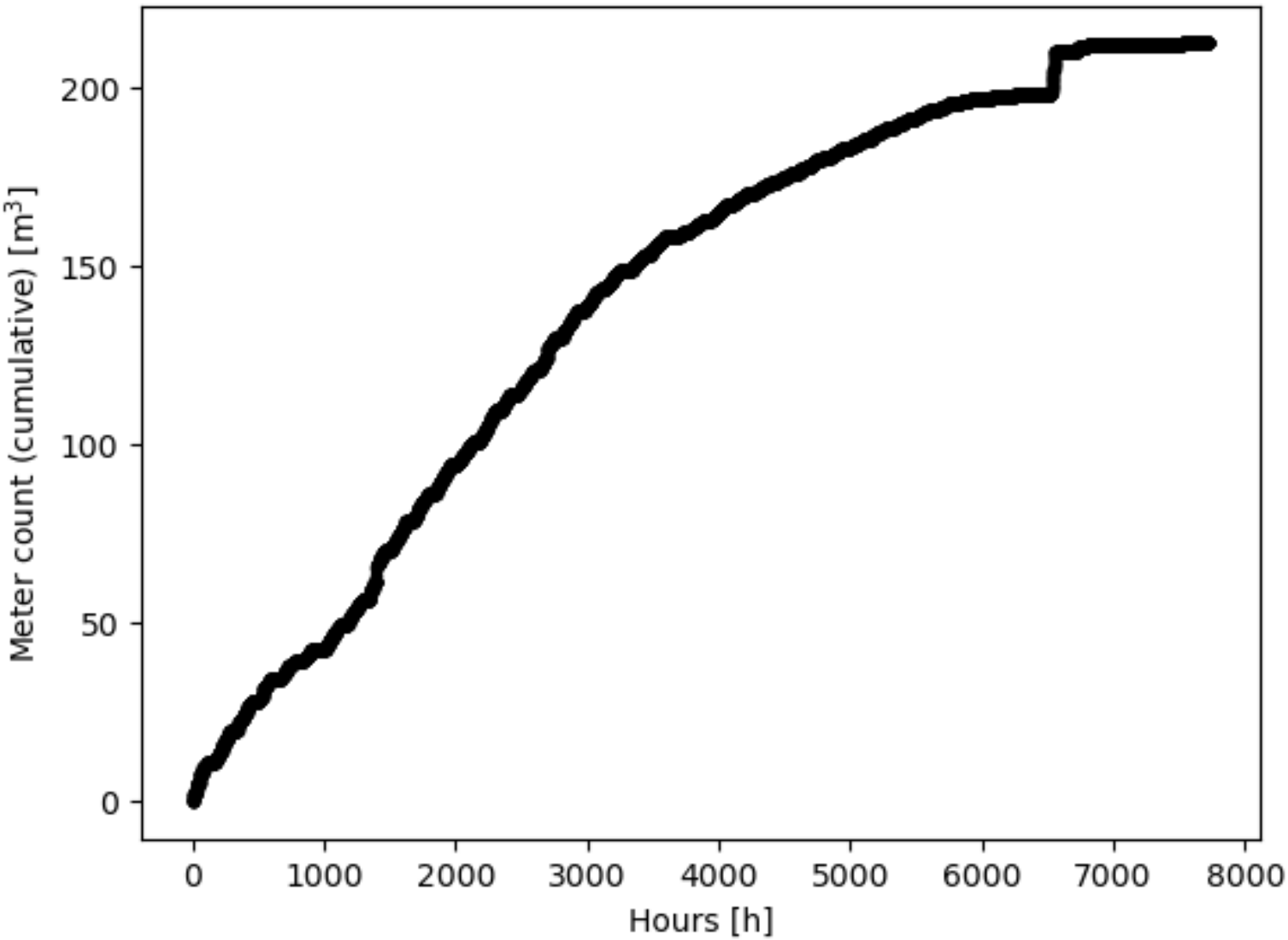
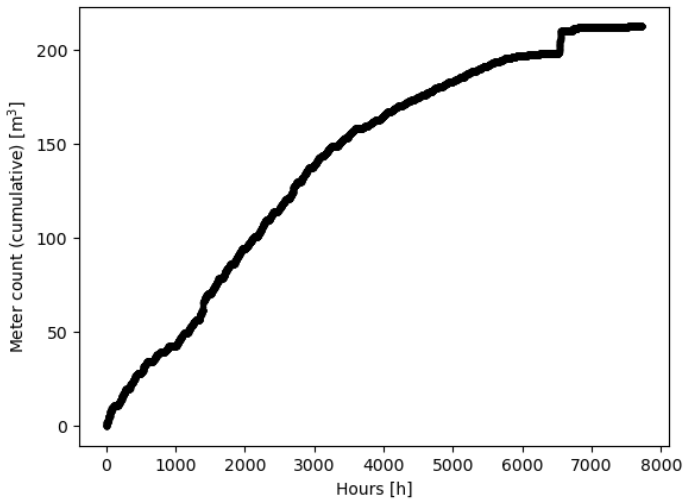
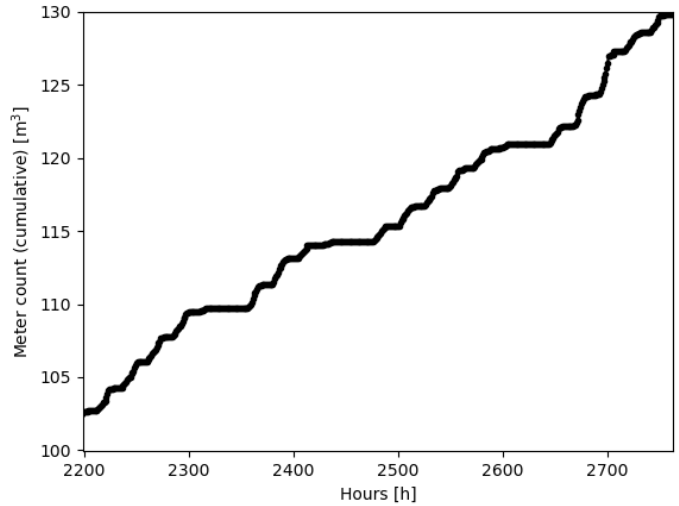
Plan wykładu

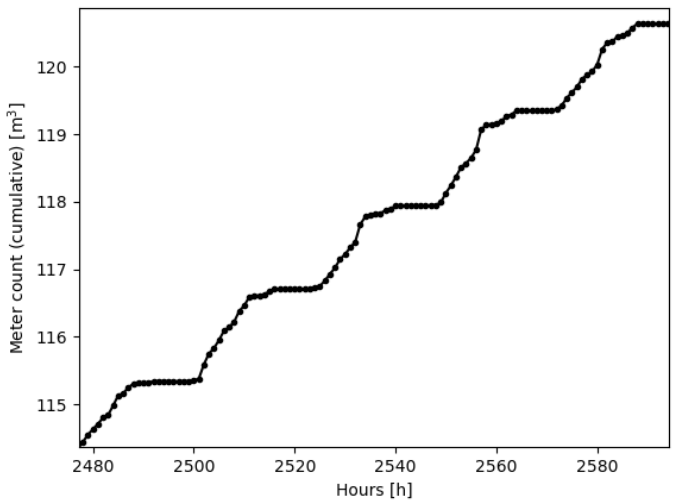
Dane – przykładowy sygnał

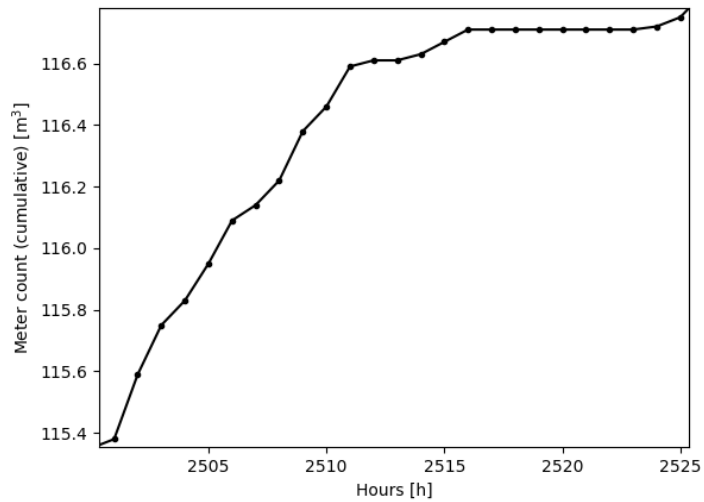


Dane – przykładowy sygnał (zoom)









Dane – przykładowy sygnał – obserwacje/cechy

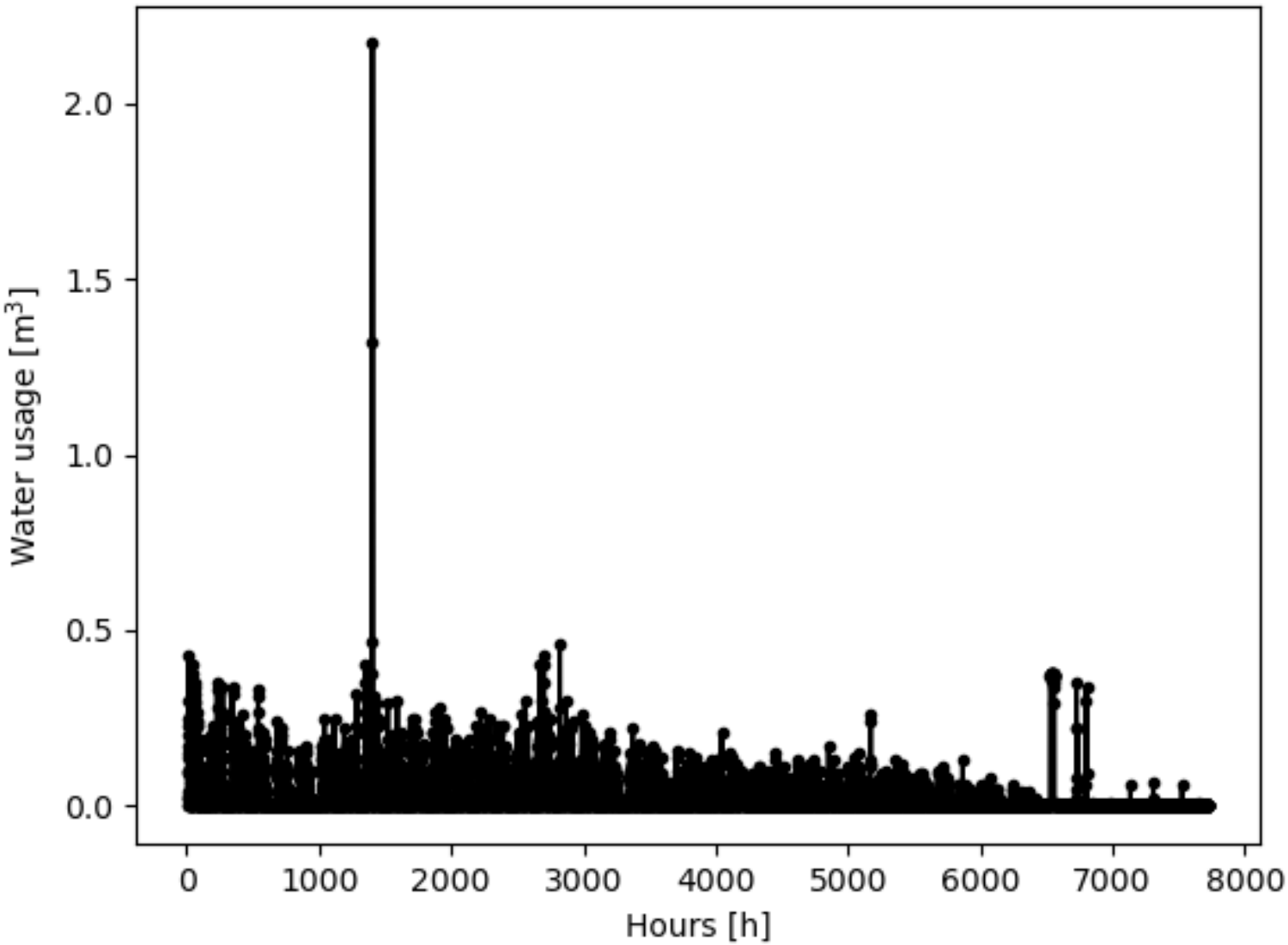
1. Szereg czasowy

Dane – przykładowy sygnał – obserwacje/cechy

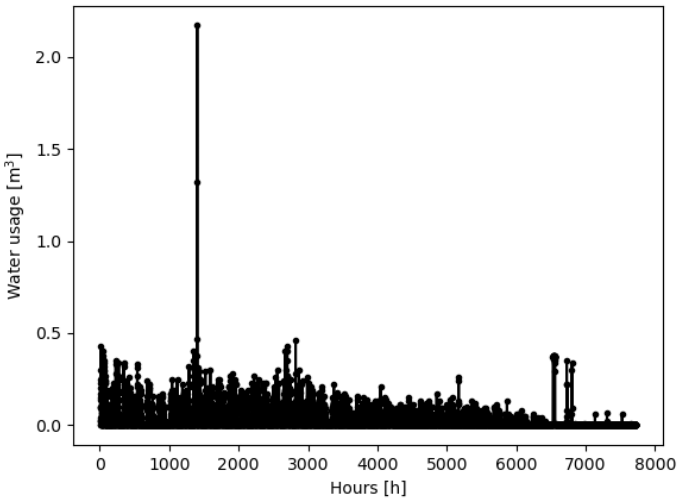
1. Szereg czasowy
2. Kumulatywne zliczanie impulsów – sygnał kumulatywny (ang. integrative) – różnicowanie (ang. differentiation)

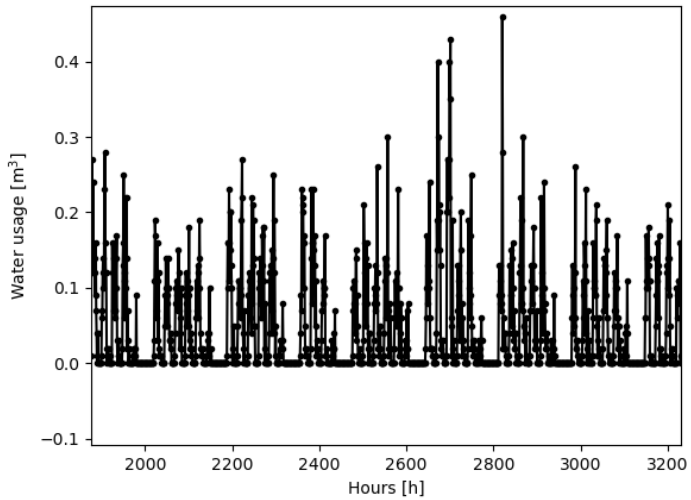
   $\longrightarrow$ Sygnał, sygnał po różnicowaniu, sygnał po scałkowaniu

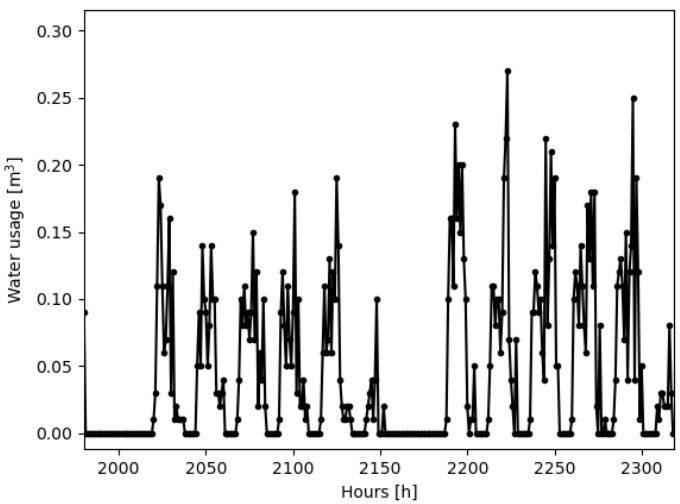
Dane – przykładowy sygnał – różnicowanie

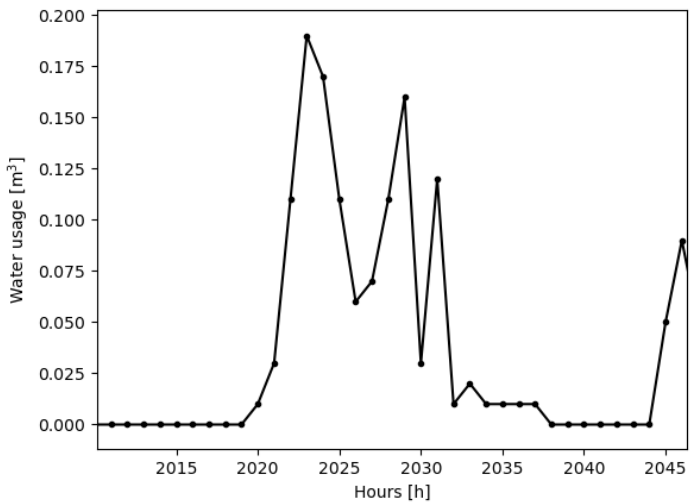


Dane – przykładowy sygnał – różnicowanie



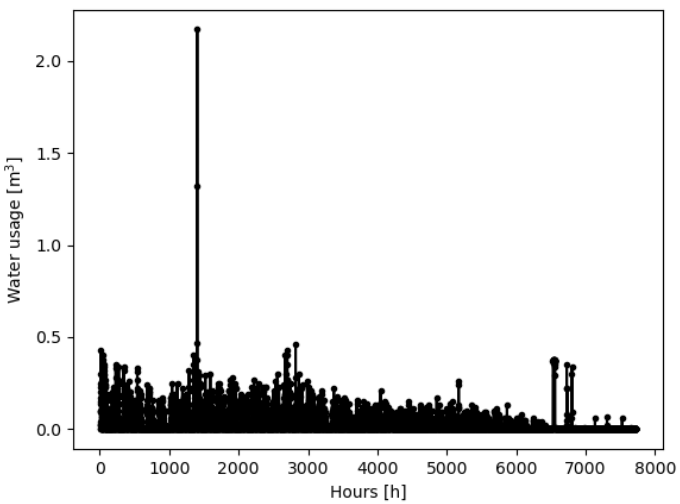


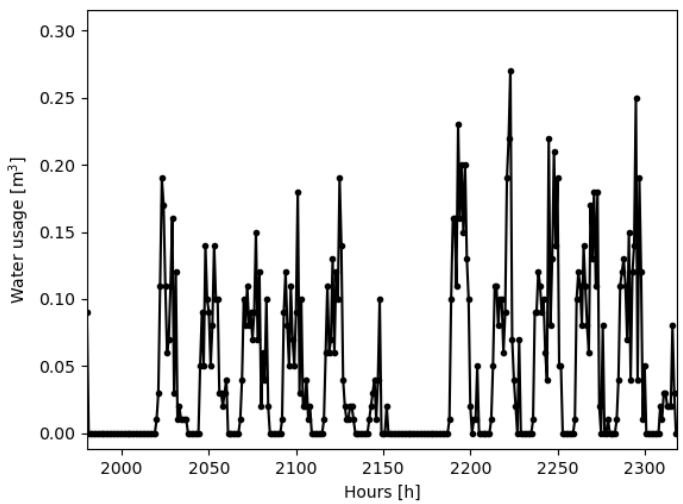




Dane – przykładowy sygnał – obserwacje/cechy

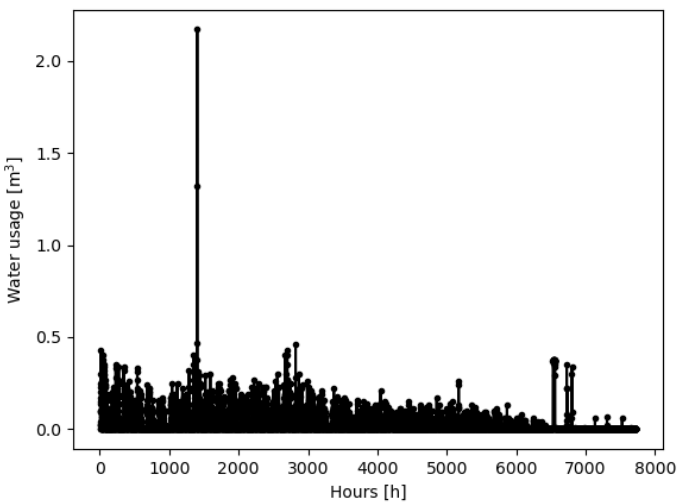
1. Szereg czasowy
2. Kumulatywne zliczanie impulsów – sygnał kumulatywny (ang. integrative) – różnicowanie (ang. differentiation)
   $\longrightarrow$ Sygnał, sygnał po różnicowaniu, sygnał po scałkowaniu
3. Cykl – dobowy, tygodniowy
   $\longrightarrow$ Cykle, sezonowość

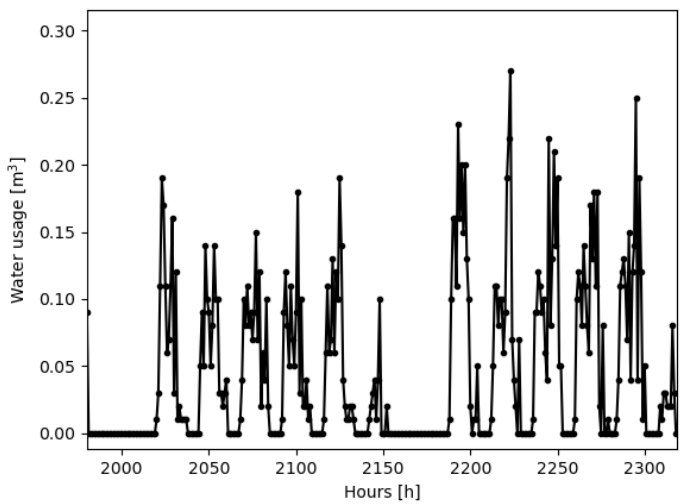




Dane – przykładowy sygnał – obserwacje/cechy

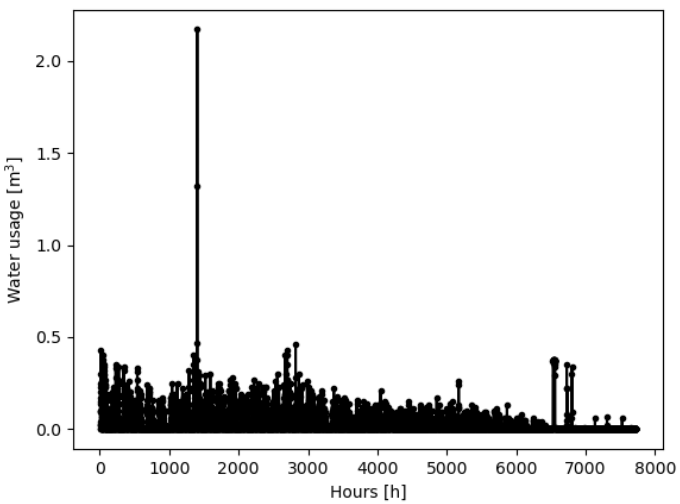
1. Szereg czasowy
2. Kumulatywne zliczanie impulsów – sygnał kumulatywny (ang. integrative) – różnicowanie (ang. differentiation)
   $\longrightarrow$ Sygnał, sygnał po różnicowaniu, sygnał po scałkowaniu
3. Cykl – dobowy, tygodniowy
   $\longrightarrow$ Cykle, sezonowość
4. Trend – zmienny

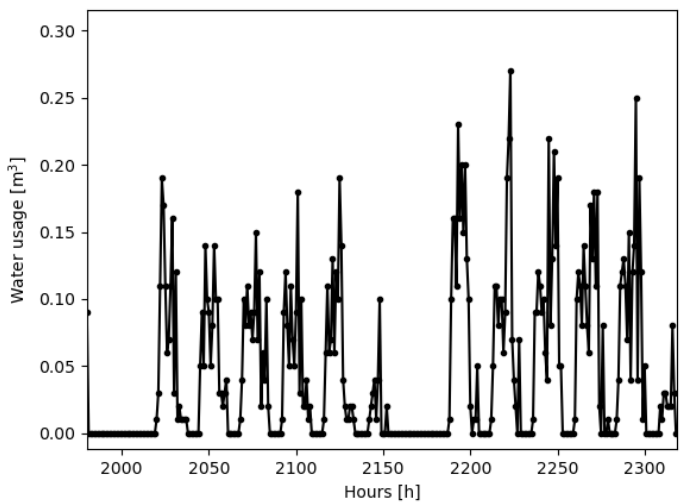




Dane – przykładowy sygnał – obserwacje/cechy

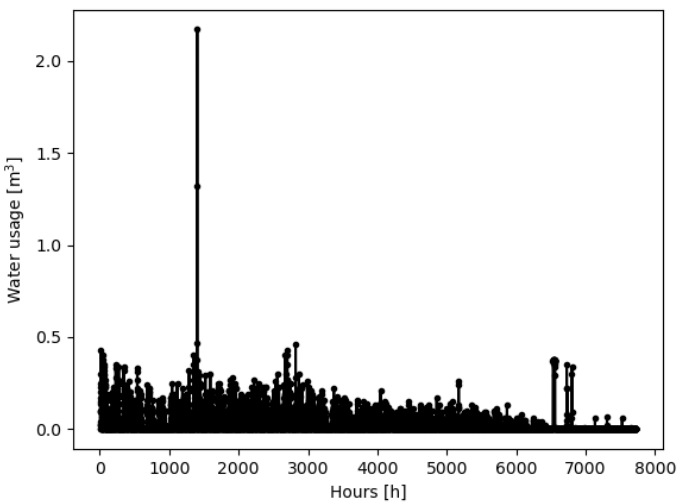
1. Szereg czasowy
2. Kumulatywne zliczanie impulsów – sygnał kumulatywny (ang. integrative) – różnicowanie (ang. differentiation)
   $\longrightarrow$ Sygnał, sygnał po różnicowaniu, sygnał po scałkowaniu
3. Cykl – dobowy, tygodniowy
   $\longrightarrow$ Cykle, sezonowość
4. Trend – zmienny
5. Pojedyncze zdarzenia – anomalie

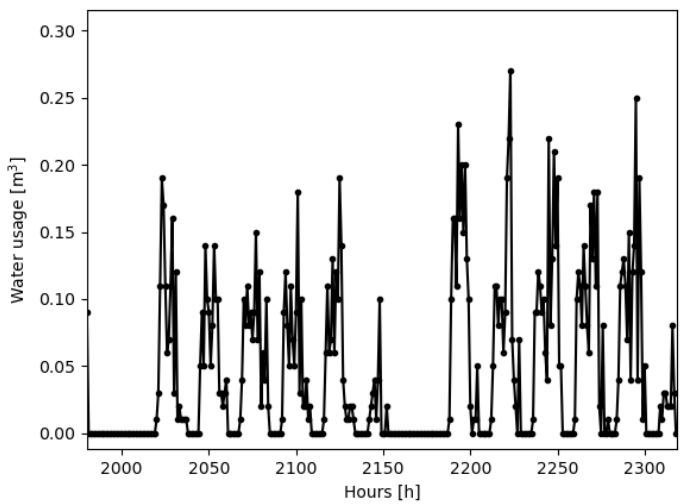




Dane – przykładowy sygnał – obserwacje/cechy

1. Szereg czasowy
2. Kumulatywne zliczanie impulsów – sygnał kumulatywny (ang. integrative) – różnicowanie (ang. differentiation)
   $\longrightarrow$ Sygnał, sygnał po różnicowaniu, sygnał po scałkowaniu
3. Cykl – dobowy, tygodniowy
   $\longrightarrow$ Cykle, sezonowość
4. Trend – zmienny
5. Pojedyncze zdarzenia – anomalie
6. Wartości liczbowe (min/max, typowe, rozkład wartości)





Bazy danych, weryfikacja danych

1. Dane, informacja, wiedza

Bazy danych, weryfikacja danych

1. Dane, informacja, wiedza
2. Reprezentacja danych – sposób zapisu w pamięci komputera

Bazy danych, weryfikacja danych

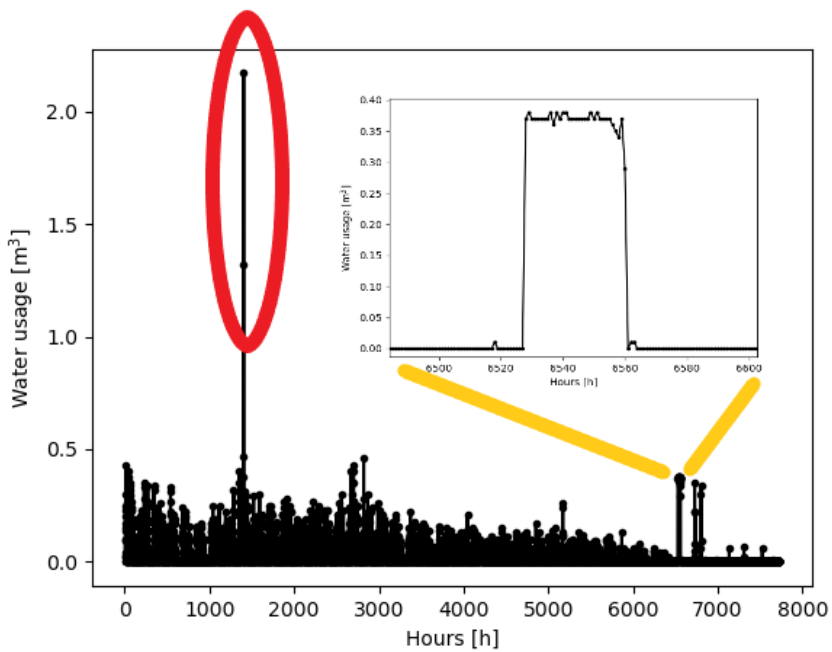
1. Dane, informacja, wiedza
2. Reprezentacja danych – sposób zapisu w pamięci komputera
3. Baza danych – zorganizowany zbiór danych

Bazy danych, weryfikacja danych

1. Dane, informacja, wiedza
2. Reprezentacja danych – sposób zapisu w pamięci komputera
3. Baza danych – zorganizowany zbiór danych
4. Weryfikacja
   - 4.1 Poprawność i spójność
   - 4.2 Weryfikacja merytoryczna – „sens" danych, znaczenie, kontekst, ...
     - 4.2.1 Anomalie – odstępstwa od „normy"
     - 4.2.2 Wystąpienie znanych problemów – detekcja/klasyfikacja

Bazy danych, weryfikacja danych

1. Dane, informacja, wiedza
2. Reprezentacja danych – sposób zapisu w pamięci komputera
3. Baza danych – zorganizowany zbiór danych
4. Weryfikacja
   - 4.1 Poprawność i spójność
   - 4.2 Weryfikacja merytoryczna – „sens" danych, znaczenie, kontekst, ...
     - 4.2.1 Anomalie – odstępstwa od „normy"
     - 4.2.2 Wystąpienie znanych problemów – detekcja/klasyfikacja



Plan wykładu

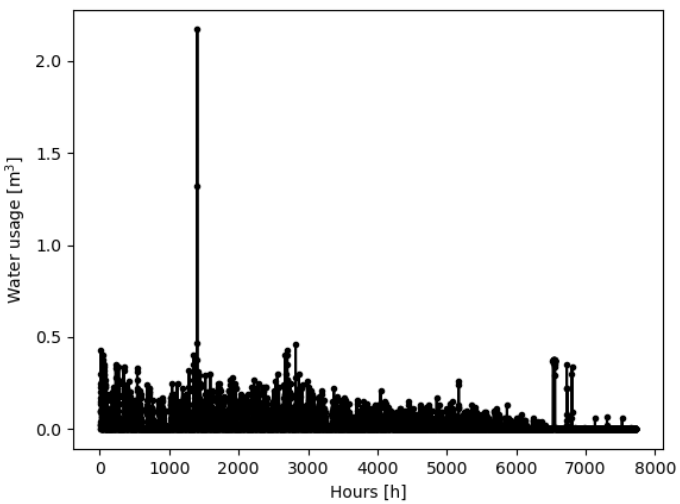
Analiza przykładowego sygnału

Weryfikacja statystyczna – MAD

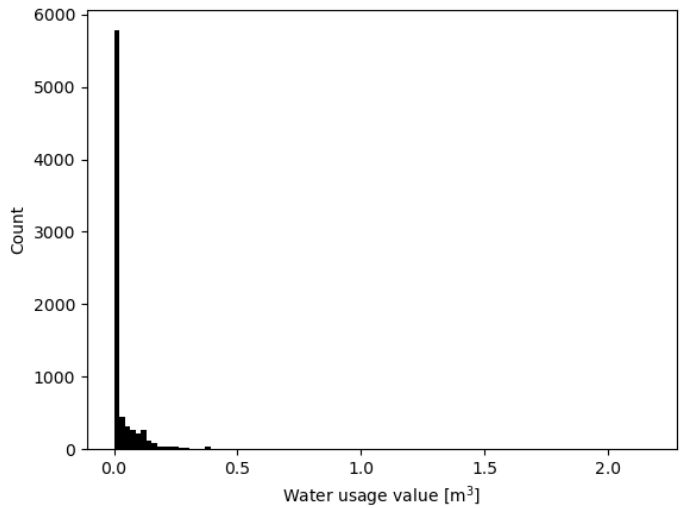
Weryfikacja statystyczna – RX

Statystyki opisowe – opis sygnału

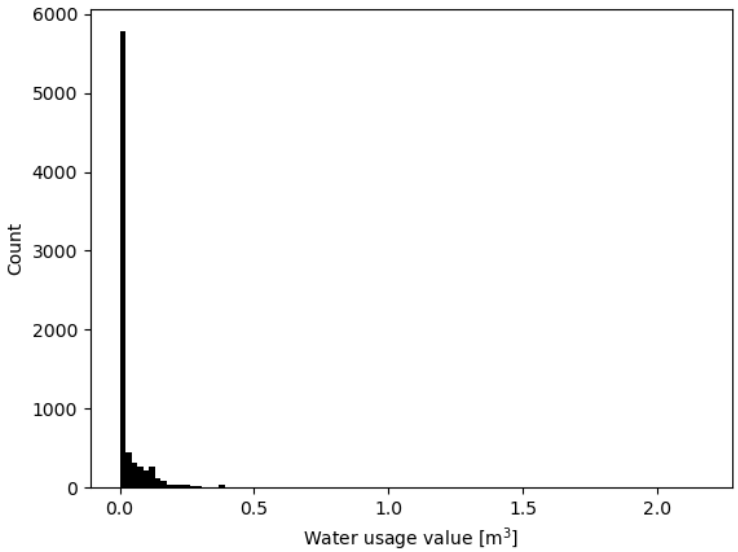
1. ~~Szereg czasowy~~ $\longrightarrow$ histogram

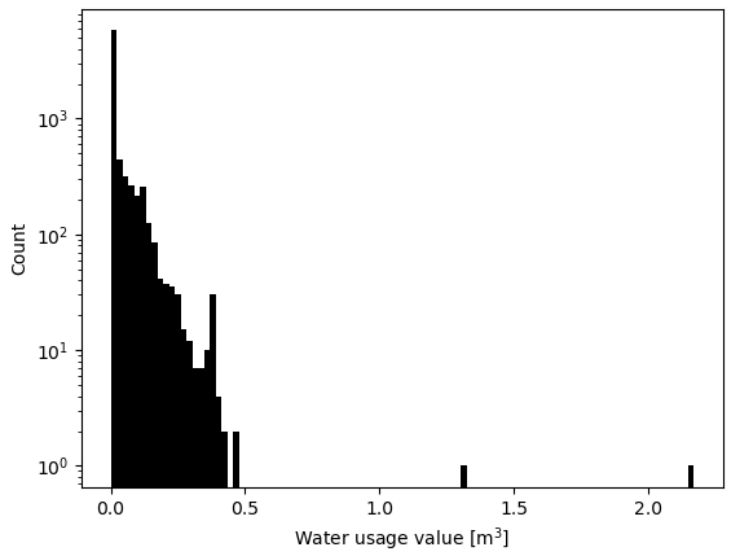


$\longrightarrow$



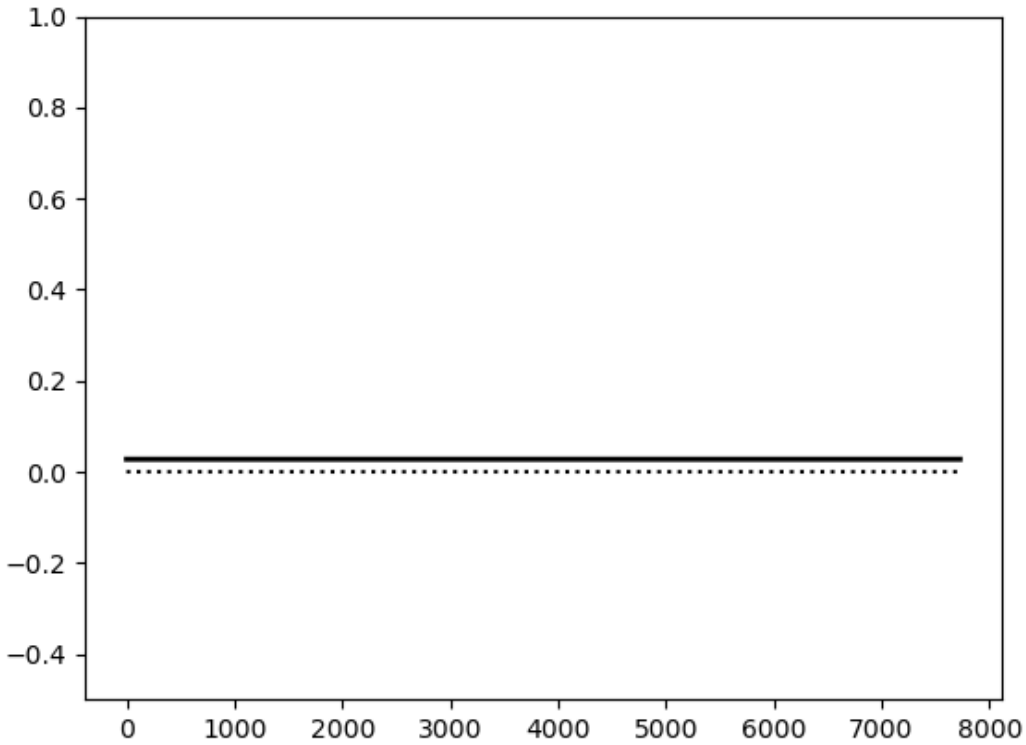
Statystyki opisowe – opis sygnału – histogram





Statystyki opisowe – opis sygnału

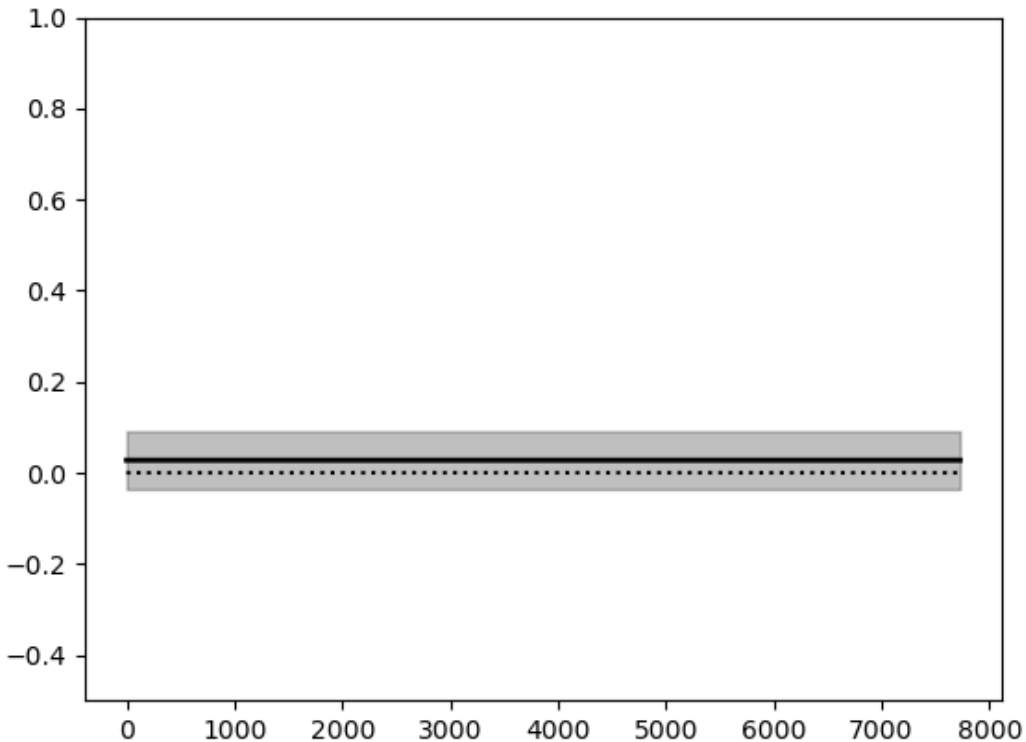
średnia $\mu = 0.028$ $\qquad \mu = \frac{1}{n}\sum_i^n x_i$



Statystyki opisowe – opis sygnału

średnia $\mu = 0.028$ $\quad \mu = \frac{1}{n}\sum_i^n x_i$

wariancja $\sigma^2 = 0.004$ $\quad \sigma^2 = \frac{1}{n}\sum_i^n (\mu - x_i)^2$ $\quad \sigma = 0.064$

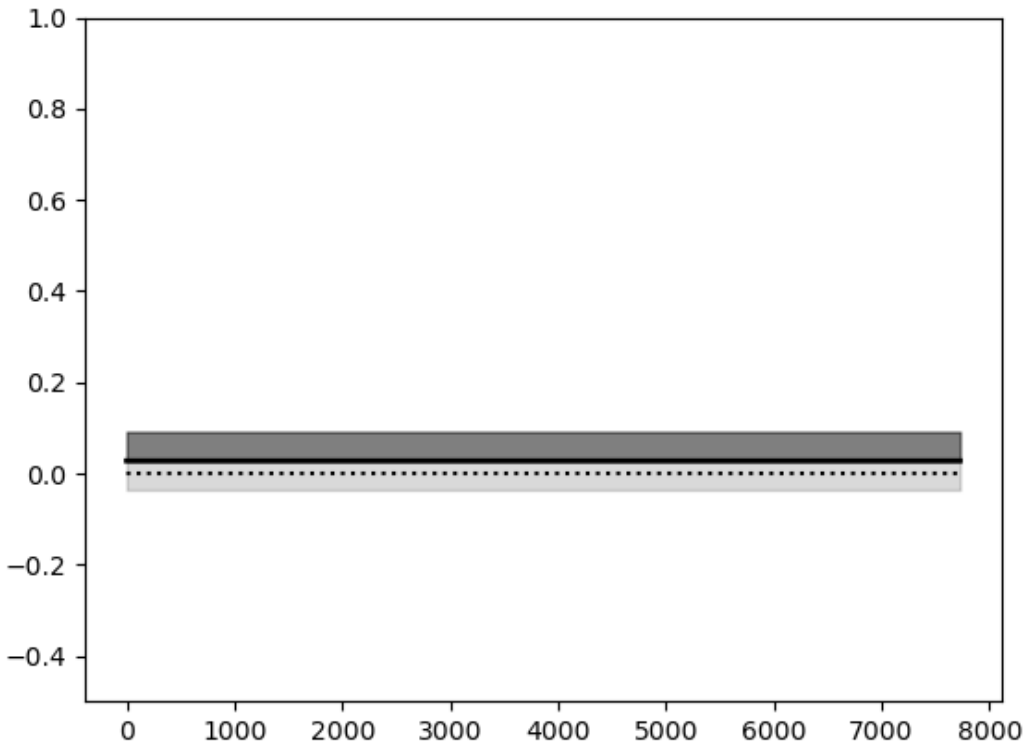


Statystyki opisowe – opis sygnału

średnia $\mu = 0.028$ $\quad \mu = \frac{1}{n}\sum_i^n x_i$

wariancja $\sigma^2 = 0.004$ $\quad \sigma^2 = \frac{1}{n}\sum_i^n (\mu - x_i)^2$ $\quad \sigma = 0.064$

skośność $s = 8.234$



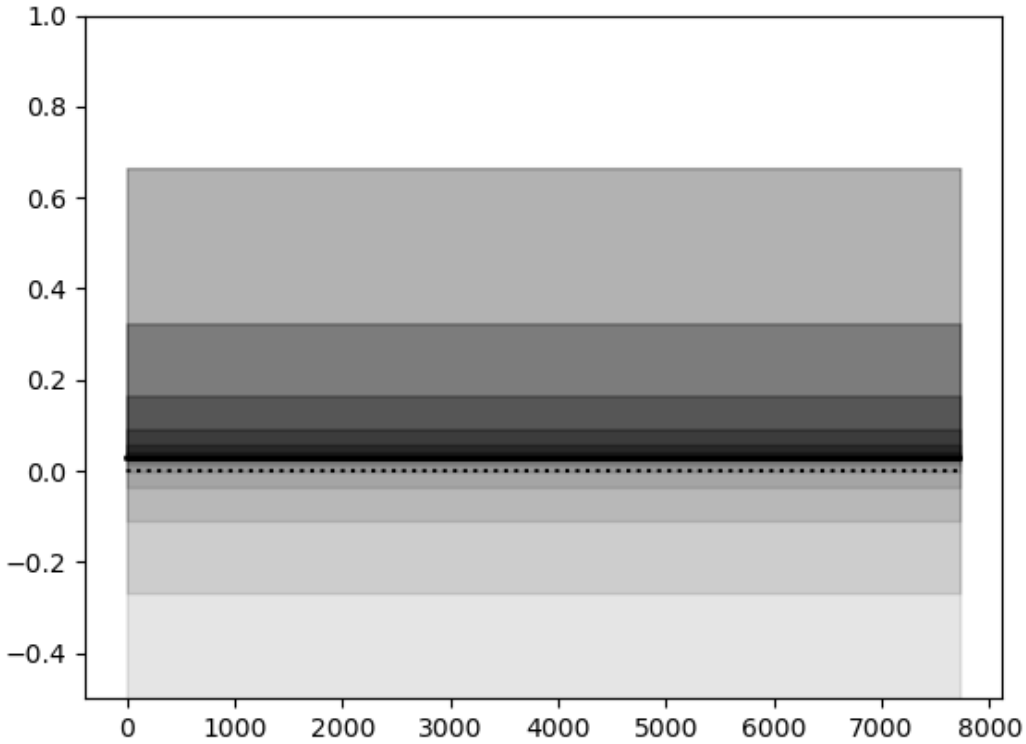
Statystyki opisowe – opis sygnału

średnia $\mu = 0.028$ $\quad \mu = \frac{1}{n}\sum_i^n x_i$

wariancja $\sigma^2 = 0.004$ $\quad \sigma^2 = \frac{1}{n}\sum_i^n (\mu - x_i)^2$ $\quad \sigma = 0.064$

skośność $s = 8.234$

kurtoza $k = 193.664$



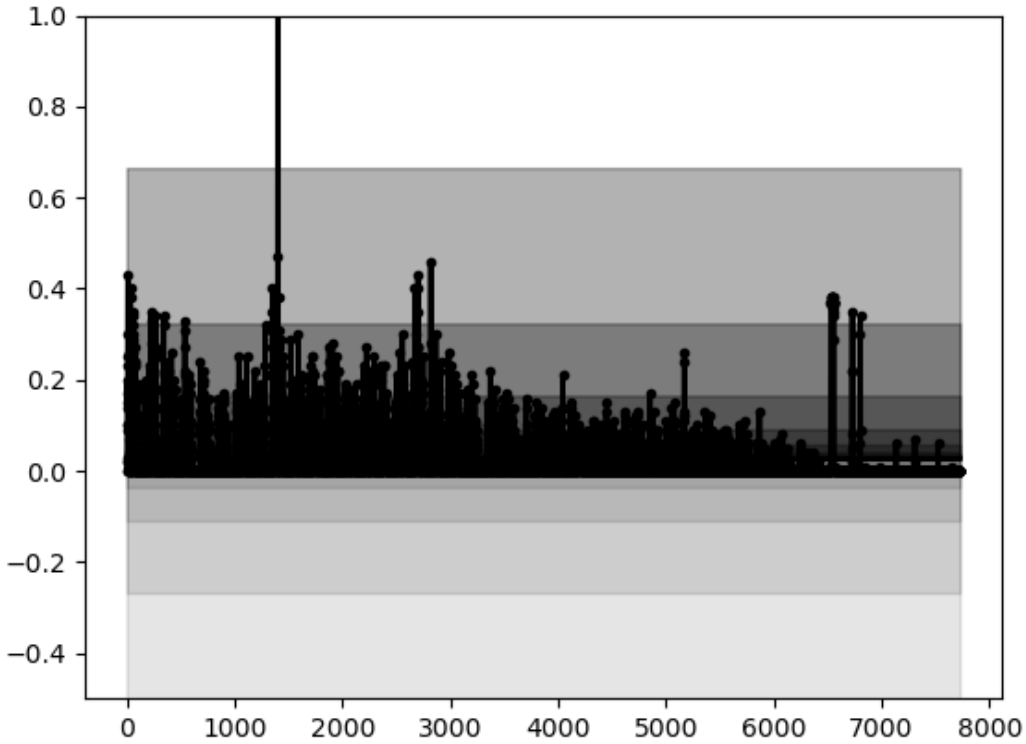
Statystyki opisowe – opis sygnału

średnia $\mu = 0.028$ $\quad \mu = \frac{1}{n}\sum_i^n x_i$

wariancja $\sigma^2 = 0.004$ $\quad \sigma^2 = \frac{1}{n}\sum_i^n (\mu - x_i)^2$ $\quad \sigma = 0.064$

skośność $s = 8.234$

kurtoza $k = 193.664$

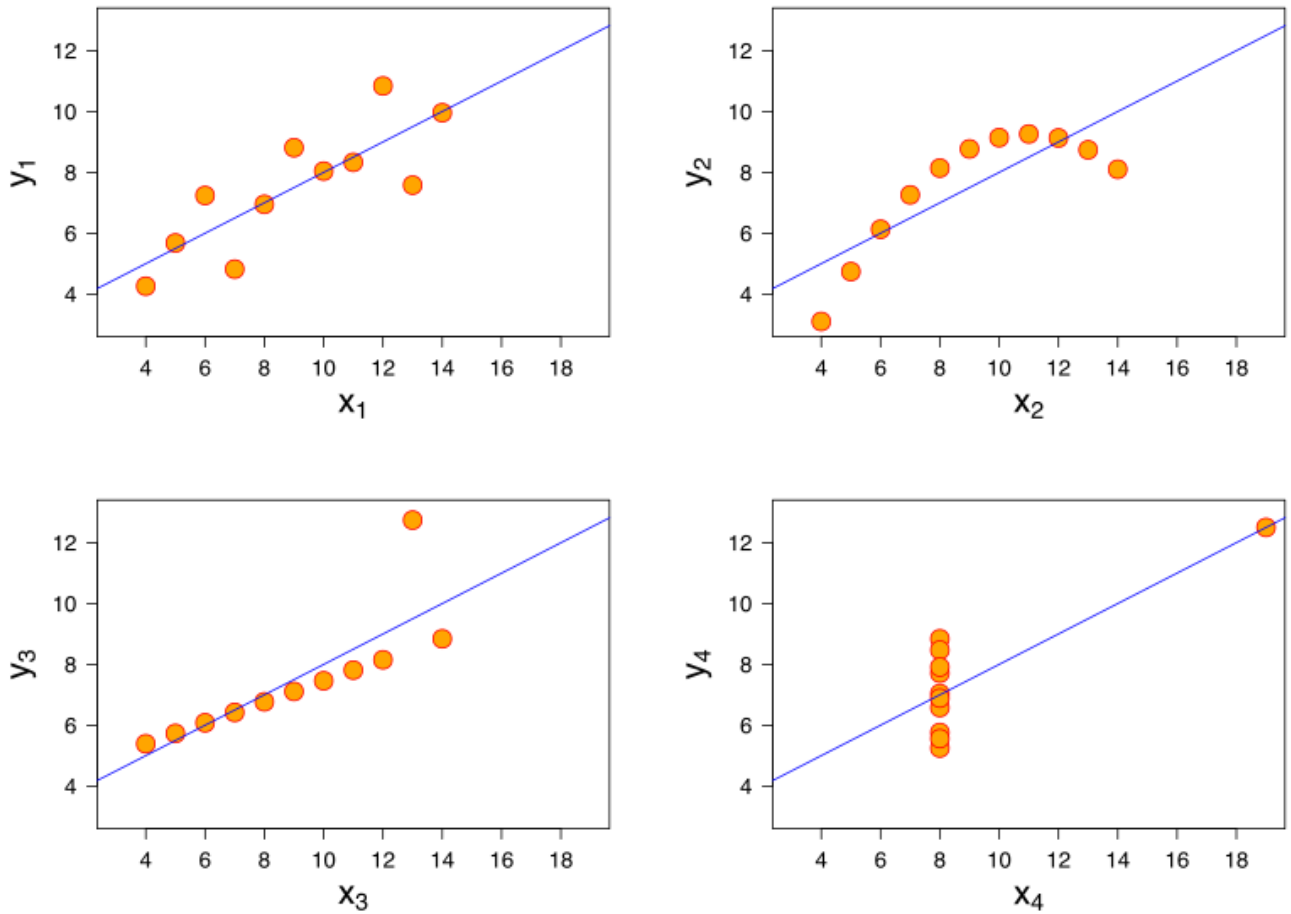


Kwartet Anscombe'a

Danych jest 11 par punktów (x, y) które mają następujące parametry statystyczne:

średnia x	9
wariancja x	11
średnia y	7.5
wariancja y	4.125
korelacja x i y	0.816
prosta regresji	$y = 3 + 0.5x$

Kwartet Anscombe'a



https://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg, User:Schutz + User:Avenue, CC SA 3

Statystki porządkowe (ang. order statistics)

1. Mediana
2. Kwartyle
3. Percentyle

$$\begin{bmatrix} -2. & 0.5 & 0.71 & 0.6 & 0.7 & -2.1 & 0.59 & 0.51 \end{bmatrix}$$

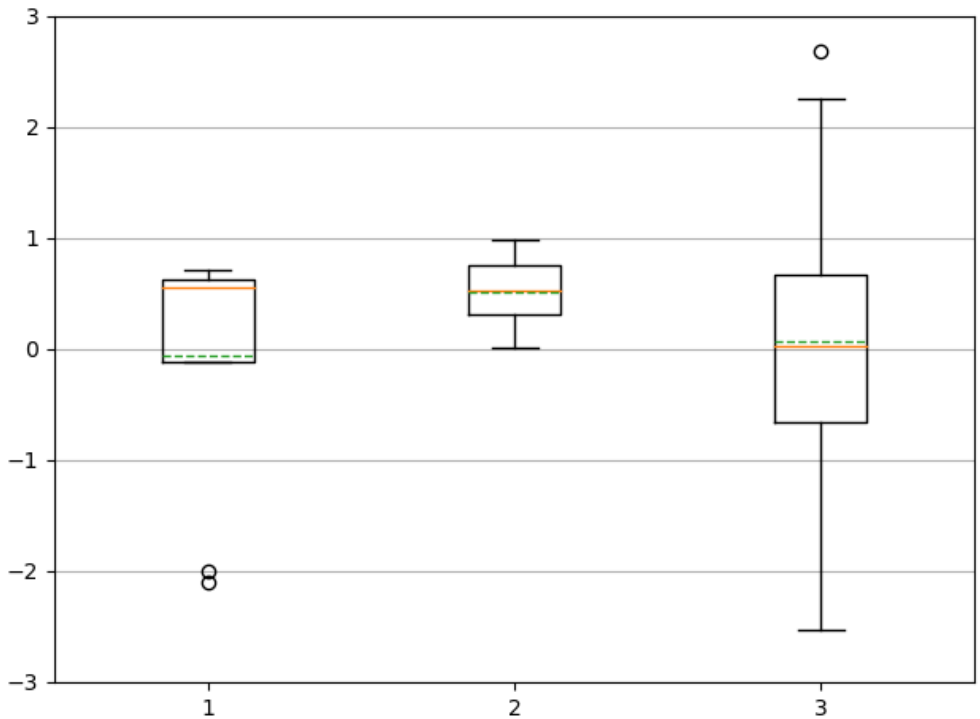
$$\begin{bmatrix} -2.1 & -2. & 0.5 & 0.51 & 0.59 & 0.6 & 0.7 & 0.71 \end{bmatrix}$$

średnia -0.06125

mediana 0.55

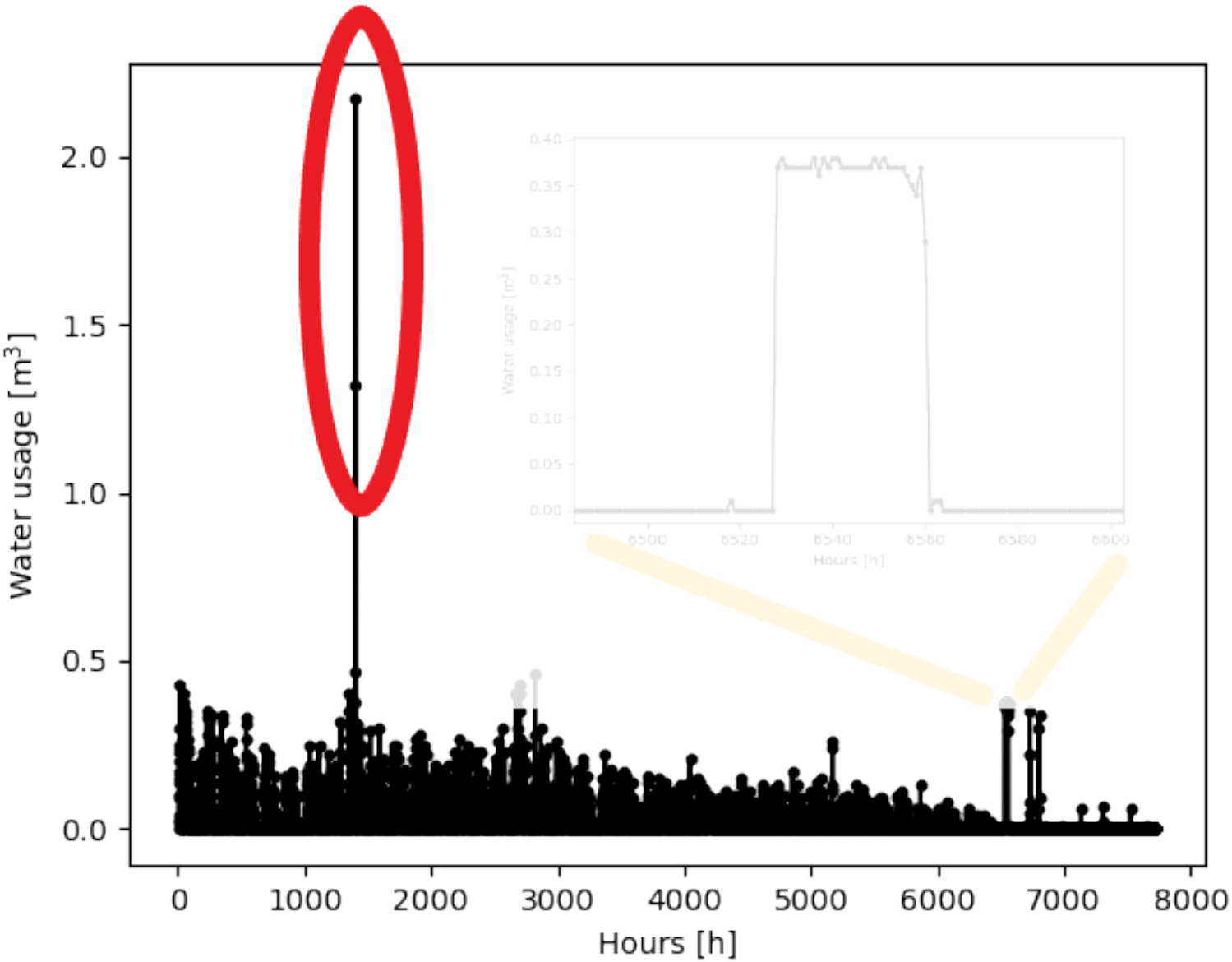
Statystki porządkowe (ang. order statistics) – wykres pudełkowy (ang. boxplot)

1. $\begin{bmatrix} -2. & 0.5 & 0.71 & 0.6 & 0.7 & -2.1 & 0.59 & 0.51 \end{bmatrix}$
2. 100 punktów z rozkładu równomiernego (ang. uniform) $\langle 0, 1 \rangle$
3. 100 punktów ze standardowego rozkładu normalnego (ang. standard normal distribution)



Weryfikacja z wykorzystaniem statystyk

→ Algorytm Median Absolute Deviation (MAD)



Algorytm Median Absolute Deviation (MAD)

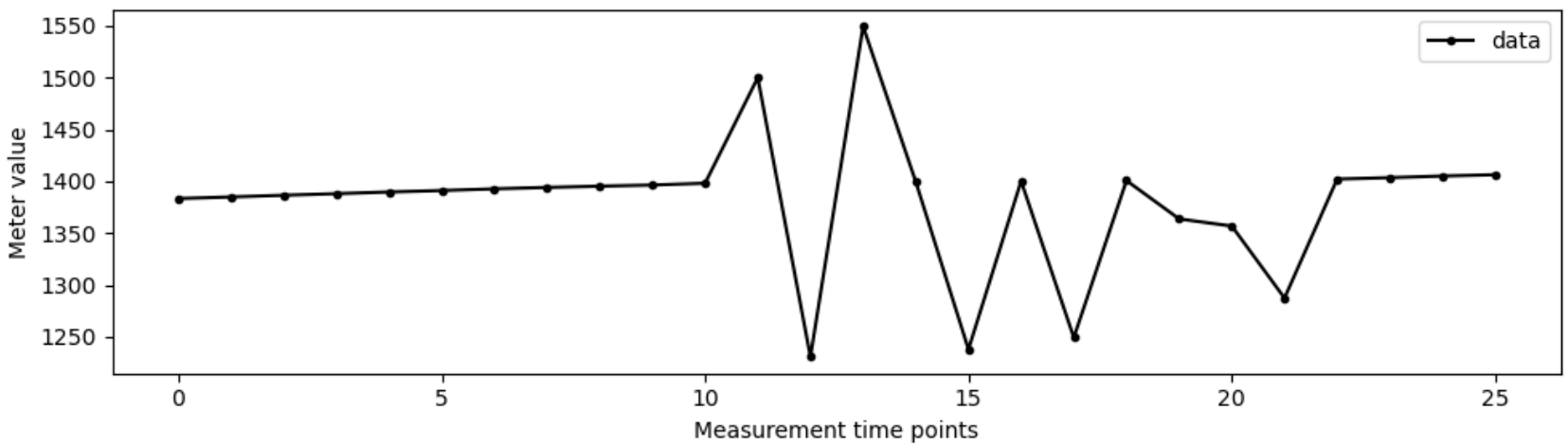
1. Dane (przykładowe odczyty) $\mathbf{x} = \Big[x_1, x_2, \ldots, x_n\Big]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \left[x_1, x_2, \ldots, x_n\right]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

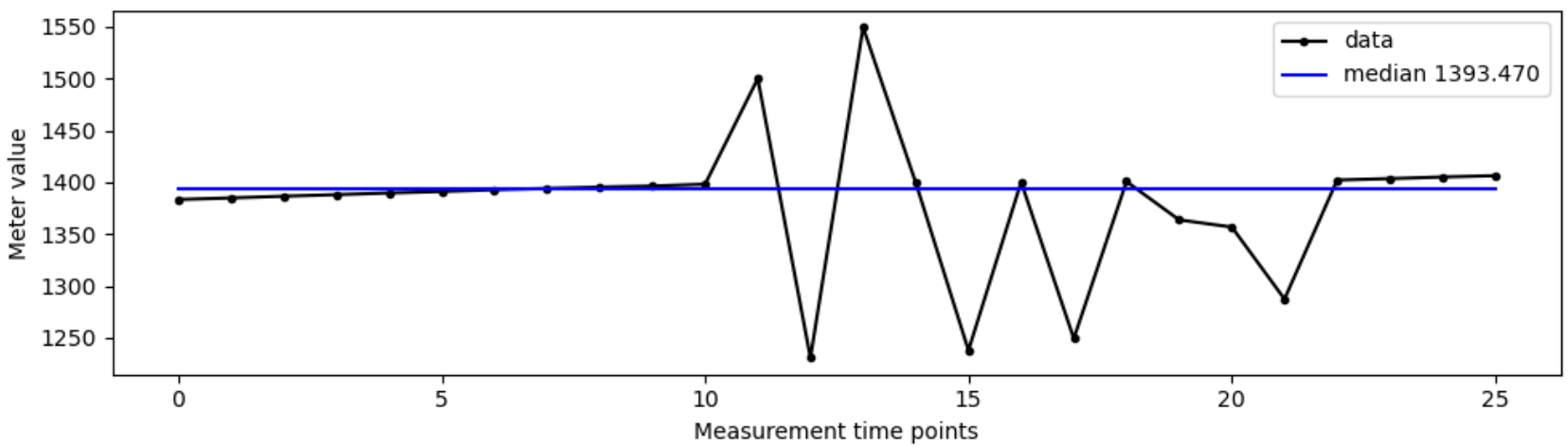


Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \left[x_1, x_2, \ldots, x_n\right]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$

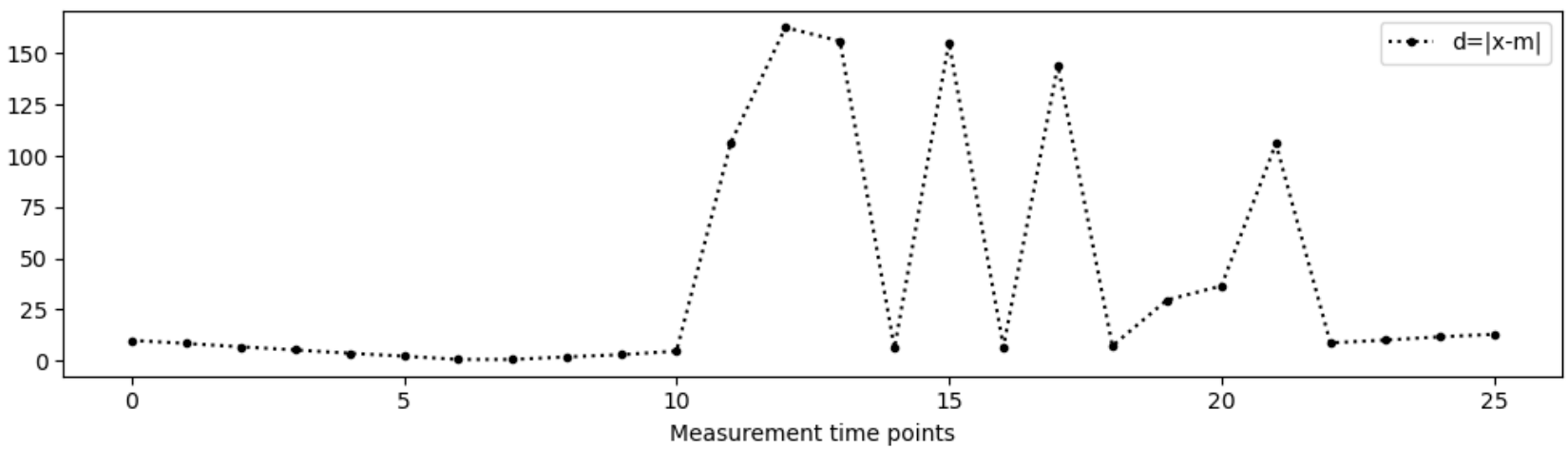


Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \Big[x_1, x_2, \ldots, x_n\Big]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$
3. $d_i = |x_i - m| \quad \mathbf{d} = \Big[d_1, d_2, \ldots, d_n\Big]$

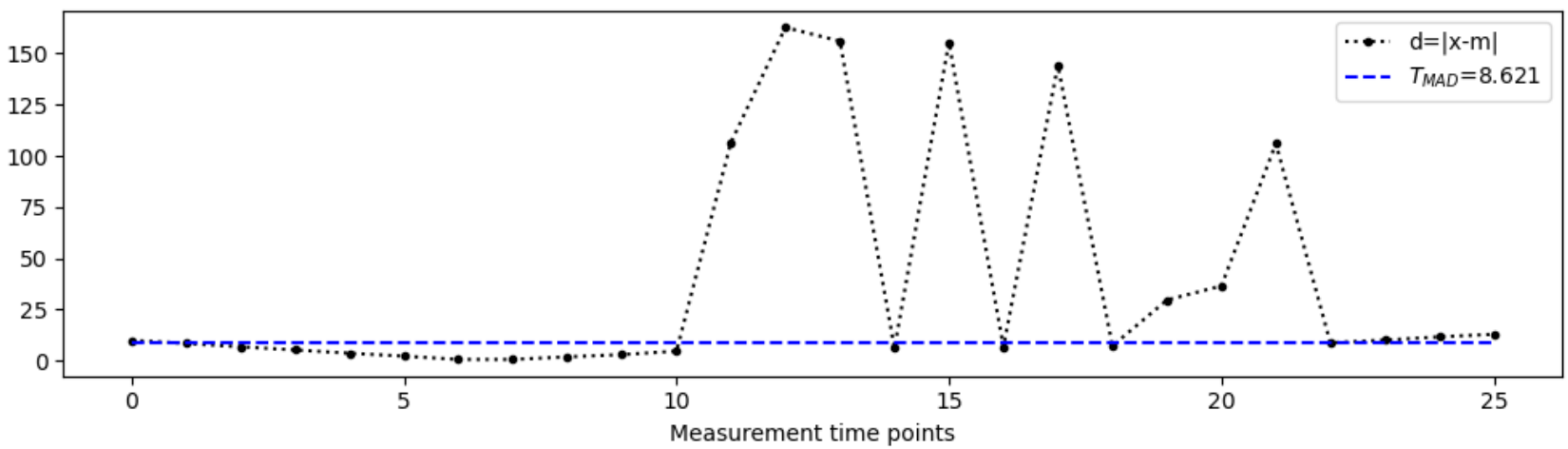


Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \Big[x_1, x_2, \ldots, x_n\Big]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$
3. $d_i = |x_i - m| \quad \mathbf{d} = \Big[d_1, d_2, \ldots, d_n\Big]$
4. $T_{\text{MAD}} = \text{median}(\mathbf{d})$

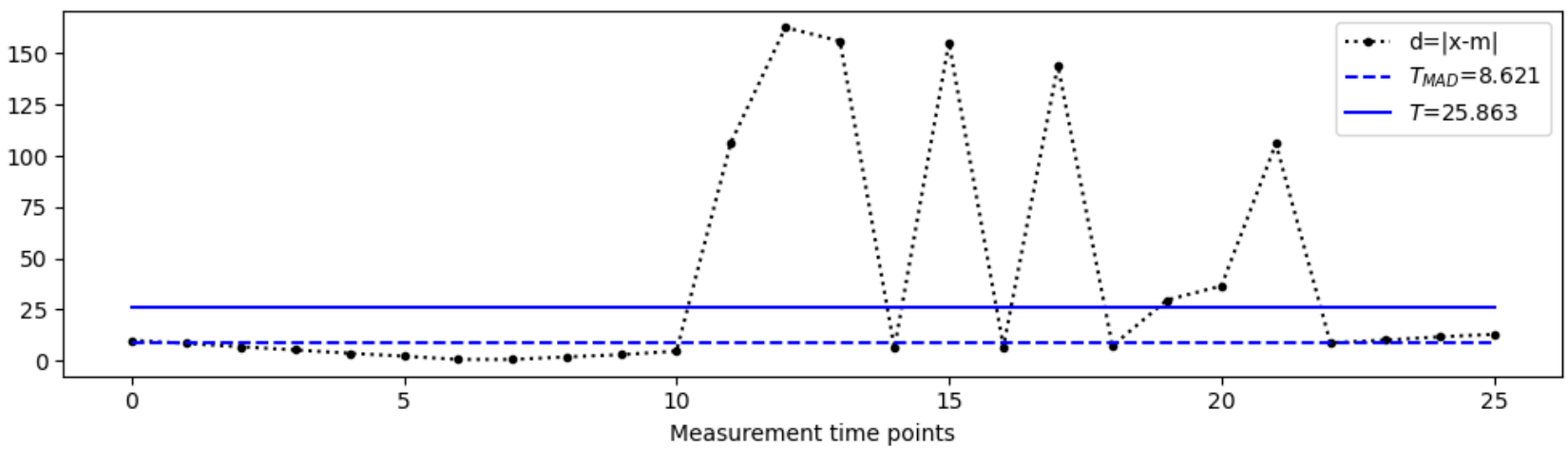


Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \left[x_1, x_2, \ldots, x_n\right]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$
3. $d_i = |x_i - m| \quad \mathbf{d} = \left[d_1, d_2, \ldots, d_n\right]$
4. $T_{\text{MAD}} = \text{median}(\mathbf{d})$
5. $T = 3 \times T_{\text{MAD}}$, $d_i > T \rightarrow$ błędny pomiar

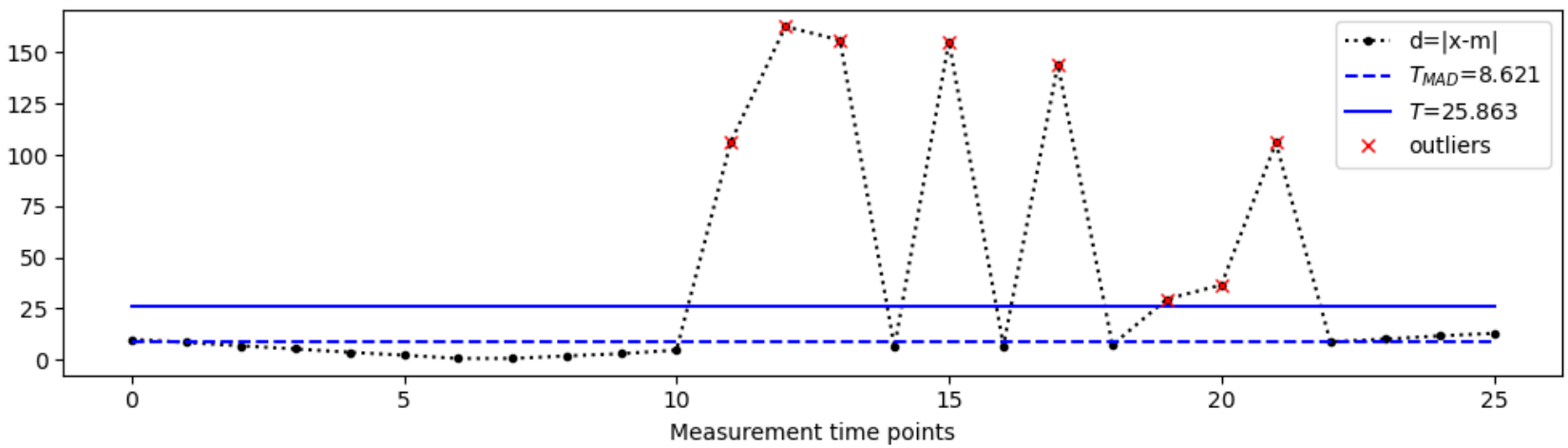


Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \left[x_1, x_2, \ldots, x_n\right]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$
3. $d_i = |x_i - m| \quad \mathbf{d} = \left[d_1, d_2, \ldots, d_n\right]$
4. $T_{\text{MAD}} = \text{median}(\mathbf{d})$
5. $T = 3 \times T_{\text{MAD}}$, $d_i > T \rightarrow$ błędny pomiar

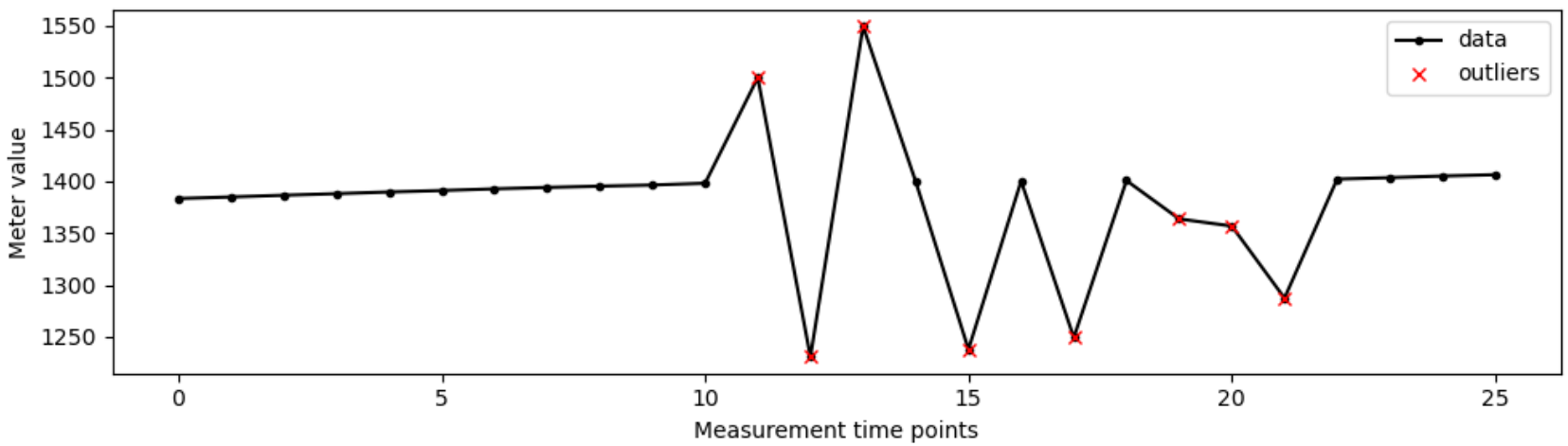


Algorytm Median Absolute Deviation (MAD)

1. Dane (przykładowe odczyty) $\mathbf{x} = \left[x_1, x_2, \ldots, x_n\right]$

```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$
3. $d_i = |x_i - m| \quad \mathbf{d} = \left[d_1, d_2, \ldots, d_n\right]$
4. $T_{\text{MAD}} = \text{median}(\mathbf{d})$
5. $T = 3 \times T_{\text{MAD}}$, $d_i > T \rightarrow$ błędny pomiar

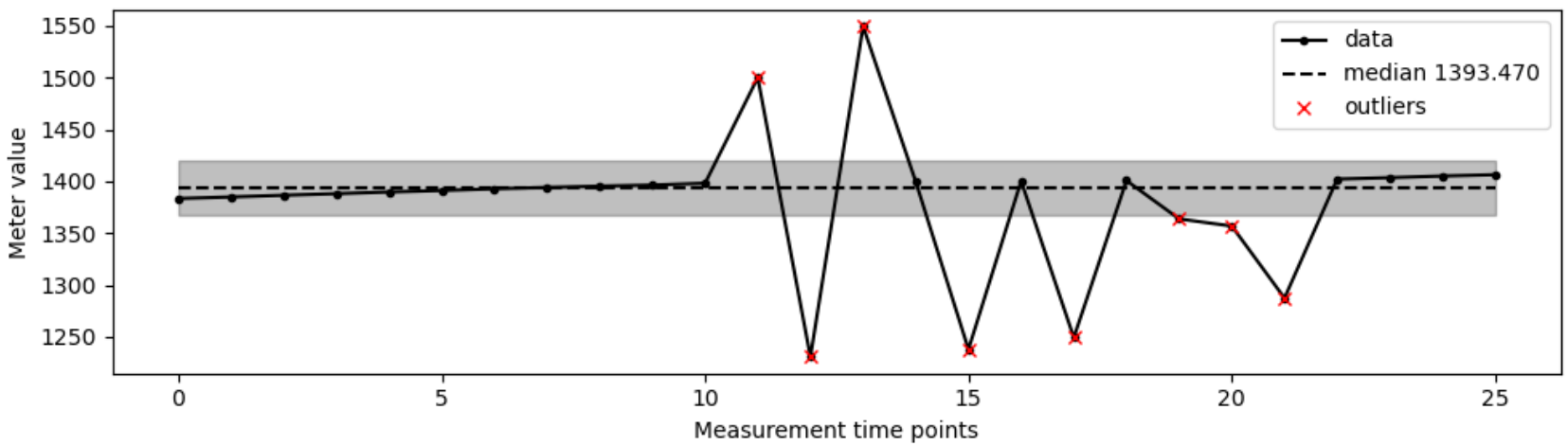


Algorytm Median Absolute Deviation (MAD)

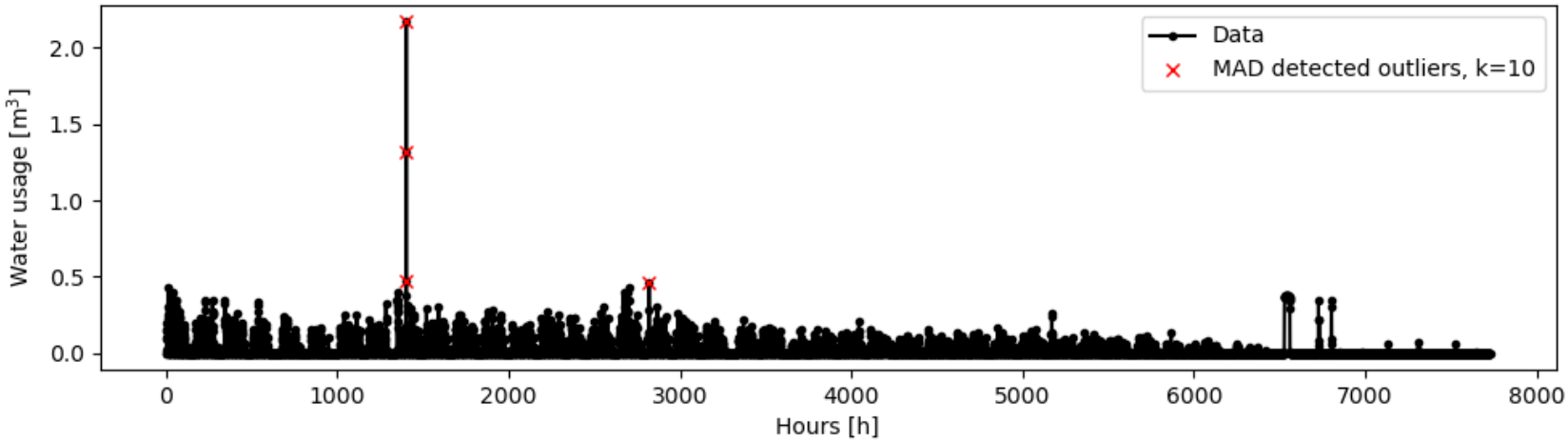
1. Dane (przykładowe odczyty) $\mathbf{x} = \left[x_1, x_2, \ldots, x_n\right]$

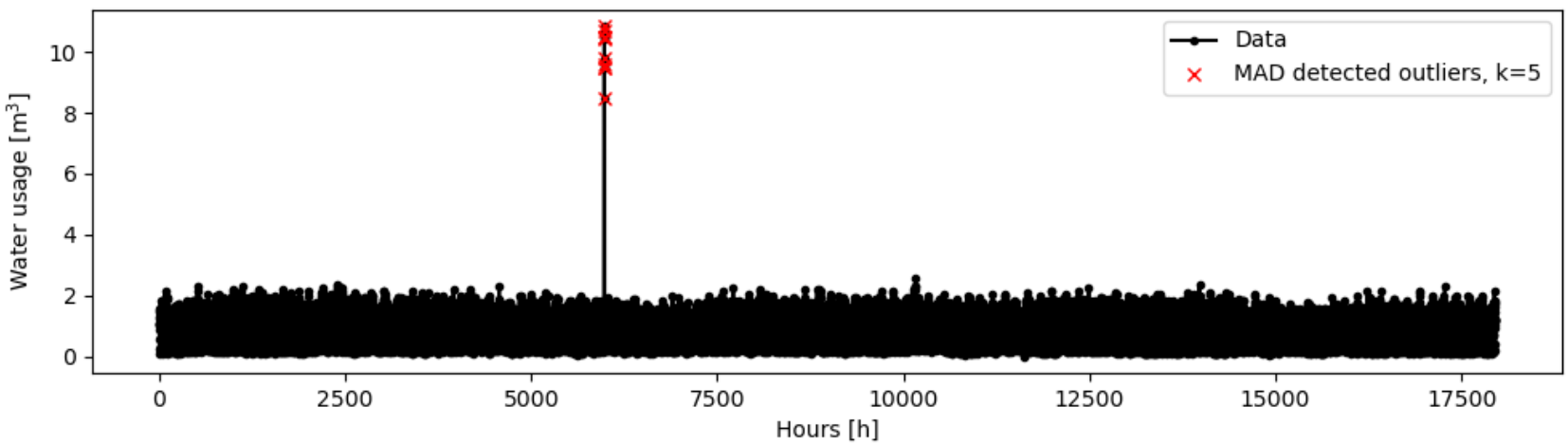
```
x = [1383.464 1384.9871 1386.6281 1388.142 1389.766 1391.188 1392.796 1394.144 1395.4139
1396.488 1398.209 1500.011 1230.793 1549.475 1399.6281 1238.033 1400.318 1249.4 1400.764
1363.806 1356.909 1287.488 1402.229 1403.588 1405.167 1406.427 ]
```

2. $m = \text{median}(\mathbf{x})$
3. $d_i = |x_i - m| \quad \mathbf{d} = \left[d_1, d_2, \ldots, d_n\right]$
4. $T_{\text{MAD}} = \text{median}(\mathbf{d})$
5. $T = 3 \times T_{\text{MAD}}$, $d_i > T \rightarrow$ błędny pomiar

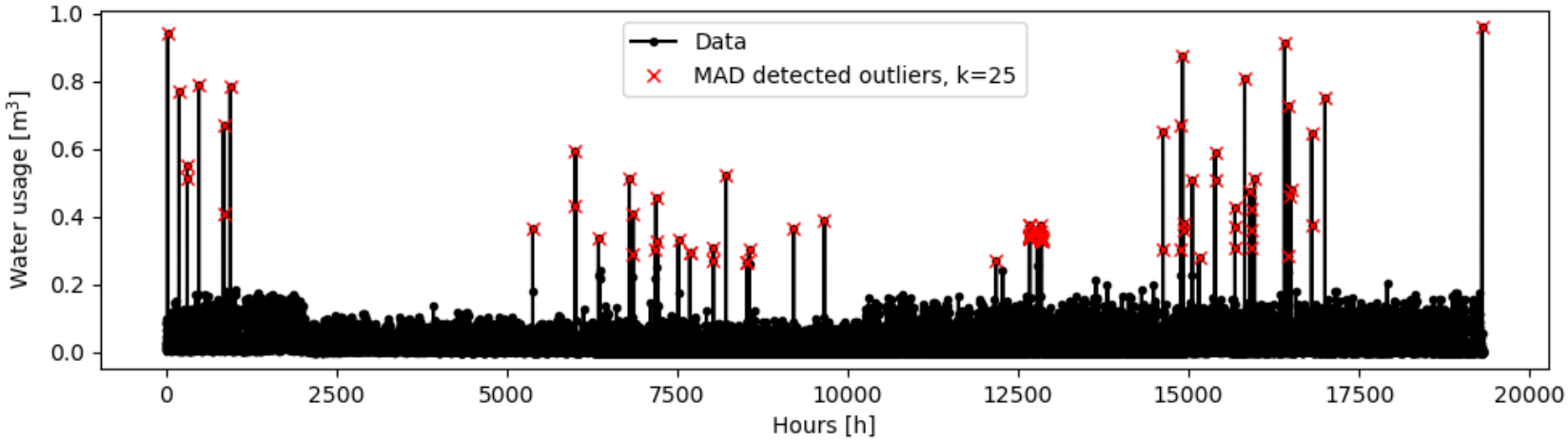


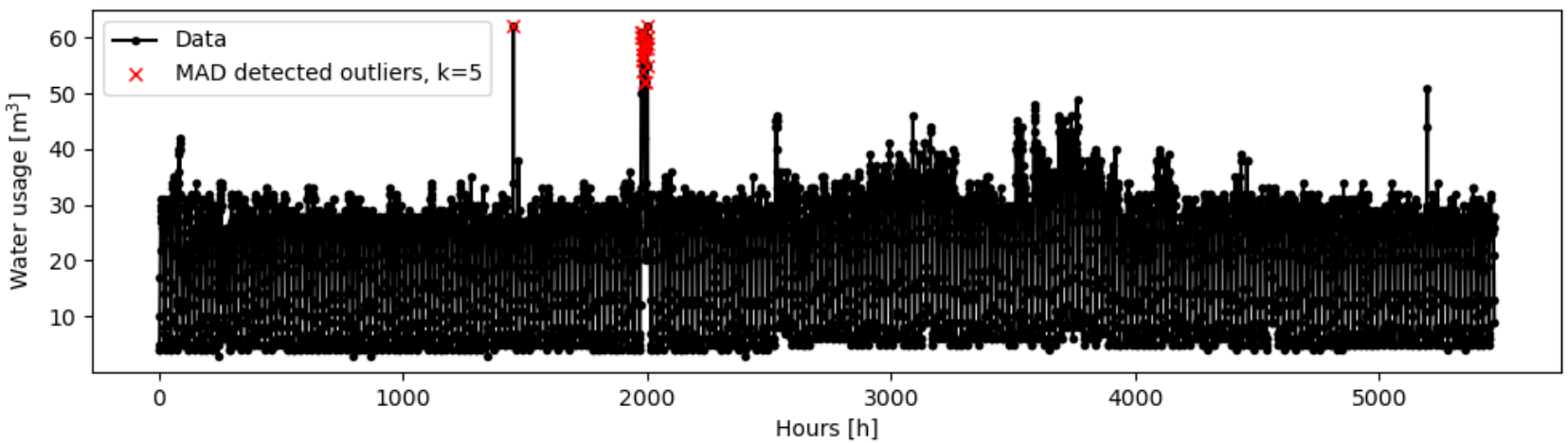
Algorytm MAD – przykłady





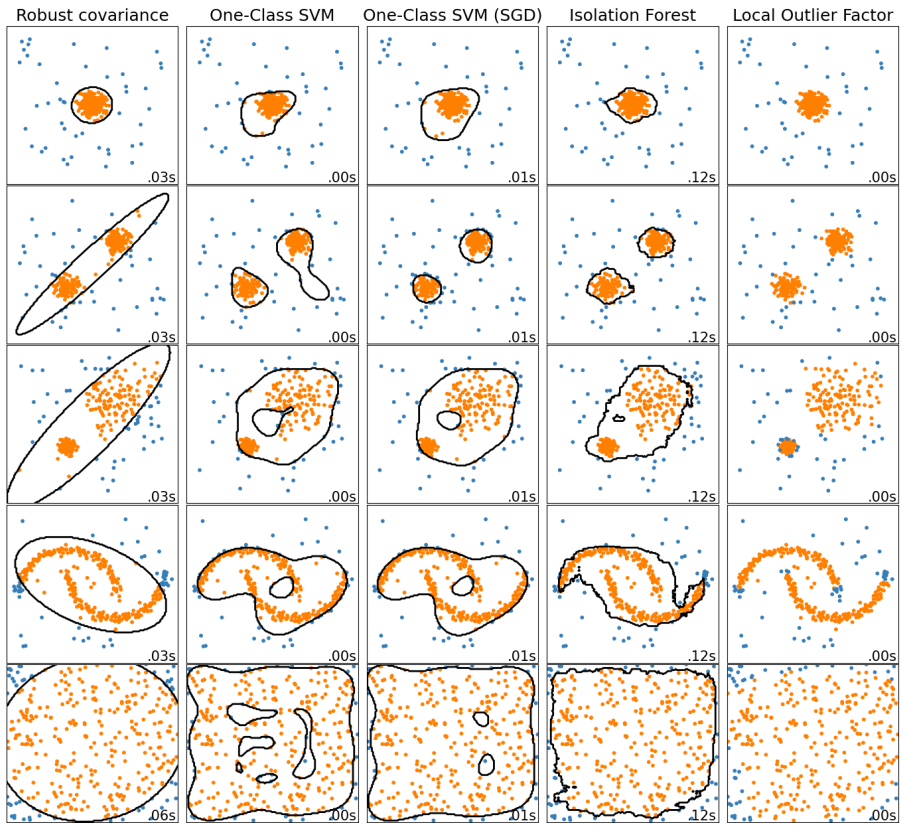
Algorytm MAD – przykłady





Weryfikacja statystyczna – outliery

1. Wykrywanie danych odstających (ang. outliers)
   - 1.1 „odstających” $\approx$ spoza zakresu danych
   - 1.2 MAD (i wiele innych algorytmów)[1]
2. Weryfikacja
   - 2.1 Wyizolowanie pomiarów błędnych – błędy urządzeń
   - 2.2 Wykrycie anomalii i zmian trendu



[1] scikit-learn.org/stable/auto_examples/miscellaneous/plot_anomaly_comparison.html, scikit-learn devs, BSD

Plan wykładu

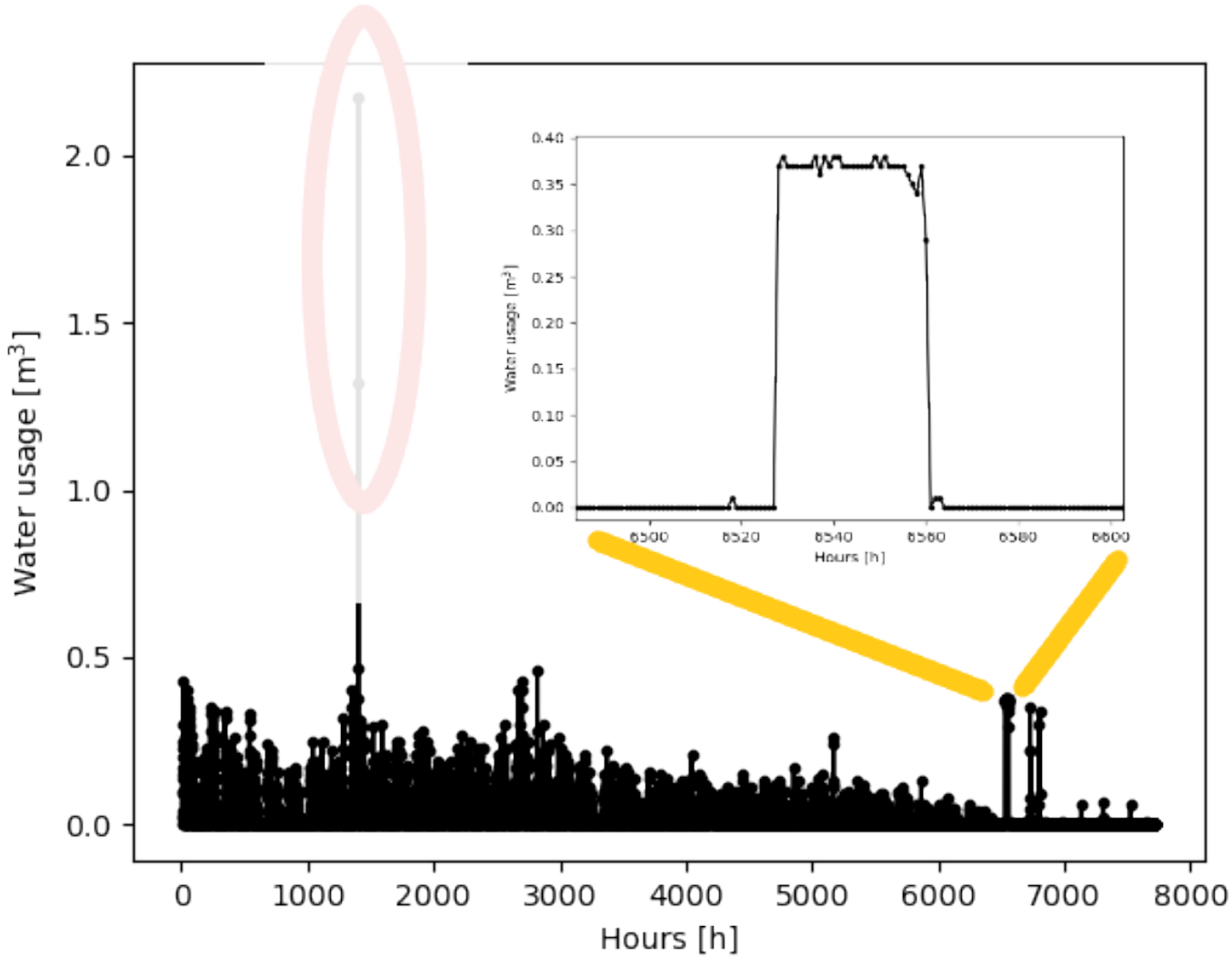
Analiza przykładowego sygnału

Weryfikacja statystyczna – MAD

Weryfikacja statystyczna – RX

Weryfikacja z wykorzystaniem statystyk „strikes back”

→ Algorytm Reed-Xiaoli (RX)



Dlaczego średnia (i inne statystyki np. wariancja) jest dobra?

Dlaczego średnia (i inne statystyki np. wariancja) jest dobra?

1. Łatwa do policzenia

Dlaczego średnia (i inne statystyki np. wariancja) jest dobra?

1. Łatwa do policzenia
2. Znane właściwości

Dlaczego średnia (i inne statystyki np. wariancja) jest dobra?

1. Łatwa do policzenia
2. Znane właściwości
3. Element wnioskowania statystycznego, część bardziej złożonych metod

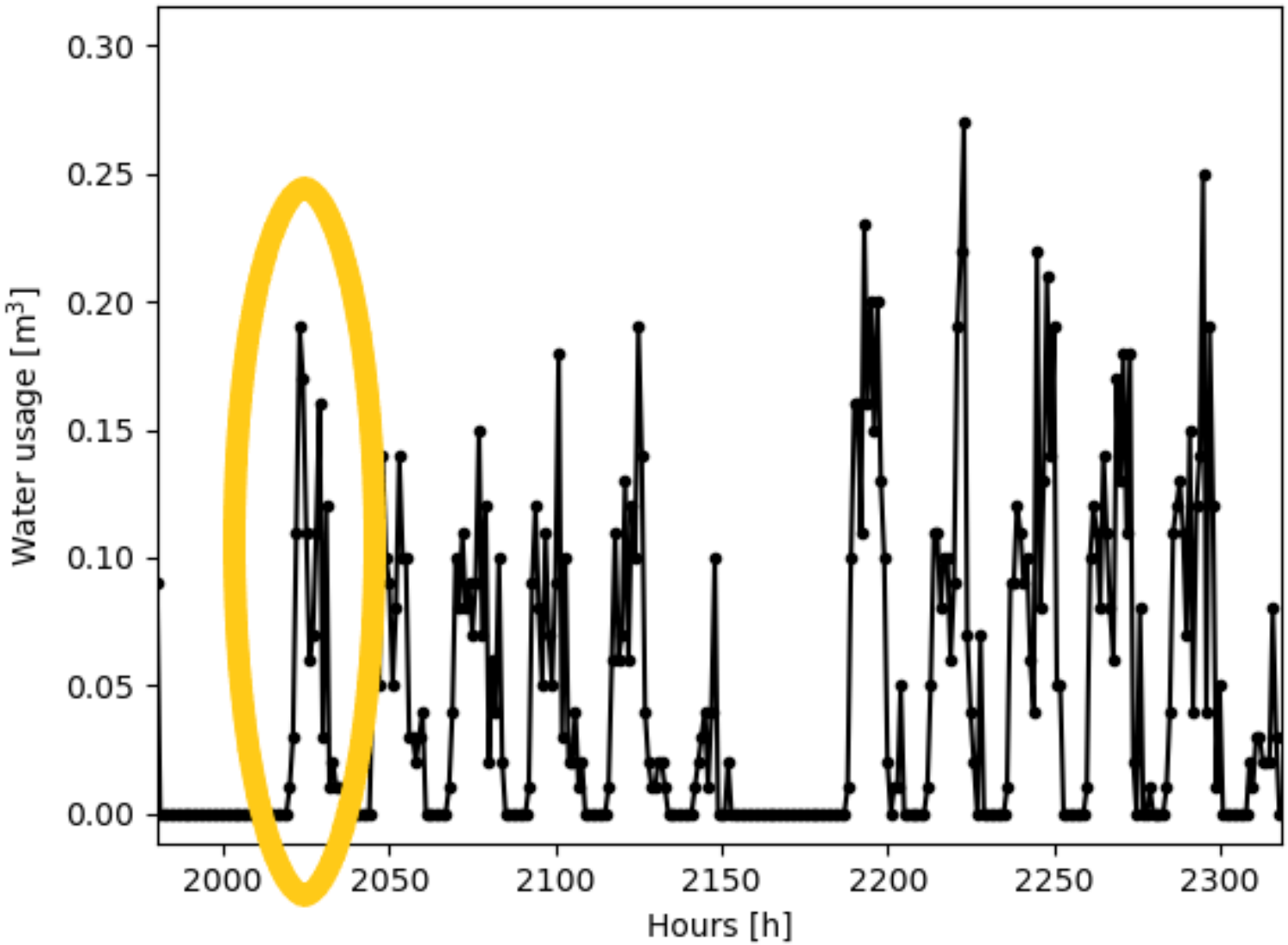
Dlaczego średnia (i inne statystyki np. wariancja) jest dobra?

1. Łatwa do policzenia
2. Znane właściwości
3. Element wnioskowania statystycznego, część bardziej złożonych metod

outlier wartość spoza rozkładu danych, prawdopodobnie błędna - np. błąd urządzenia rejestrującego

anomalia wartość z rozkładu danych, ale rzadka – np. wyciek z instalacji

Wyjście poza pojedyncze punkty danych



Okresy dobowe

1. Wykorzystujemy właściwości sygnału – okresowość dobową

```
[[4.43 4.77 6.1  ... 7.97 4.85 4.29]
[4.22 4.18 5.18 ... 9.1  6.37 5.06]
[4.86 5.9  6.54 ... 8.92 6.54 5.19]
...
[2.72 2.52 2.66 ... 9.17 6.59 4.8 ]
[3.51 2.9  2.52 ... 8.79 5.34 3.59]
[2.62 2.72 2.57 ... 9.71 6.01 3.41]]
```

$X = \left[x_{ij}\right]$ i dni j godziny, $\mathbf{x}_i$ i-ta doba

Okresy dobowe

1. Wykorzystujemy właściwości sygnału – okresowość dobową

```
[[4.43 4.77 6.1  ... 7.97 4.85 4.29]
[4.22 4.18 5.18 ... 9.1  6.37 5.06]
[4.86 5.9  6.54 ... 8.92 6.54 5.19]
...
[2.72 2.52 2.66 ... 9.17 6.59 4.8 ]
[3.51 2.9  2.52 ... 8.79 5.34 3.59]
[2.62 2.72 2.57 ... 9.71 6.01 3.41]]
```

   $X = \left[x_{ij}\right]$ i dni j godziny, $\mathbf{x}_i$ i-ta doba

2. Anomalia $\rightarrow$ „nietypowe zachowanie”

Okresy dobowe

1. Wykorzystujemy właściwości sygnału – okresowość dobową

```
[[4.43 4.77 6.1  ... 7.97 4.85 4.29]
[4.22 4.18 5.18 ... 9.1  6.37 5.06]
[4.86 5.9  6.54 ... 8.92 6.54 5.19]
...
[2.72 2.52 2.66 ... 9.17 6.59 4.8 ]
[3.51 2.9  2.52 ... 8.79 5.34 3.59]
[2.62 2.72 2.57 ... 9.71 6.01 3.41]]
```

   $X = \left[x_{ij}\right]$ i dni j godziny, $\mathbf{x}_i$ i-ta doba

2. Anomalia $\longrightarrow$ „nietypowe zachowanie”
3. „Zachowanie” $\longrightarrow$ średnia dobowa $\bar{\mathbf{x}} = \left[\bar{x}_j\right]$ $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$

Okresy dobowe

1. Wykorzystujemy właściwości sygnału – okresowość dobową

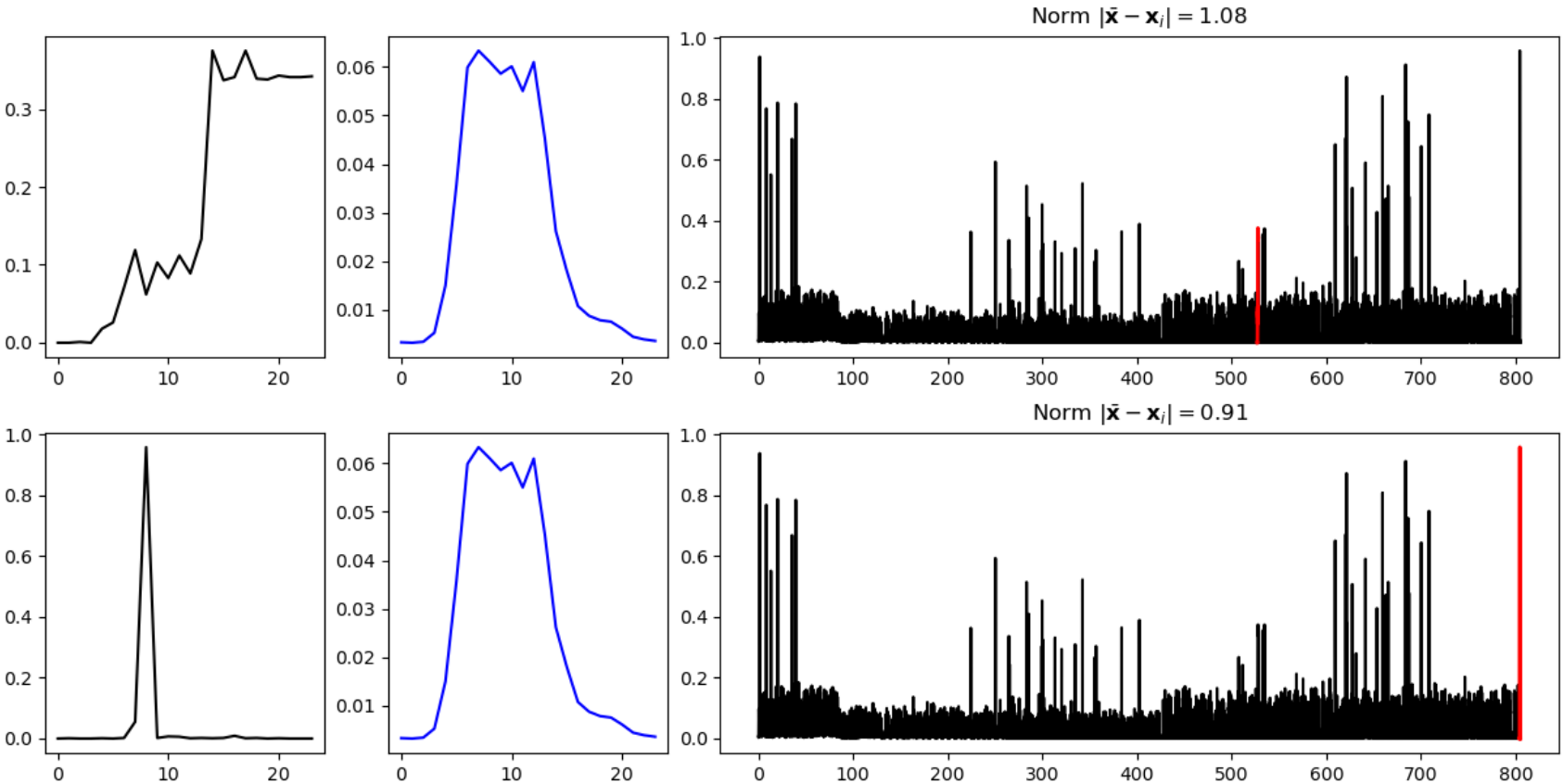
```
[[4.43 4.77 6.1  ... 7.97 4.85 4.29]
[4.22 4.18 5.18 ... 9.1  6.37 5.06]
[4.86 5.9  6.54 ... 8.92 6.54 5.19]
...
[2.72 2.52 2.66 ... 9.17 6.59 4.8 ]
[3.51 2.9  2.52 ... 8.79 5.34 3.59]
[2.62 2.72 2.57 ... 9.71 6.01 3.41]]
```

   $X = \left[x_{ij}\right]$ i dni j godziny, $\mathbf{x}_i$ i-ta doba

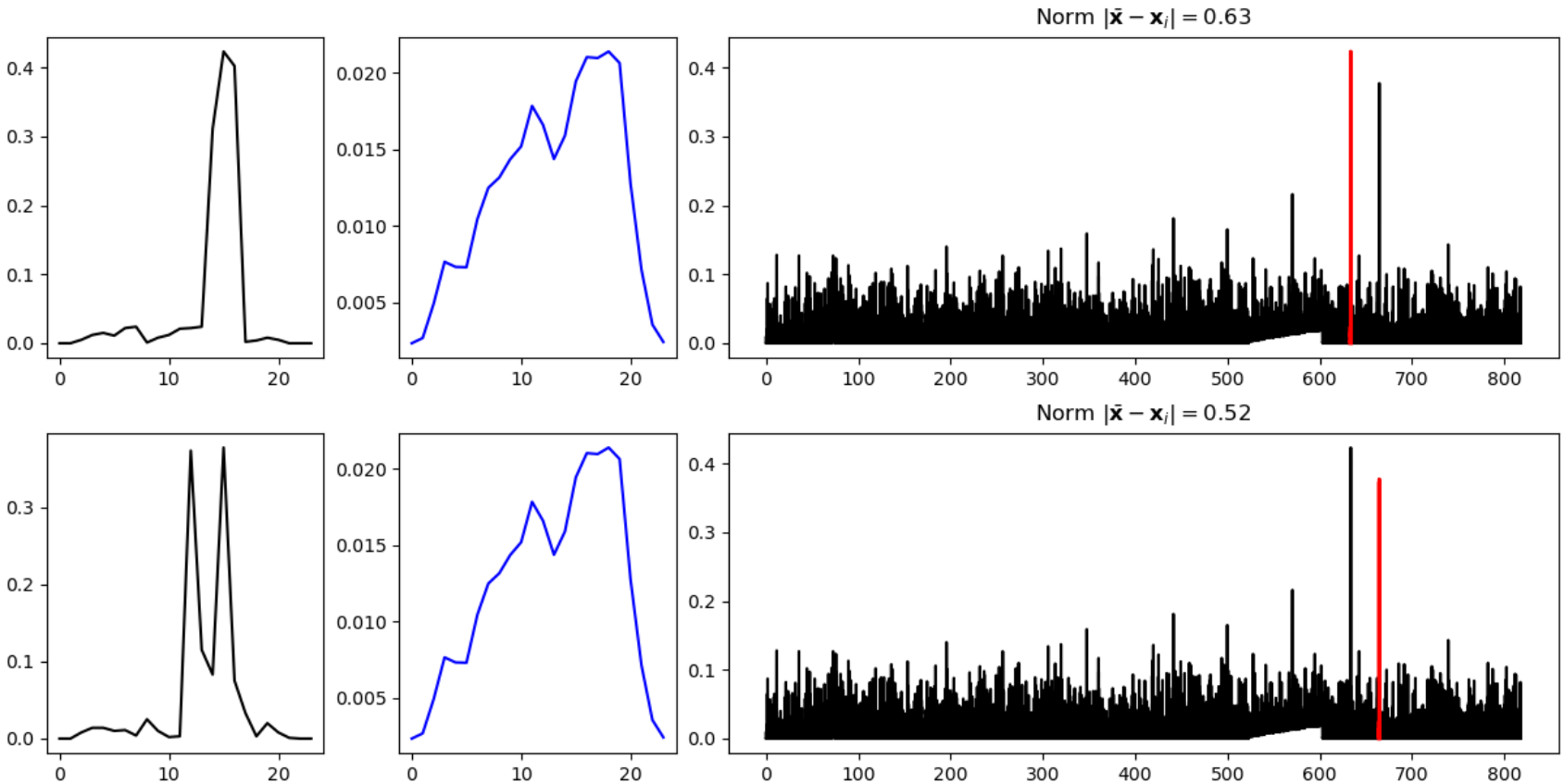
2. Anomalia $\longrightarrow$ „nietypowe zachowanie”
3. „Zachowanie” $\longrightarrow$ średnia dobowa $\bar{\mathbf{x}} = \left[\bar{x}_j\right]$ $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
4. „Nietypowe” $\longrightarrow$ duża odległość od średniej

   $\|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24}(\bar{x}_j - x_{ij})^2}$

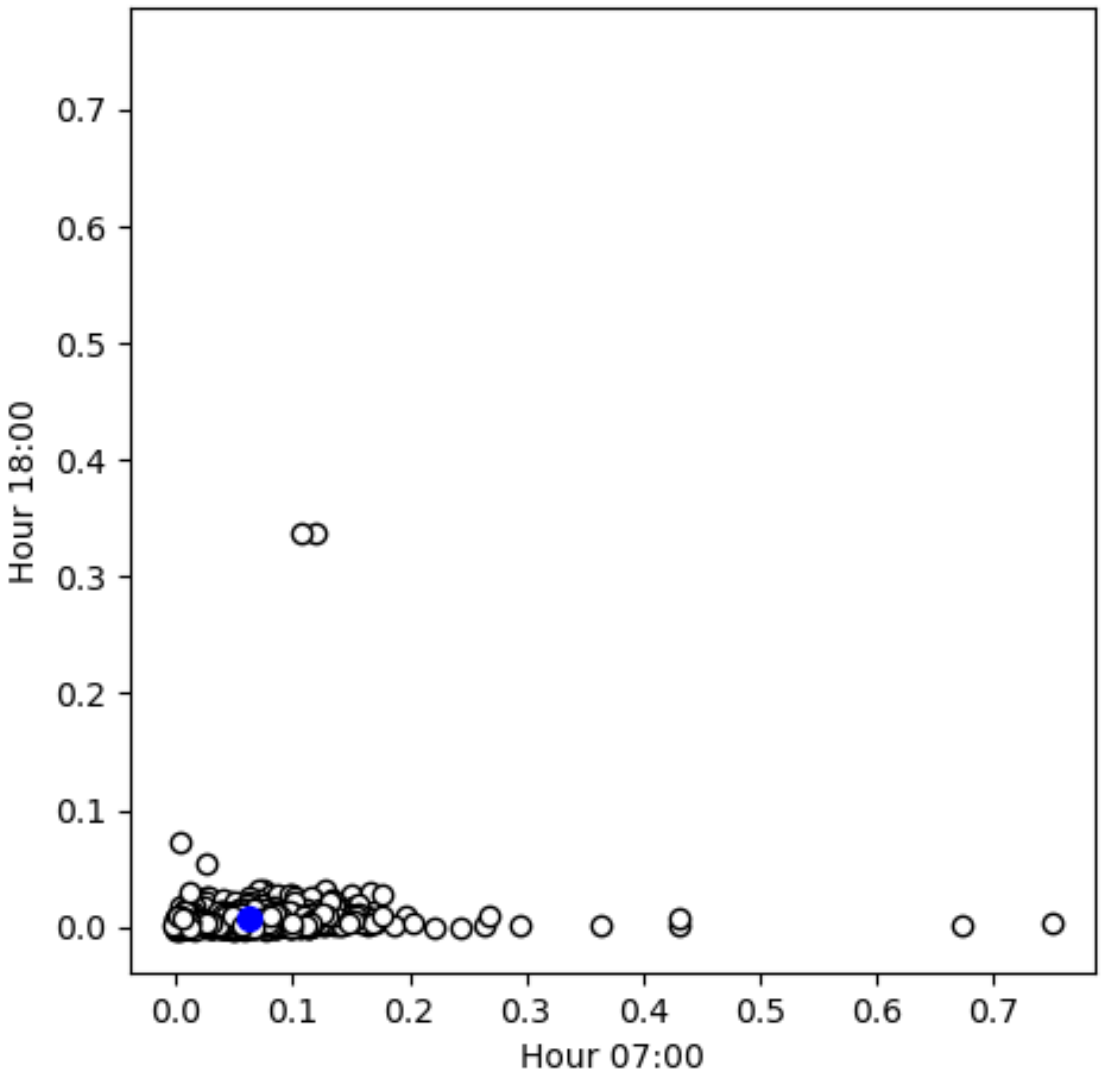
Problem?


Norm $|\bar{\mathbf{x}} - \mathbf{x}_i| = 1.08$
Norm $|\bar{\mathbf{x}} - \mathbf{x}_i| = 0.91$

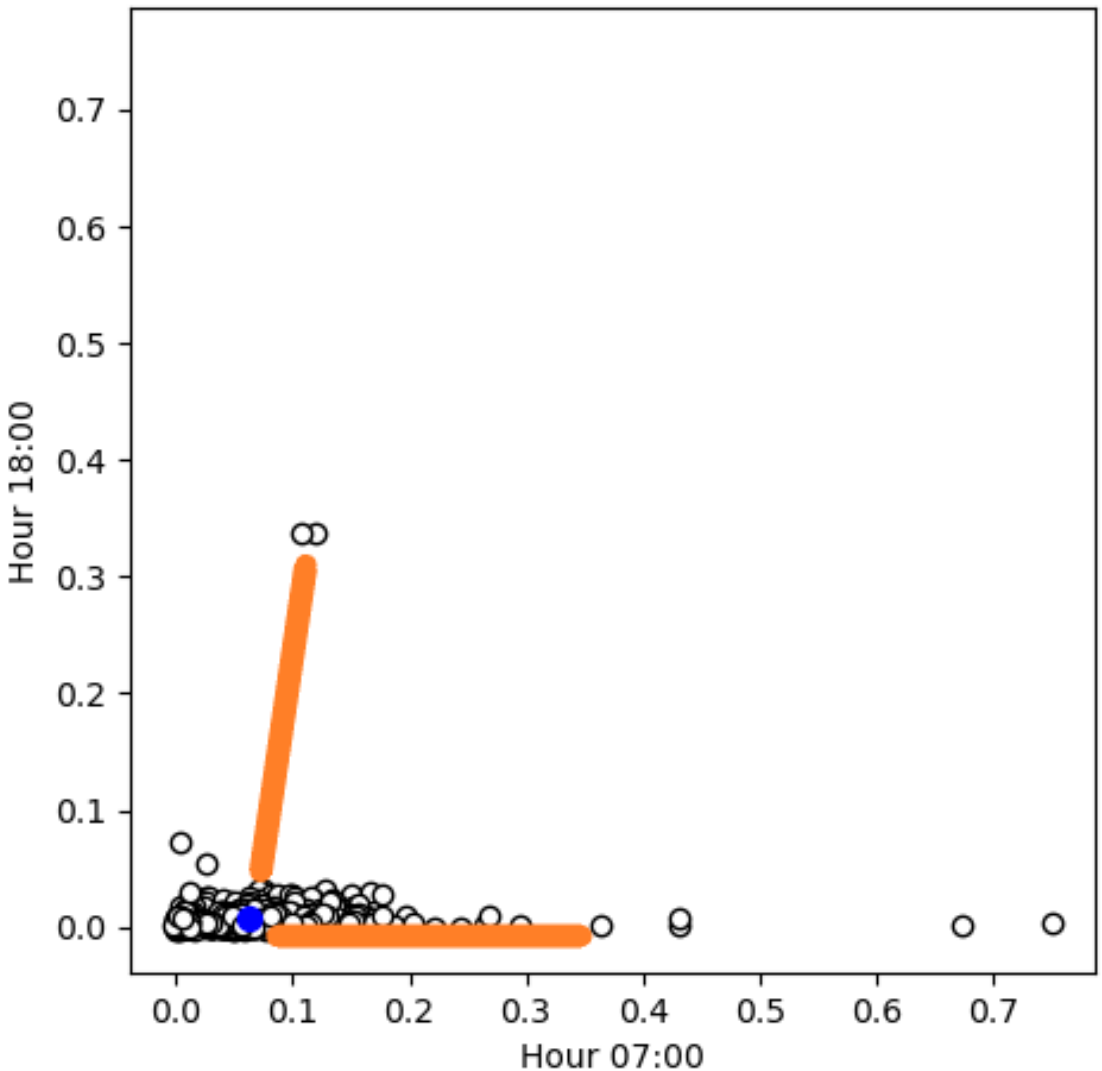
Problem?


Norm $|\bar{\mathbf{x}} - \mathbf{x}_i| = 0.63$
Norm $|\bar{\mathbf{x}} - \mathbf{x}_i| = 0.52$

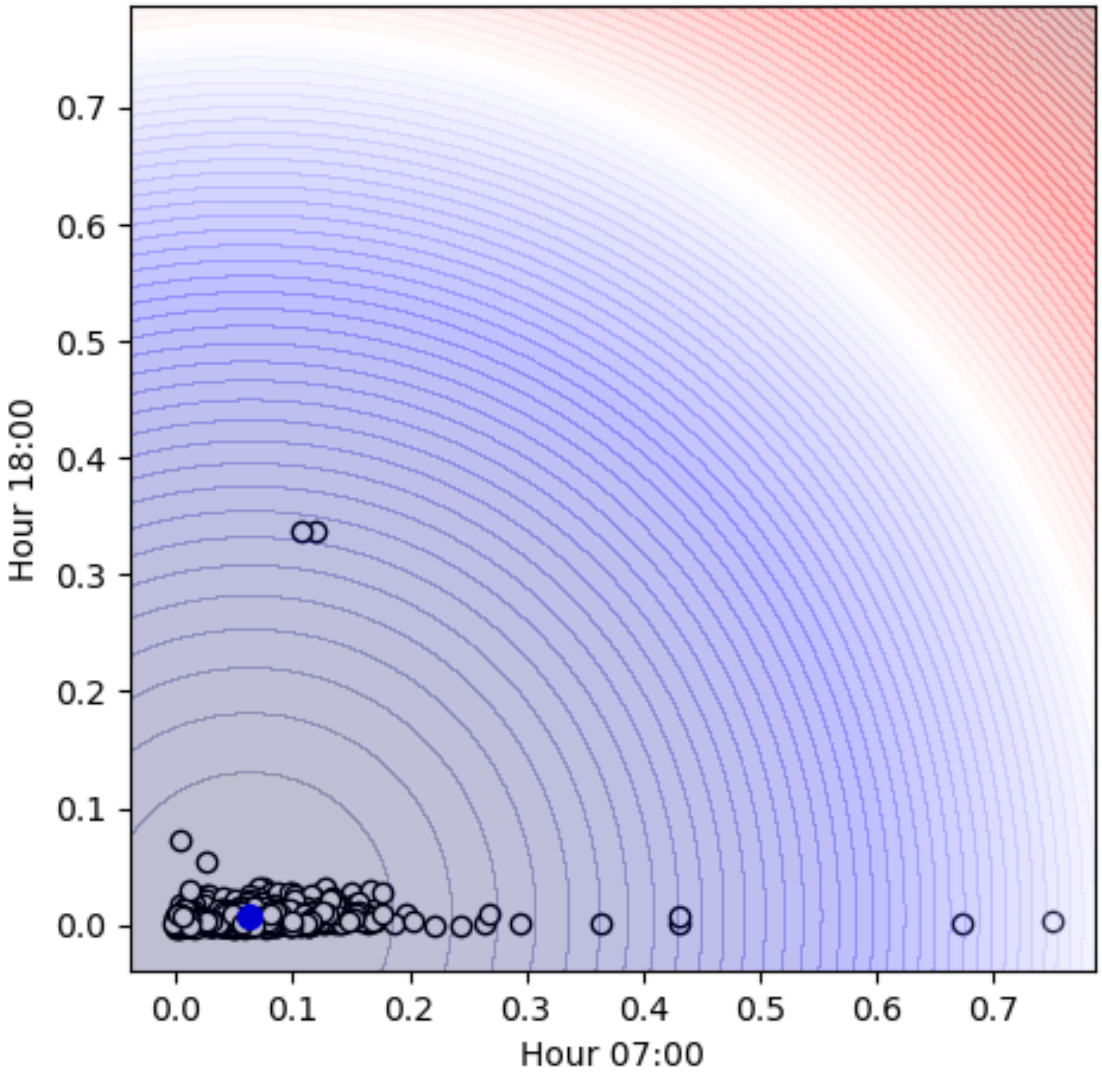
Odległość od średniej a rozkład danych



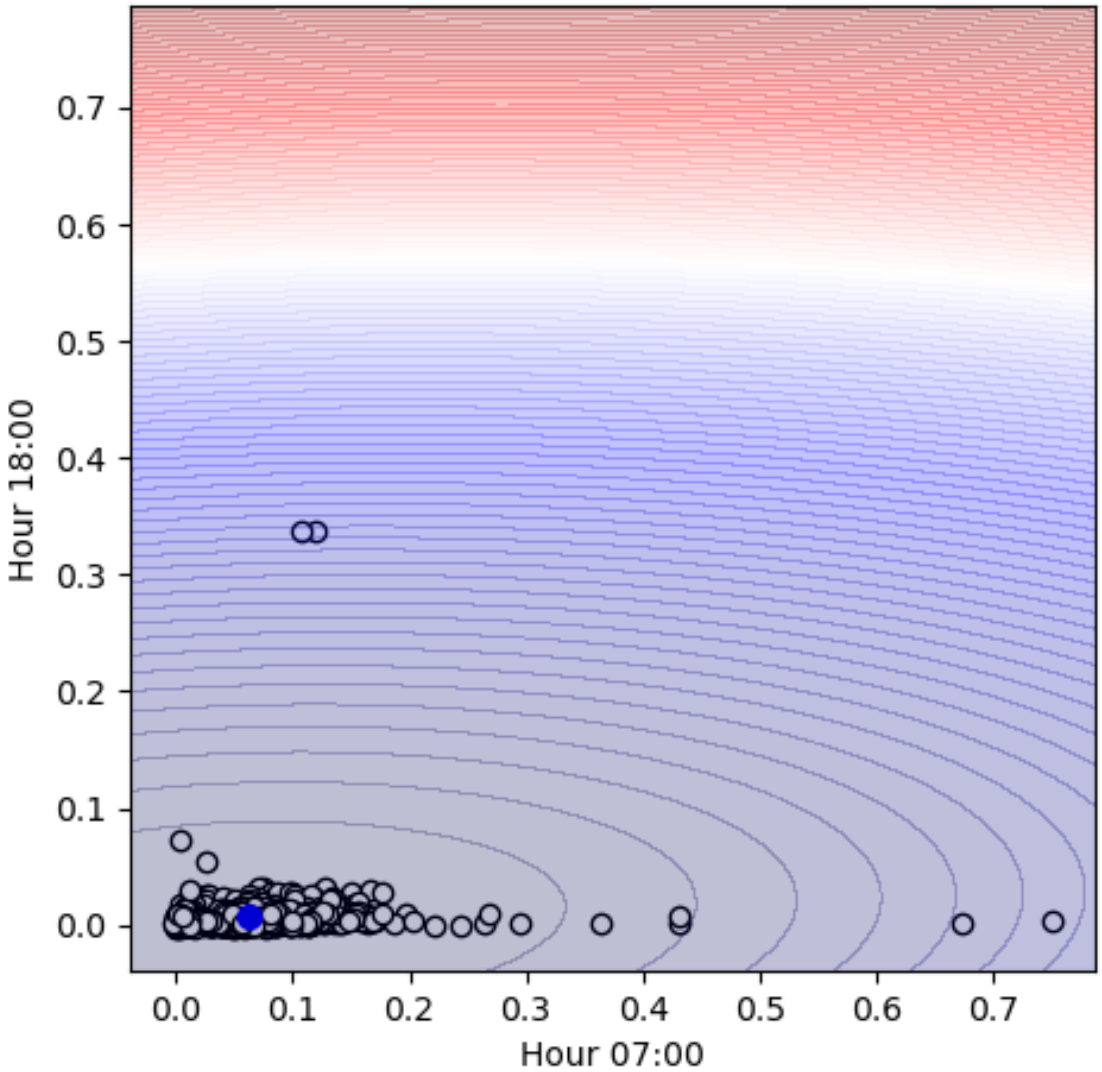
Odległość od średniej a rozkład danych



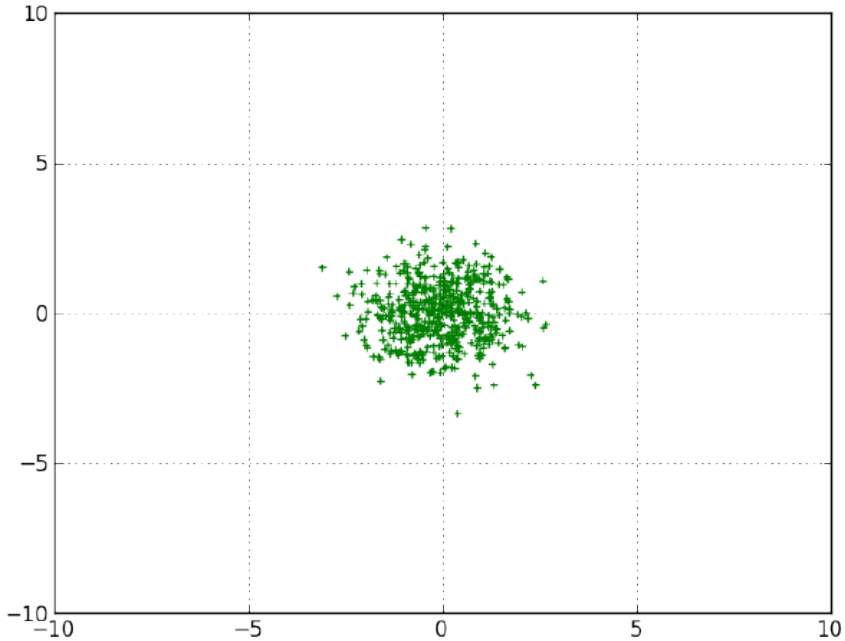
Odległość od średniej a rozkład danych



Odległość od średniej a rozkład danych



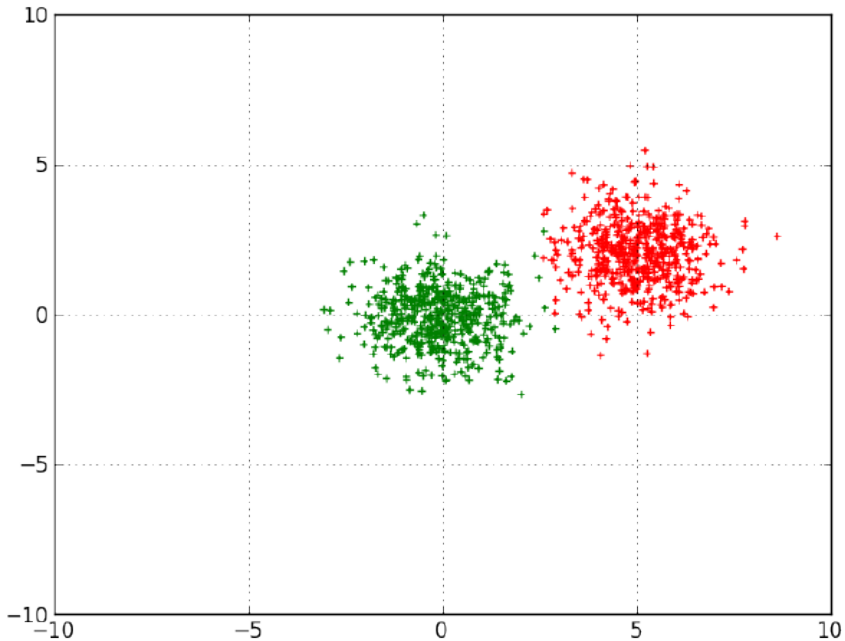
Opis rozkładu – macierz kowariancji



Parametry rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

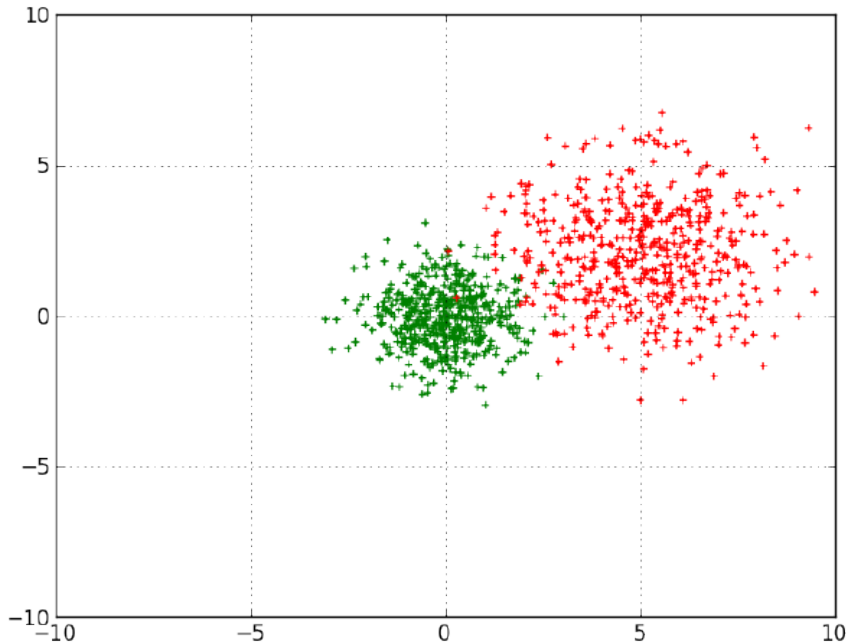
Opis rozkładu – macierz kowariancji



Parametry rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

$$\mu = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

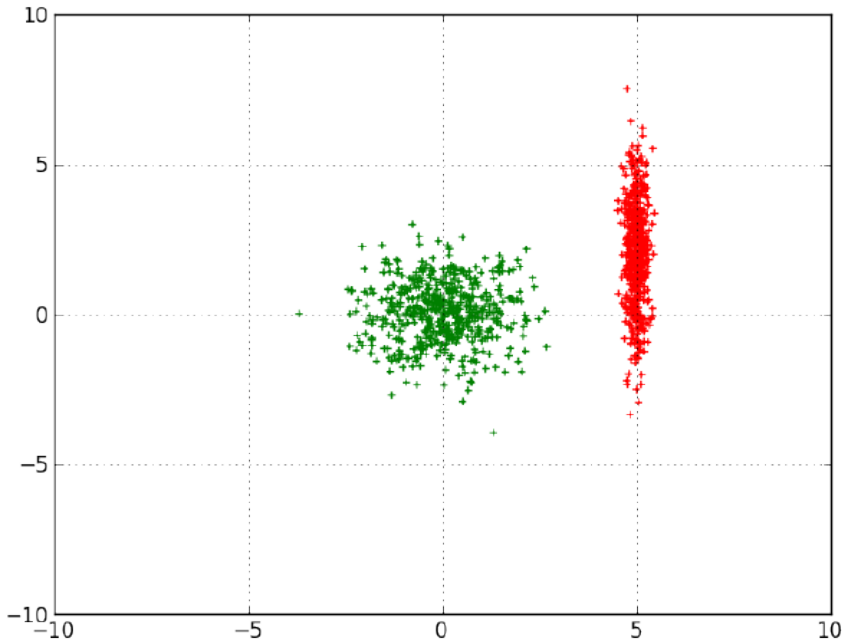
Opis rozkładu – macierz kowariancji



Parametry rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

$$\mu = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

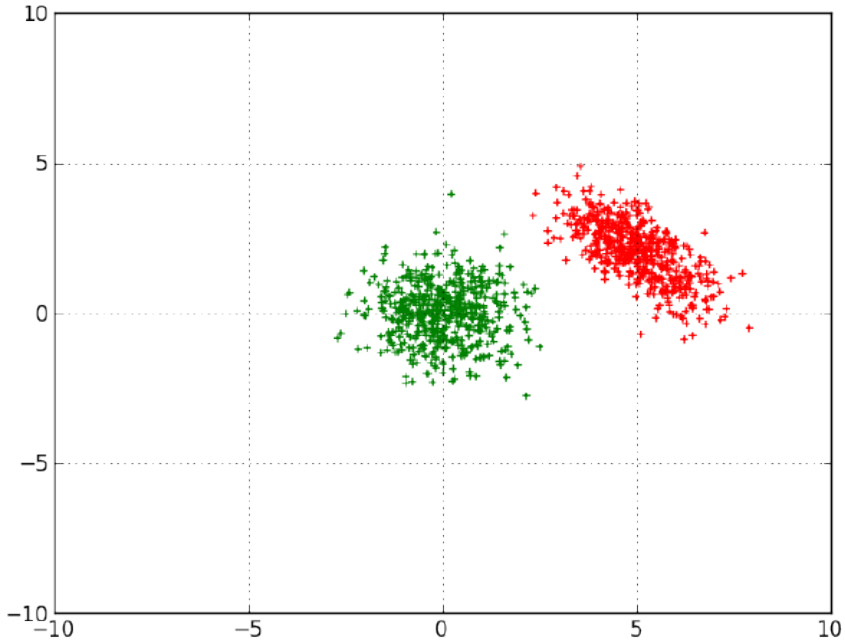
Opis rozkładu – macierz kowariancji



Parametry rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

$$\mu = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0.03 & 0 \\ 0 & 3 \end{bmatrix}$$

Opis rozkładu – macierz kowariancji

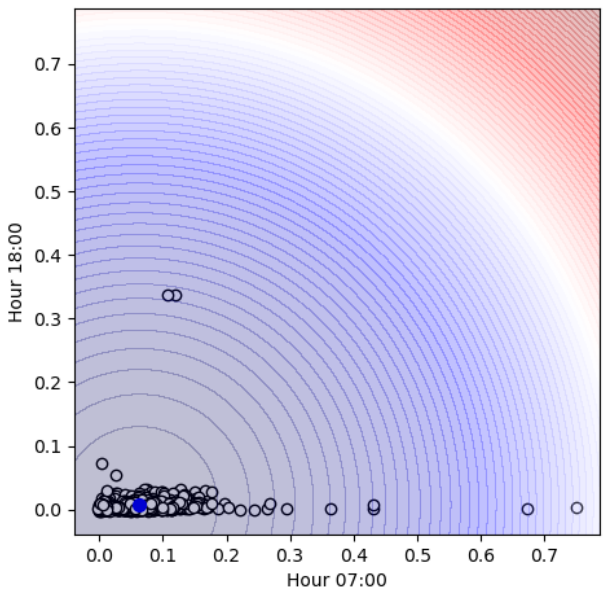


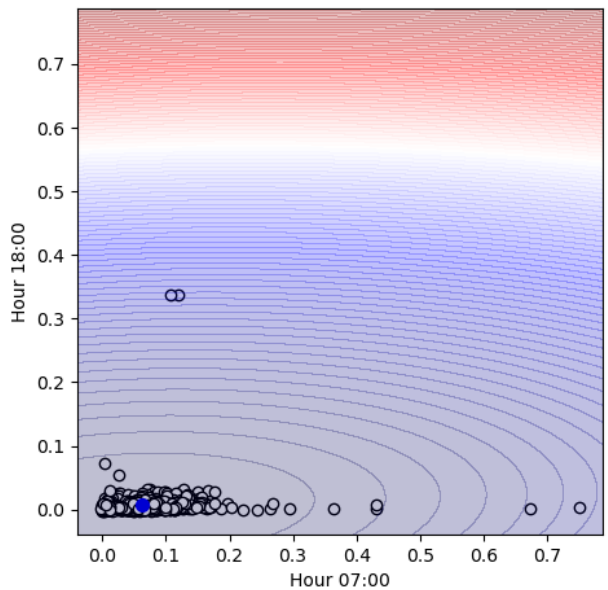
Parametry rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

$$\mu = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.69 \\ -0.69 & 1 \end{bmatrix}$$

Odległość Mahalanobisa

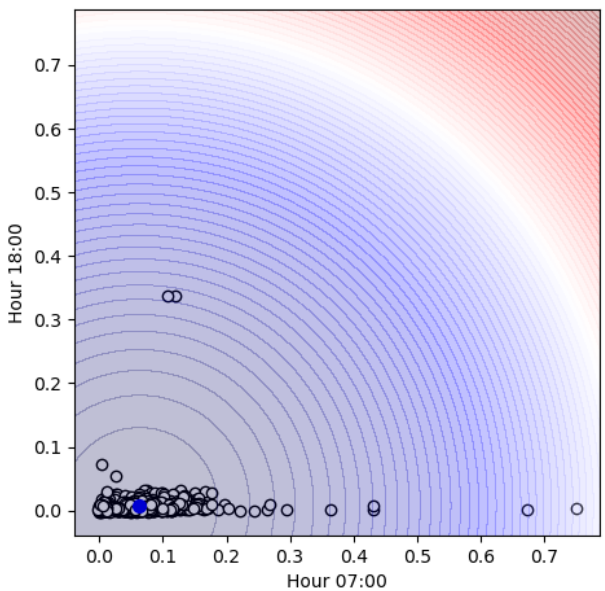
$$d_E(\bar{\mathbf{x}}, \mathbf{x}_i) = \|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24} (\bar{x}_j - x_{ij})^2} =$$

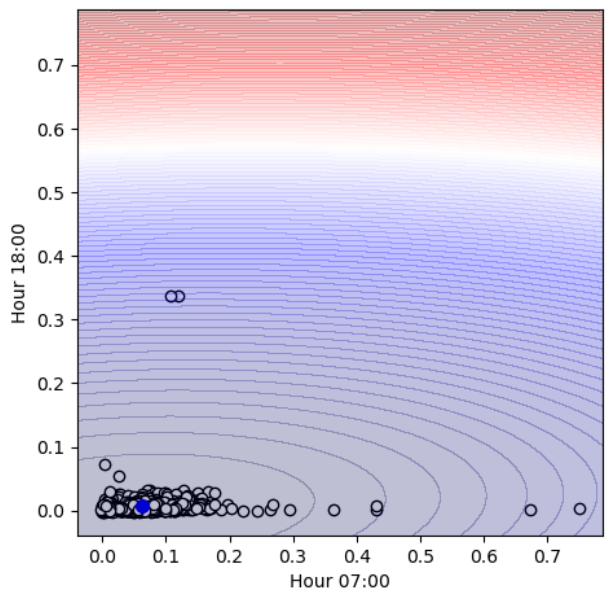




Odległość Mahalanobisa

$$d_E(\bar{\mathbf{x}}, \mathbf{x}_i) = \|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24} (\bar{x}_j - x_{ij})^2} =$$
$$\sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)(\bar{\mathbf{x}} - \mathbf{x}_i)^\top} =$$

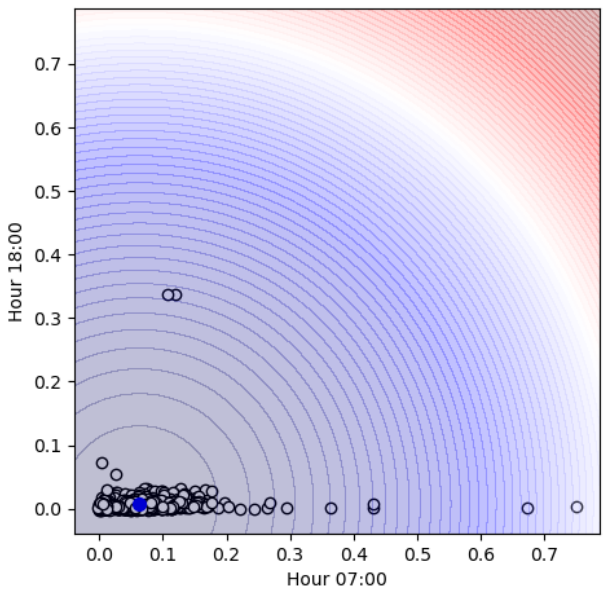


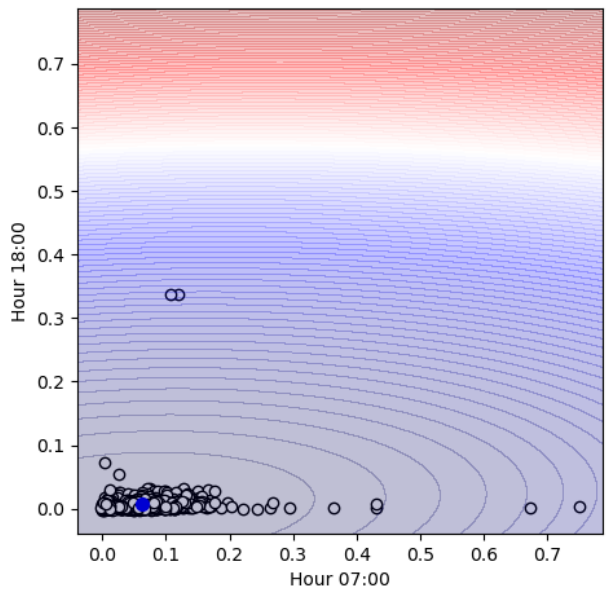


Odległość Mahalanobisa

$$d_E(\bar{\mathbf{x}}, \mathbf{x}_i) = \|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24} (\bar{x}_j - x_{ij})^2} =$$
$$\sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)(\bar{\mathbf{x}} - \mathbf{x}_i)^\top} = \sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i) I (\bar{\mathbf{x}} - \mathbf{x}_i)^\top}$$

I – macierz identyczności



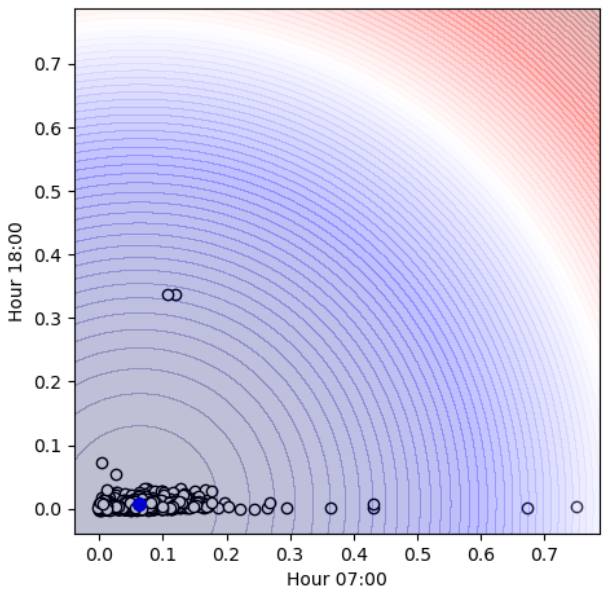


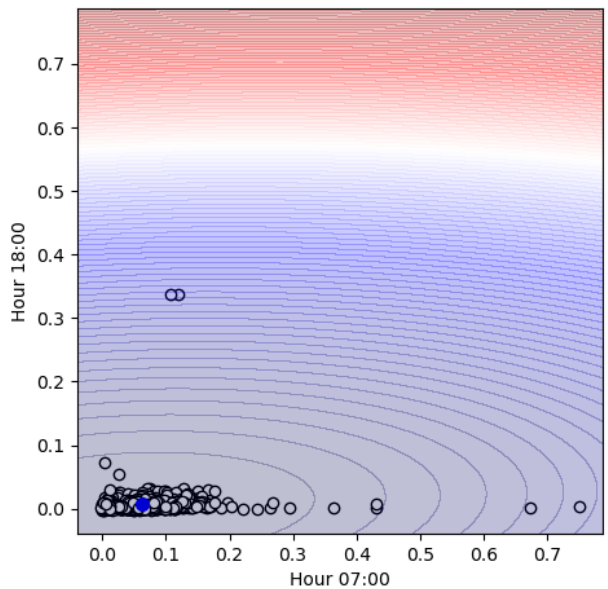
Odległość Mahalanobisa

$$d_E(\bar{\mathbf{x}}, \mathbf{x}_i) = \|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24} (\bar{x}_j - x_{ij})^2} =$$
$$\sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)(\bar{\mathbf{x}} - \mathbf{x}_i)^\top} = \sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i) I (\bar{\mathbf{x}} - \mathbf{x}_i)^\top}$$

I – macierz identyczności

$C = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}})$ – macierz kowariancji danych (ang. sample covariance)





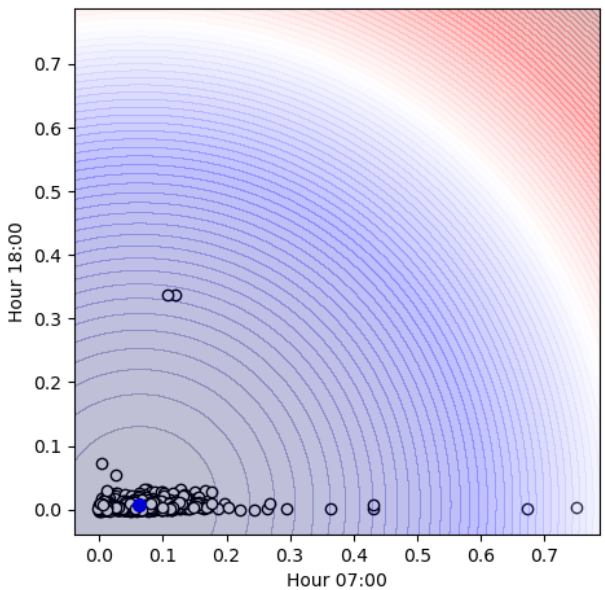
Odległość Mahalanobisa

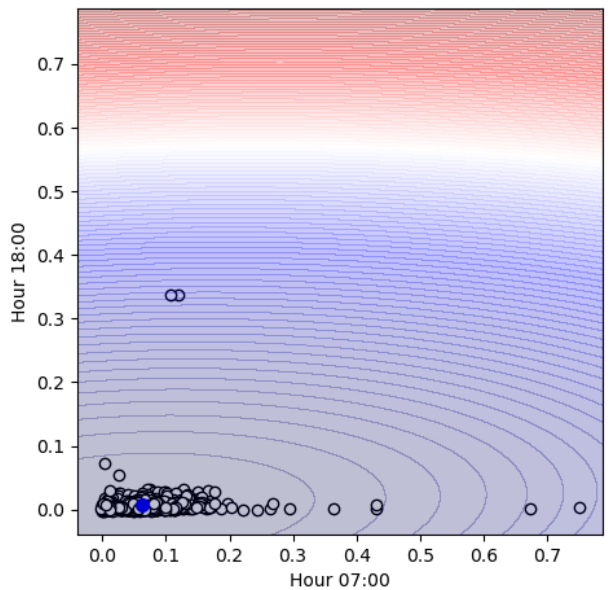
$$d_E(\bar{\mathbf{x}}, \mathbf{x}_i) = \|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24} (\bar{x}_j - x_{ij})^2} =$$
$$\sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)(\bar{\mathbf{x}} - \mathbf{x}_i)^\top} = \sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i) I (\bar{\mathbf{x}} - \mathbf{x}_i)^\top}$$

I – macierz identyczności

$C = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}})$ – macierz kowariancji danych (ang. sample covariance)

C^{-1} – odwrotność macierzy kowariancji





Odległość Mahalanobisa

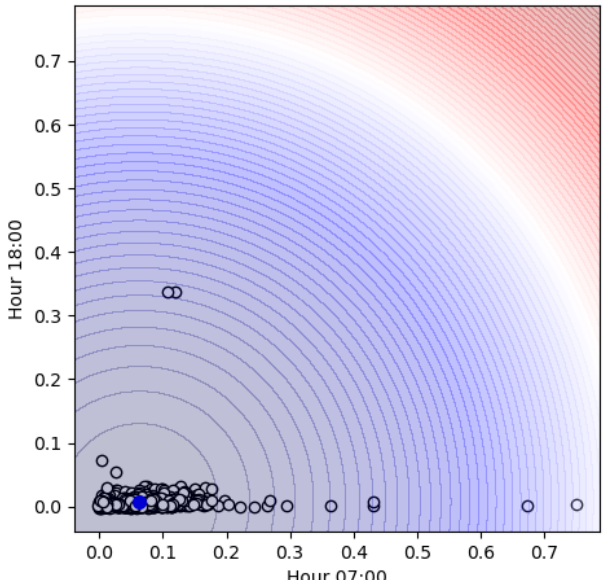
$$d_E(\bar{\mathbf{x}}, \mathbf{x}_i) = \|\bar{\mathbf{x}} - \mathbf{x}_i\| = \sqrt{\sum_{j=1}^{24}(\bar{x}_j - x_{ij})^2} =$$
$$\sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)(\bar{\mathbf{x}} - \mathbf{x}_i)^\top} = \sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)I(\bar{\mathbf{x}} - \mathbf{x}_i)^\top}$$

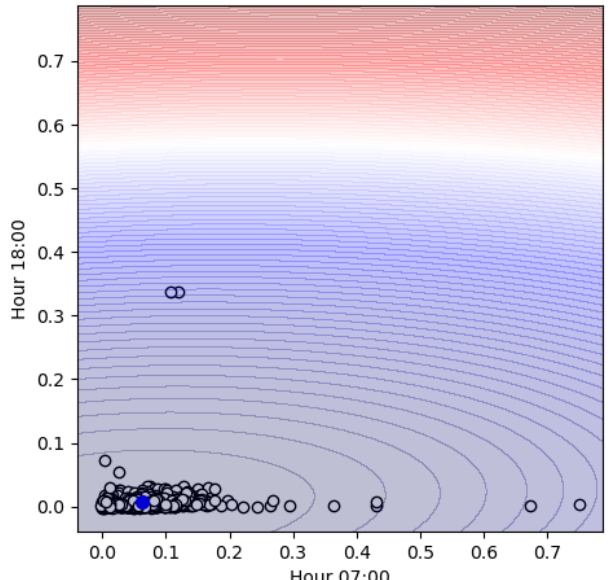
I – macierz identyczności

$C = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top(\mathbf{x}_i - \bar{\mathbf{x}})$ – macierz kowariancji danych (ang. sample covariance)

C^{-1} – odwrotność macierzy kowariancji

$$d_M(\bar{\mathbf{x}}, \mathbf{x}_i) = \sqrt{(\bar{\mathbf{x}} - \mathbf{x}_i)C^{-1}(\bar{\mathbf{x}} - \mathbf{x}_i)^\top}$$





Wykrywanie anomalii – odległość Mahalanobisa

1. Wykorzystujemy właściwości sygnału – okresowość dobową

```
[[4.43 4.77 6.1  ... 7.97 4.85 4.29]
[4.22 4.18 5.18 ... 9.1  6.37 5.06]
[4.86 5.9  6.54 ... 8.92 6.54 5.19]
...
[2.72 2.52 2.66 ... 9.17 6.59 4.8 ]
[3.51 2.9  2.52 ... 8.79 5.34 3.59]
[2.62 2.72 2.57 ... 9.71 6.01 3.41]]
```

   $X = \left[x_{ij}\right]$ i dni j godziny, $\mathbf{x}_i$ i-ta doba

2. Anomalia $\longrightarrow$ „nietypowe zachowanie"
3. „Zachowanie" $\longrightarrow$ średnia dobowa $\bar{\mathbf{x}} = \left[\bar{x}_j\right]$ $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ i macierz kowariancji
4. „Nietypowe" $\longrightarrow$ duża odległość od średniej liczona odległością Mahalanobisa $d_M(\bar{\mathbf{x}}, \mathbf{x}_i)$

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$ . . . ale jak wyliczyć T?

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$ . . . ale jak wyliczyć T?
   . . . dla różnych wartości średniej i m. kowariancji różne T :(

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$ . . . ale jak wyliczyć T? . . . dla różnych wartości średniej i m. kowariancji różne T :(
3. Chcielibyśmy wyrazić próg w postaci prawdopodobieństwa, np. anomalia jest jeżeli $\mathbf{x}_i$ ma prawdopodobieństwo np. $\alpha = 1\%$, albo 0.1%

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$ . . . ale jak wyliczyć T? . . . dla różnych wartości średniej i m. kowariancji różne T :(
3. Chcielibyśmy wyrazić próg w postaci prawdopodobieństwa, np. anomalia jest jeżeli $\mathbf{x}_i$ ma prawdopodobieństwo np. $\alpha = 1\%$, albo 0.1% . . . a dla α wyznaczyć T automatycznie

Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

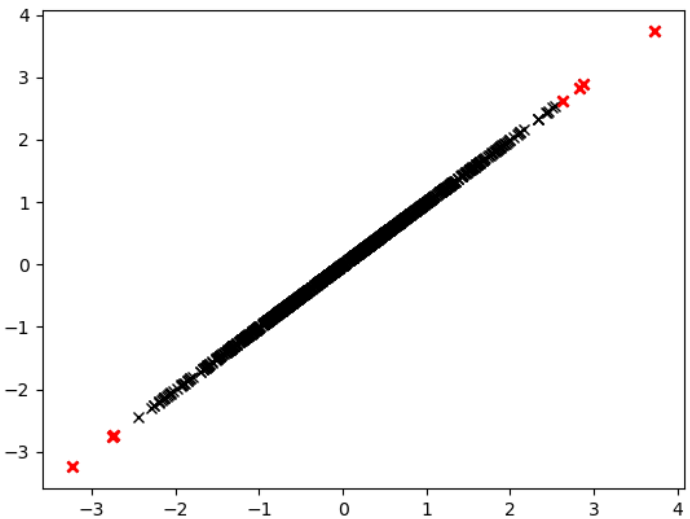
1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$ . . . ale jak wyliczyć T? . . . dla różnych wartości średniej i m. kowariancji różne T :(
3. Chcielibyśmy wyrazić próg w postaci prawdopodobieństwa, np. anomalia jest jeżeli $\mathbf{x}_i$ ma prawdopodobieństwo np. $\alpha = 1\%$, albo 0.1% . . . a dla α wyznaczyć T automatycznie
4. Wiemy, że jeżeli punkty w X mają rozkład normalny, to $d_M(\bar{\mathbf{x}}, \mathbf{x}_i)$ ma rozkład χ^2 z d-stopniami swobody (w naszym wypadku $d = 24$)

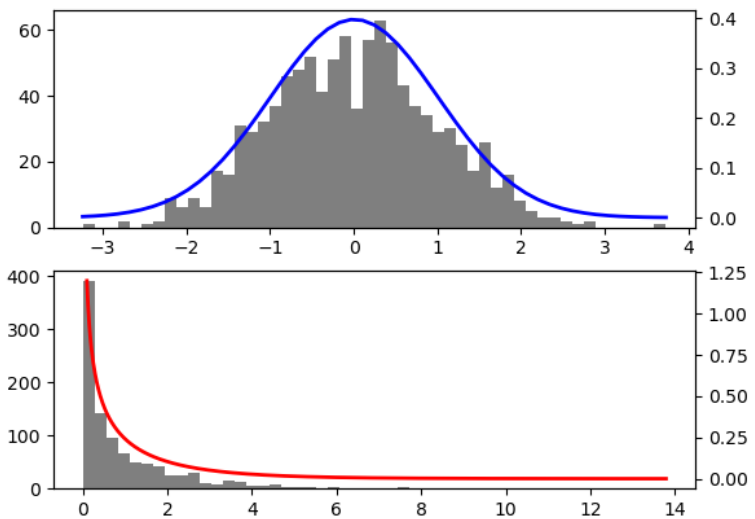
Wykrywanie anomalii – odległość Mahalanobisa $\longrightarrow$ algorytm RX

1. Mamy dane $X \longrightarrow$ średnia $\bar{\mathbf{x}}$, macierz kowariancji C
2. $\mathbf{x}_i$ jest anomalią jeżeli $d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T$ . . . ale jak wyliczyć T? . . . dla różnych wartości średniej i m. kowariancji różne T :(
3. Chcielibyśmy wyrazić próg w postaci prawdopodobieństwa, np. anomalia jest jeżeli $\mathbf{x}_i$ ma prawdopodobieństwo np. $\alpha = 1\%$, albo 0.1% . . . a dla α wyznaczyć T automatycznie
4. Wiemy, że jeżeli punkty w X mają rozkład normalny, to $d_M(\bar{\mathbf{x}}, \mathbf{x}_i)$ ma rozkład χ^2 z d-stopniami swobody (w naszym wypadku $d = 24$)
5. Znając rozkład – w tym wypadku χ^2 – możemy wyznaczyć T, dla którego prawdopodobieństwo przekroczenia wartości jest α, $P(d_M(\bar{\mathbf{x}}, \mathbf{x}_i) > T) = \alpha$

Algorytm RX – przykłady

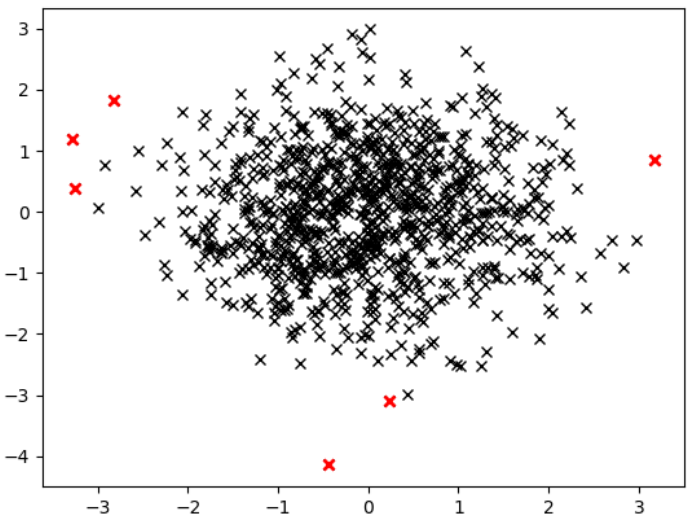
Dane sztucznie wygenerowane, rozkład normalny, $d = 1$

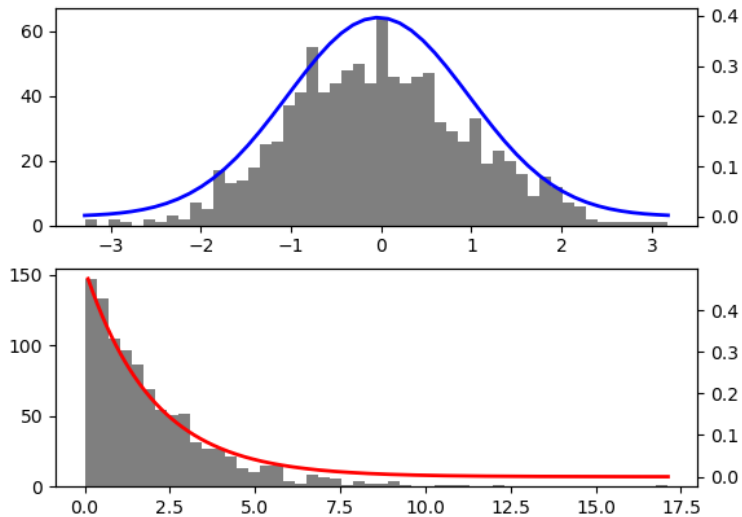




Algorytm RX – przykłady

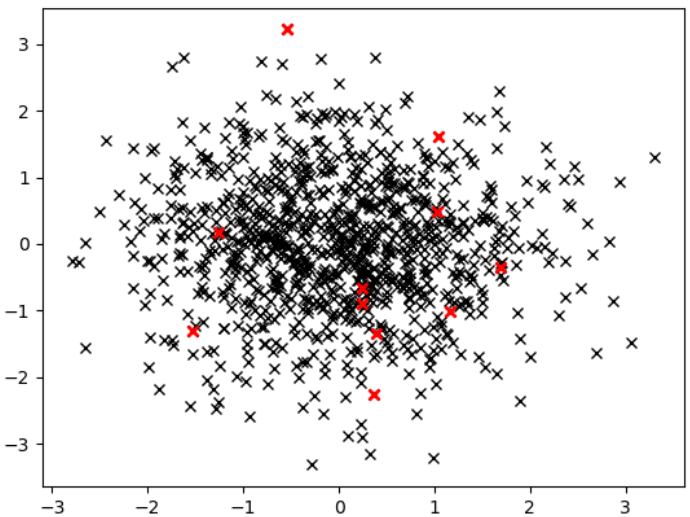
Dane sztucznie wygenerowane, rozkład normalny, $d = 2$

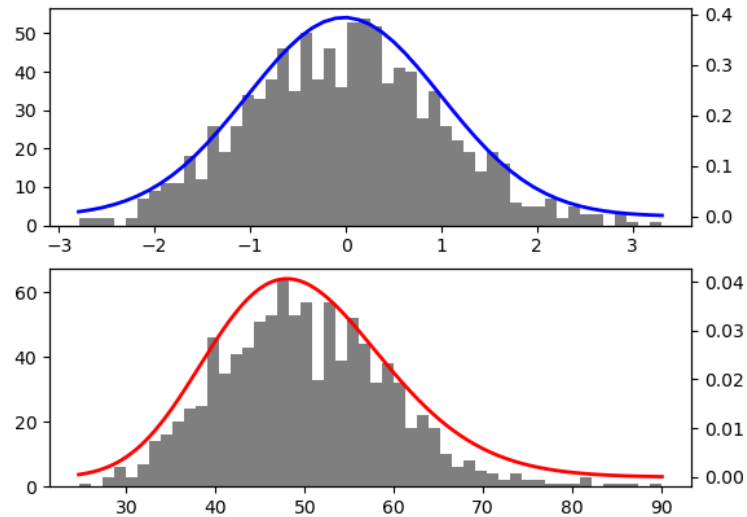




Algorytm RX – przykłady

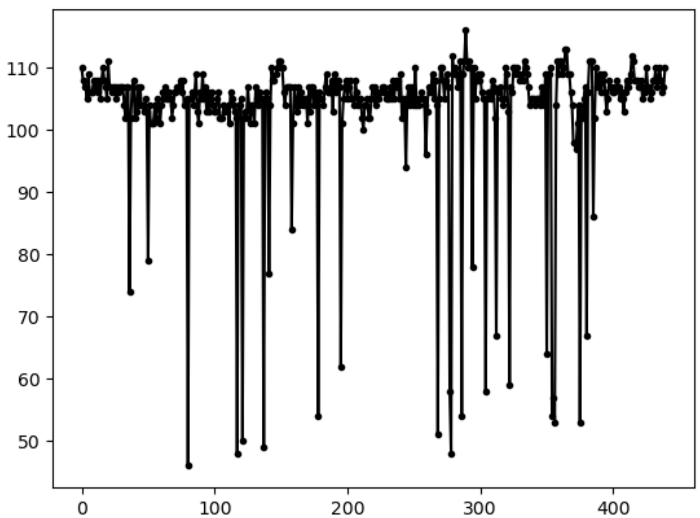
Dane sztucznie wygenerowane, rozkład normalny, $d = 50$

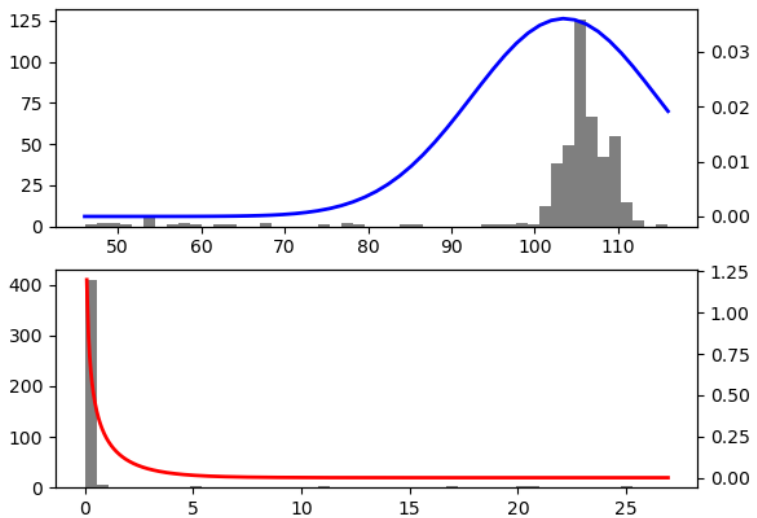




Algorytm RX – przykłady

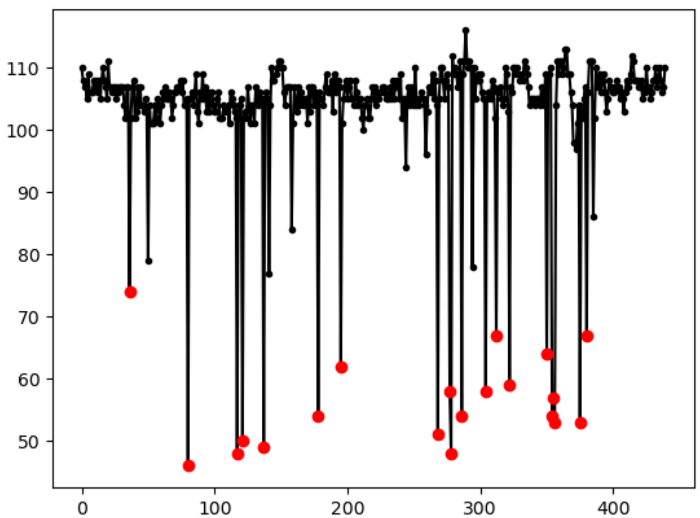
Dane rzeczywiste (monitoring przemysłowy), rozkład – ?, $d = 1$

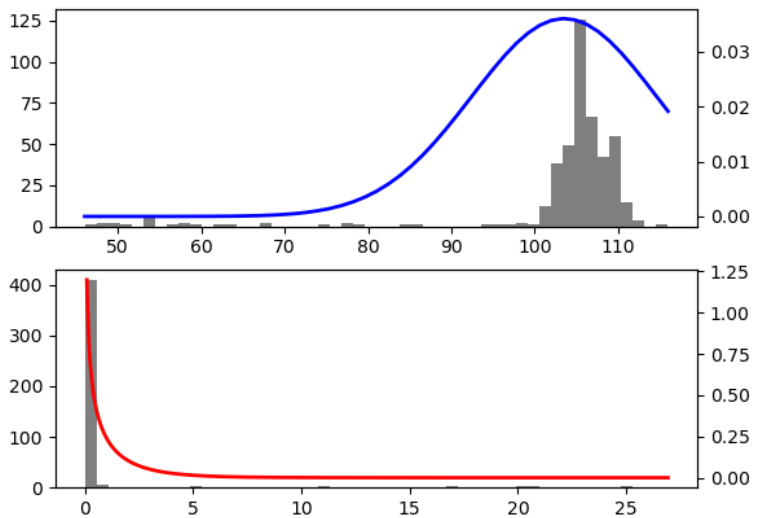




Algorytm RX – przykłady

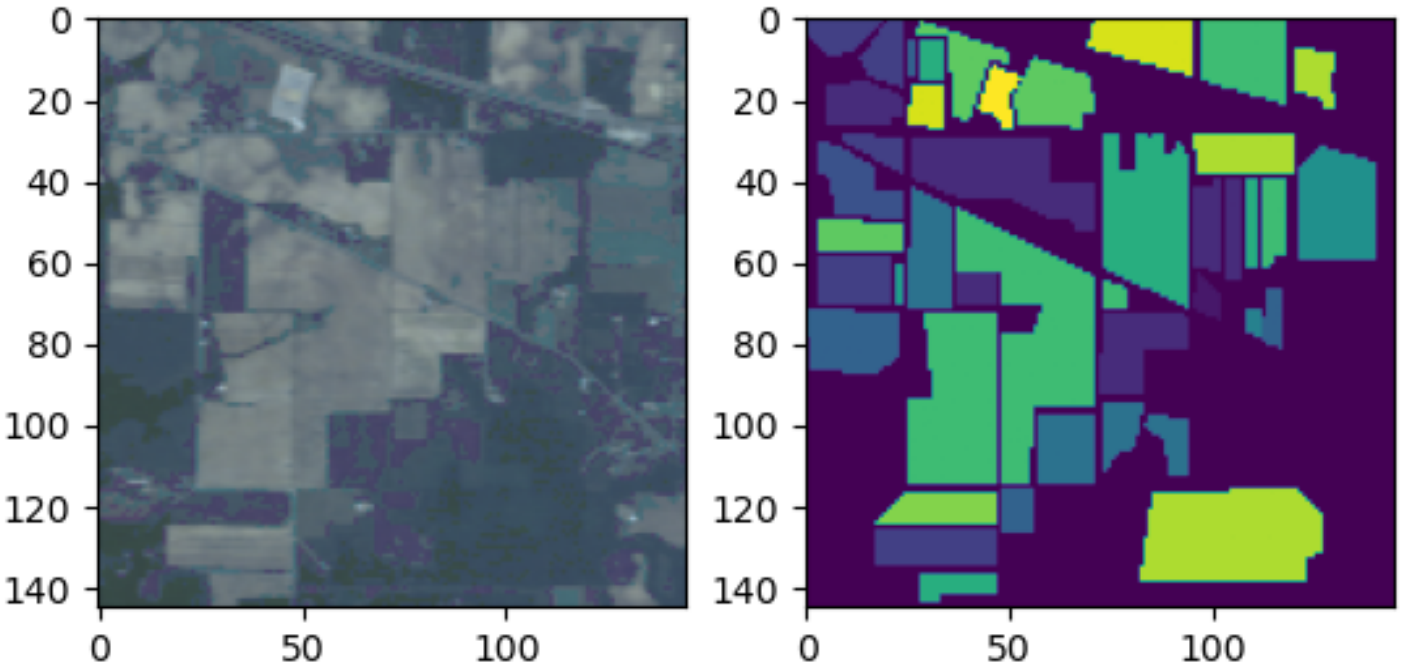
Dane rzeczywiste (monitoring przemysłowy), rozkład – ?, $d = 1$





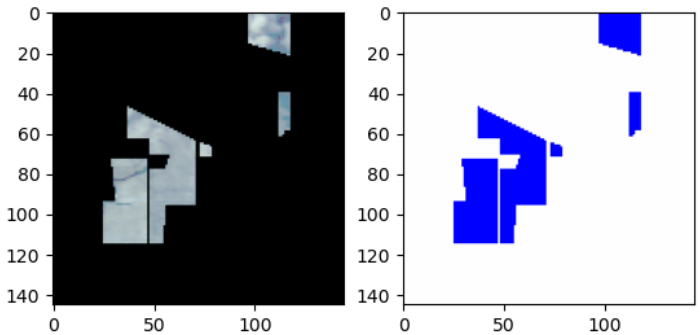
Algorytm RX – przykłady

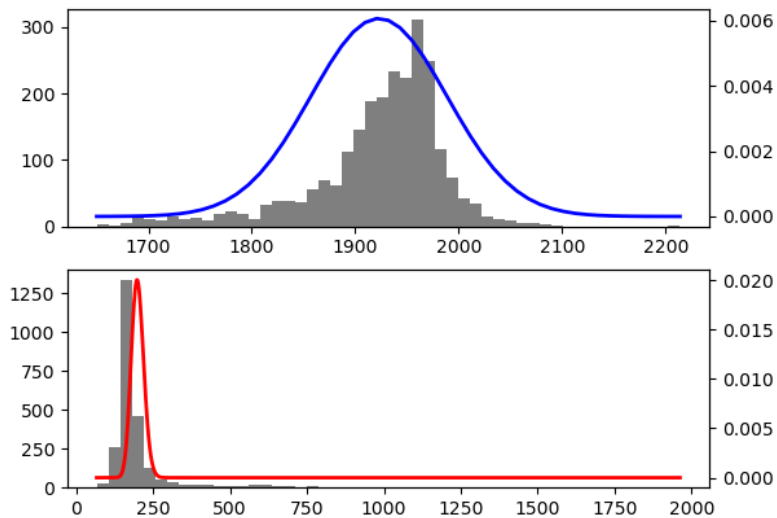
Dane rzeczywiste (zdjęcie hiperspektralne), rozkład – ?, $d = 100$



Algorytm RX – przykłady

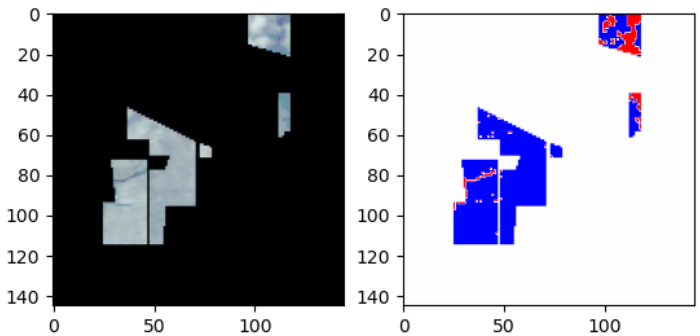
Dane rzeczywiste (zdjęcie hiperspektralne), rozkład – ?, $d = 100$

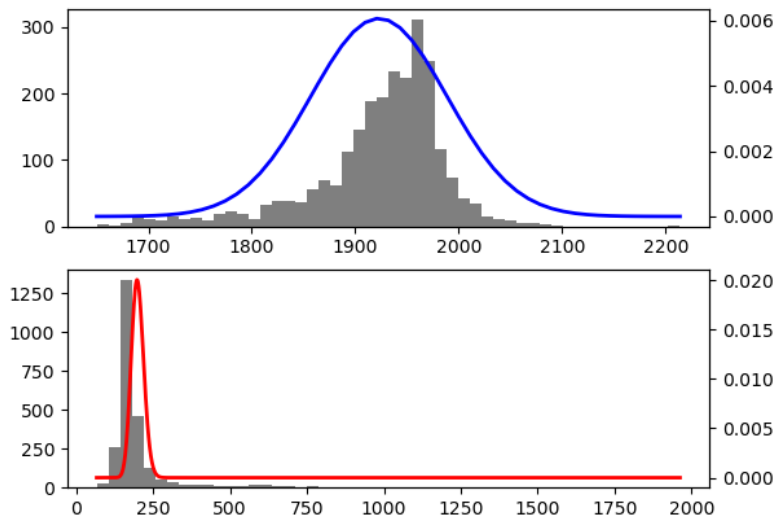




Algorytm RX – przykłady

Dane rzeczywiste (zdjęcie hiperspektralne), rozkład – ?, $d = 100$





Wykrywanie anomalii, algorytm RX, podsumowanie

1. Anomalie a outliery

Wykrywanie anomalii, algorytm RX, podsumowanie

1. Anomalie a outliery
2. Elementy (RX):
   - 2.1 Model – rozkład normalny, opisany średnią i macierzą kowariancji (parametry)
   - 2.2 Algorytm – odległość Mahalanobisa, wykorzystanie znanego rozkładu χ^2 (parametry wejściowe, wyjście)
   - 2.3 Teoria – prawdopodobieństwo i statystyka

Wykrywanie anomalii, algorytm RX, podsumowanie

1. Anomalie a outliery
2. Elementy (RX):
   - 2.1 Model – rozkład normalny, opisany średnią i macierzą kowariancji (parametry)
   - 2.2 Algorytm – odległość Mahalanobisa, wykorzystanie znanego rozkładu χ^2 (parametry wejściowe, wyjście)
   - 2.3 Teoria – prawdopodobieństwo i statystyka
3. Rozwinięcia
   - 3.1 Mixture of Gaussians (MoG)
   - 3.2 Principal Component Analysis (PCA)
   - 3.3 Hidden Markov Models (HMM)

Wykrywanie anomalii, algorytm RX, podsumowanie

1. Anomalie a outliery
2. Elementy (RX):
   - 2.1 Model – rozkład normalny, opisany średnią i macierzą kowariancji (parametry)
   - 2.2 Algorytm – odległość Mahalanobisa, wykorzystanie znanego rozkładu χ^2 (parametry wejściowe, wyjście)
   - 2.3 Teoria – prawdopodobieństwo i statystyka
3. Rozwinięcia
   - 3.1 Mixture of Gaussians (MoG)
   - 3.2 Principal Component Analysis (PCA)
   - 3.3 Hidden Markov Models (HMM)
4. Alternatywy – np. sieci neuronowe