART	m = 30, α = var	1,22	1,36
Fuzzy CART	m = 1, α = var	1,22	1,33
Fuzzy CART	$m = m_{CART}$, α = var	1,13	1,31
Fuzzy CART	m = 30, α = var	1,23	1,42
Fuzzy CART	m = 1, α = var	1,23	1,35
Fuzzy CART	$m = m_{CART}$, α_1, …, α_m = var	0,74	1,37

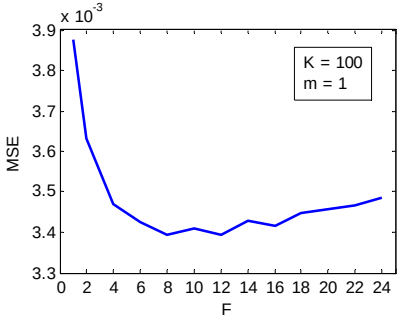


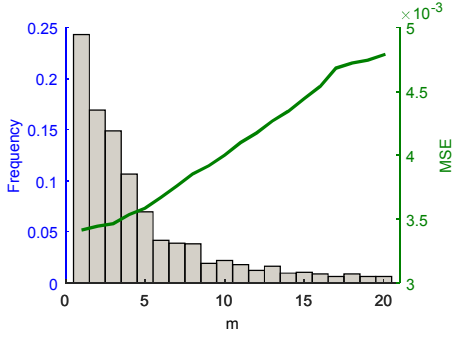
Histogram optymalnych kątów nachylenia funkcji przynależności

Las losowy

Parametry - liczba drzew *K*, liczba składowych *F*, *m* (przegląd zupełny, out-of-bag)



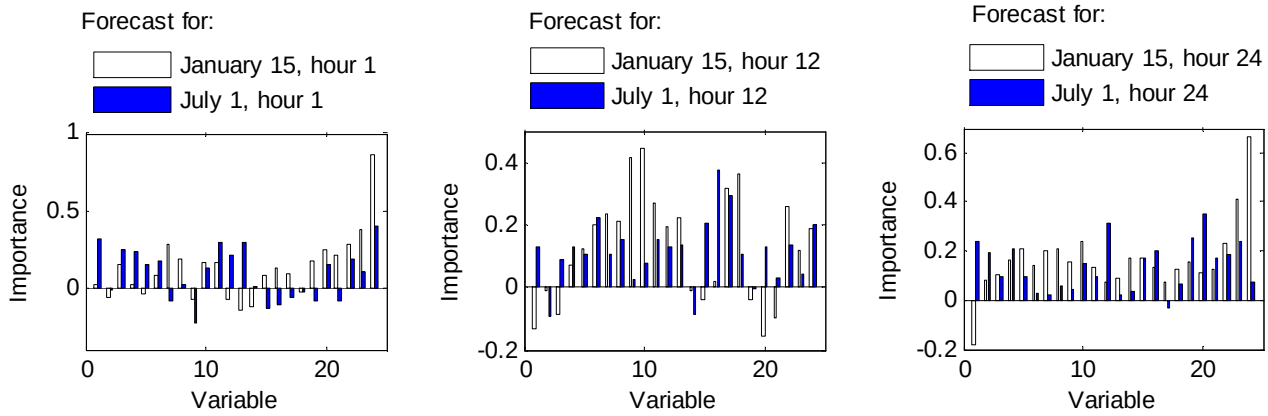




⇒ Dudek G.: **Short-Term Load Forecasting using Random Forests.** In: Filev D. et al. (eds.): Intelligent Systems'2014, Advances in Intelligent Systems and Computing 323, pp. 821-828, 2015.

Las losowy

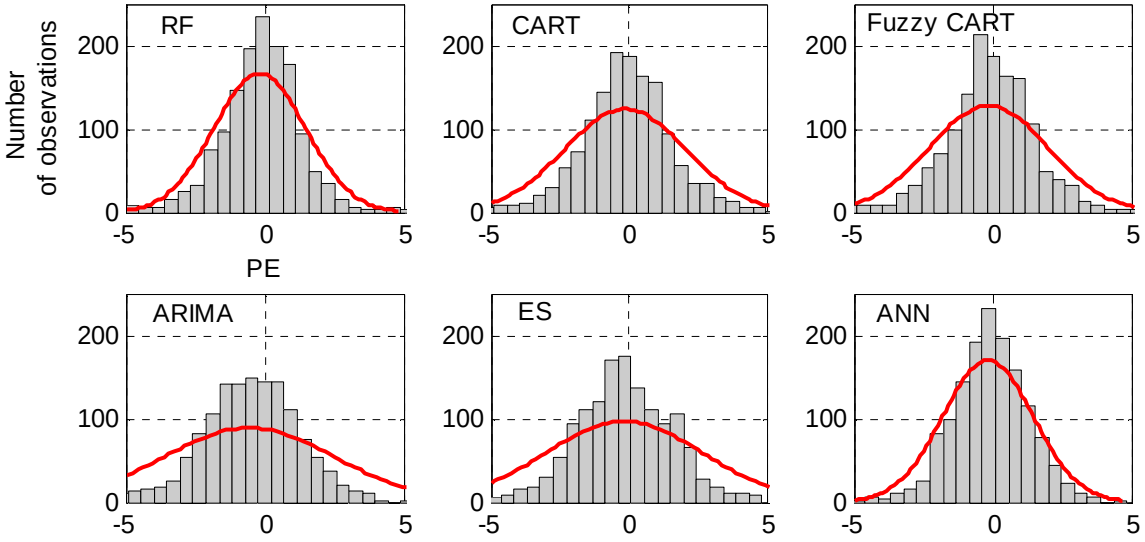
Ważność składowych



Wyniki

Model	Styczeń		Lipiec		Średni	
	$MAPE_{tst}$	*IQR*	$MAPE_{tst}$	*IQR*	$MAPE_{tst}$	*IQR*
Las losowy	1.42	1.39	0.92	0.98	1.16	1.17
CART	1.70	1.58	1.16	1.17	1.42	1.39
Fuzzy CART	1.62	1.47	1.13	1.12	1.37	1.35
ARIMA	2.64	2.34	1.21	1.24	1.91	1.67
Wygładzanie wykładnicze	2.35	1.88	1.19	1.30	1.76	1.56
Sieć neuronowa	1.32	1.30	0.97	1.01	1.14	1.15
Prognoza naiwna	6.37	5.36	1.29	1.20	3.78	3.82

Rozkład błędów



WNIOSKI

- Reprezentacja szeregów czasowych za pomocą obrazów cykli sezonowych ułatwia prognozowanie szeregów niestacjonarnych z trendem i wieloma cyklami wahań sezonowych

- Model prognostyczne oparte na drzewach regresyjnych wyróżnia prosta i zrozumiała budowa oraz niewielka liczba parametrów

- Rozmyta wersja drzew regresyjnych pozwala sterować równowagą między obciążeniem i wariancją modelu

- Lasy losowe jako komitet słabych uczniów pozwalają zredukować błąd prognozy i uzyskać stabilniejsze rezultaty

Dziękuję za uwagę