Wybrane metody identyfikacji obserwacji odstających w szeregach czasowych

Piotr Migdałek

Politechnika Wrocławska

11.07.2022

Postać szeregu czasowego z interwencją

Szereg czasowy Y_t^*, który jest poddany działaniu niepowtarzającego się efektu zewnętrznego (w chwili T) można zapisać jako:

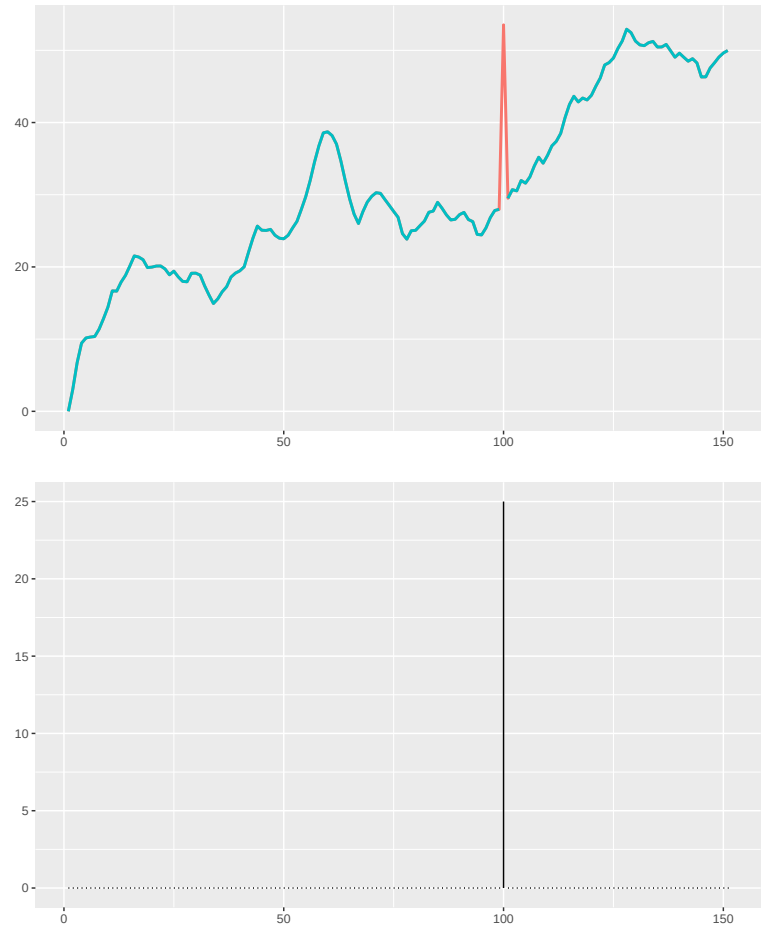
$$Y_t^* = Y_t + \omega \frac{A(B)}{G(B)H(B)} P_t^{(T)},$$

gdzie Y_t jest w ogólności procesem SARIMA. $P_t^{(T)} = 1$, gdy $t = T$ oraz 0, gdy $t \neq T$. Parametr ω odpowiada za siłę efektu obserwacji odstającej, natomiast wyrażenie $A(B)/\{G(B)H(B)\}$ modeluje jego dynamikę (B to operator przesunięcia wstecz).

Addytywna obserwacja odstająca

Addytywną obserwacją odstającą (ang. *additive outlier*) nazwiemy obserwację, która wpływa na szereg czasowy jedynie w chwili $t = T$. Jej dynamikę możemy przedstawić następująco:

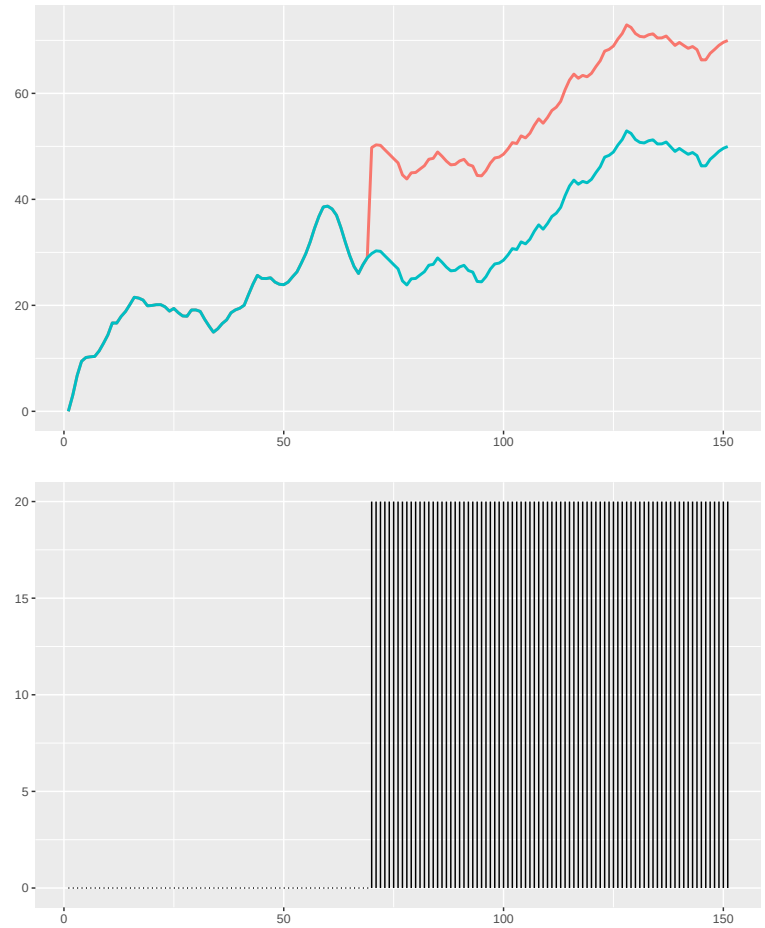
$$\frac{A(B)}{G(B)H(B)} = 1.$$



Zmiana poziomu

Zmianą poziomu (ang. *level shift*) nazwiemy obserwację odstającą, która generuje nagłe oraz trwałe przesunięcie poziomu szeregu. W tym przypadku, dynamika efektu obserwacji odstającej może być przedstawiona jako:

$$\frac{A(B)}{G(B)H(B)} = \frac{1}{1-B}.$$



Tymczasowa zmiana

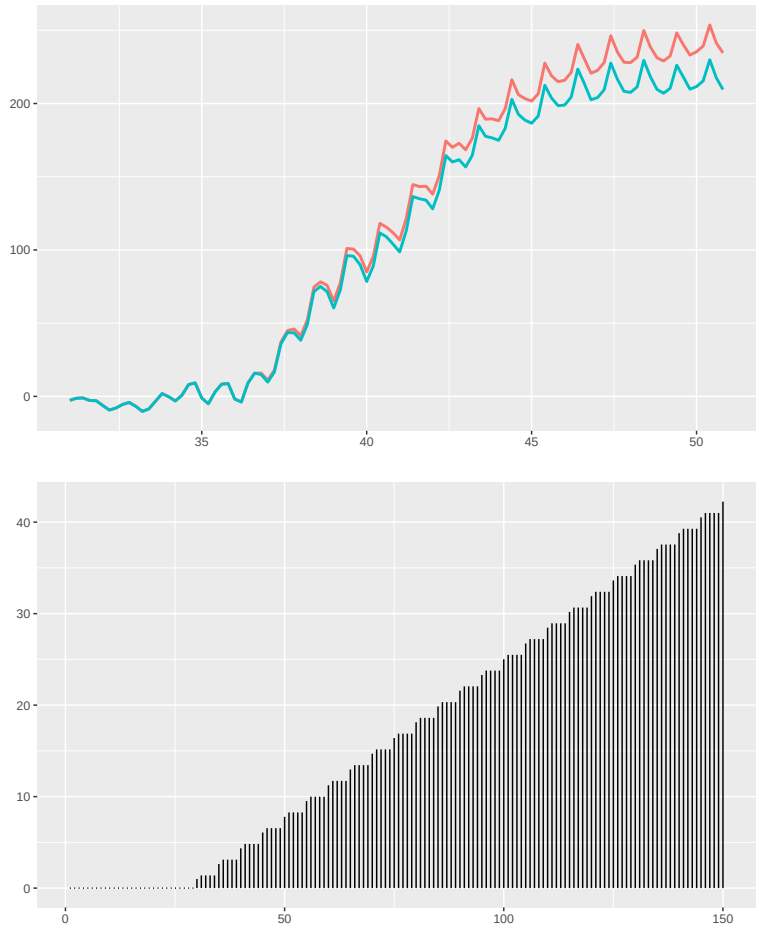
Tymczasową zmianą (ang. *temporary change*) nazwiemy obserwację odstającą, która generuje nagłe przesunięcie poziomu szeregu, które stopniowo wygasa. Jej dynamikę możemy przedstawić następująco:

$$\frac{A(B)}{G(B)H(B)} = \frac{1}{1-\delta B}, \ \delta \in [0,1].$$

Innowacyjna obserwacja odstającą

Innowacyjną obserwację odstającą (ang. *innovative outlier*) nazwiemy obserwację, która ma nagły i trwały wpływ na postać szeregu, a sama jej dynamika zależy od modelu wybranego dla Y_t. Jej dynamikę możemy przedstawić jako:

$$\frac{A(B)}{G(B)H(B)} = \frac{\theta(B)\Theta(B^s)}{\nabla^d \nabla_s^D \phi(B)\Phi(B^s)}.$$



Model ARIMA w przypadku obecności wielu obserwacji odstających

Ogólny model dla szeregu Y_t^* poddanego wpływowi m obserwacji odstających pojawiających się w chwilach T_1, $\ldots T_m$ jest postaci:

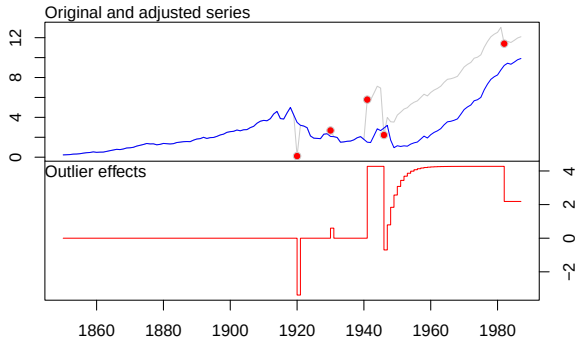
$$Y_t^* = \sum_{j=1}^{m} \omega_j \frac{A_j(B)}{G_j(B)H_j(B)} P_t^{(T_j)} + \frac{\theta(B)\Theta(B^s)}{\nabla^d \nabla_s^D \phi(B)\Phi(B^s)} Z_t, \qquad (1)$$

gdzie $Z_t \sim WN(\sigma^2)$.

Metoda estymacji parametrów modelu ARIMA w przypadku obecności wielu obserwacji odstających

- Na początku przeprowadzana jest wstępna estymacja parametrów modelu i na jej podstawie wyznaczane są pozycje T_j oraz efekty $A_j(B)/\{G_j(B)H_j(B)\}$ obserwacji odstających,
- Drugim etapem jest wspólna estymacja parametrów modelu oraz efektów generowanych przez obserwacje odstające $(\widehat{\omega}_j)$, wykorzystująca wyniki uzyskane w poprzednim kroku,
- W trzecim etapie procedury obserwacje odstające oraz ich efekty zostają ponownie estymowane na podstawie uaktualnionych estymatorów parametrów, które są w mniejszym stopniu obciążone efektami anomalii.

Studium przypadku: modelowanie ARIMA uwzględniające efekty obserwacji odstających



Prognozowanie emisji CO2 w Polsce

Głównym celem analizy było zastosowanie modeli (1) do skonstruowania prognoz oraz porównanie ich skuteczności na tle predykcji wyznaczonych wykorzystując modele ARIMA oraz ETS.

Dokładność prognoz emisji CO2 w Polsce wykorzystując modele ARIMAX, ARIMA oraz ETS

	Dane treningowe		Dane testowe	
	RMSE	MAPE	RMSE	MAPE
Bez transformacji potęgowej				
ARIMA(0, 1, 1)	0.744	36.503	3.184	34.999
ARIMA(1, 1, 3) + (AO + LS + TC)	**0.167**	**6.773**	3.832	41.476
ARIMA(0, 1, 1) + (AO + LS + TC)	**0.257**	**6.782**	**1.381**	**14.211**
ARIMA(1, 1, 2) + (AO + LS + TC + IO)	**0.341**	**8.868**	24.589	275.082
ARIMA(0, 1, 0) + (AO + LS)	0.527	11.201	**1.381**	**9.592**
metoda naiwna	0.766	35.351	3.225	35.463
ETS(M, N, N)	0.765	38.120	3.119	34.265
Uwzględniając transformację potęgową				
ARIMA(0, 2, 1)	0.782	34.773	4.943	53.636
ARIMA(0, 1, 5)	0.747	37.738	3.198	35.159
ARIMA(1, 1, 1) + (AO + LS + TC)	0.805	34.367	**1.739**	**18.823**
ARIMA(0, 1, 1) + AO	0.595	30.874	4.097	44.796
metoda naiwna	0.766	35.351	3.225	35.463
ETS(A, N, N)	0.743	35.165	3.222	35.421

Metody oparte na ruchomym oknie

Funkcje `OSWM` – (*one-sided window method*) oraz `TSWM` – (*two-sided window method*) klasyfikują obserwacje jako anomalie, gdy zachodzi poniższa nierówność:

$$|Y_t - \hat{Y}_t| > \tau.$$

Dla `OSWM`:

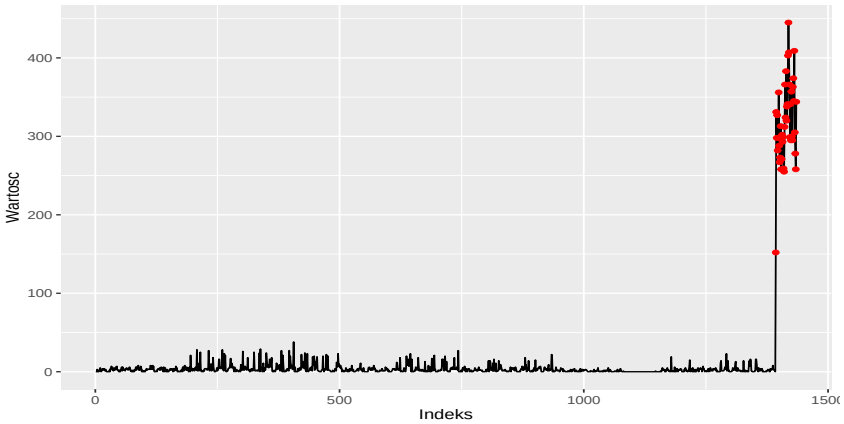
$$\widehat{Y}_t = p.func(\{Y_{t-m}, \ldots, Y_{t-1}\}).$$

Dla `TSWM`:

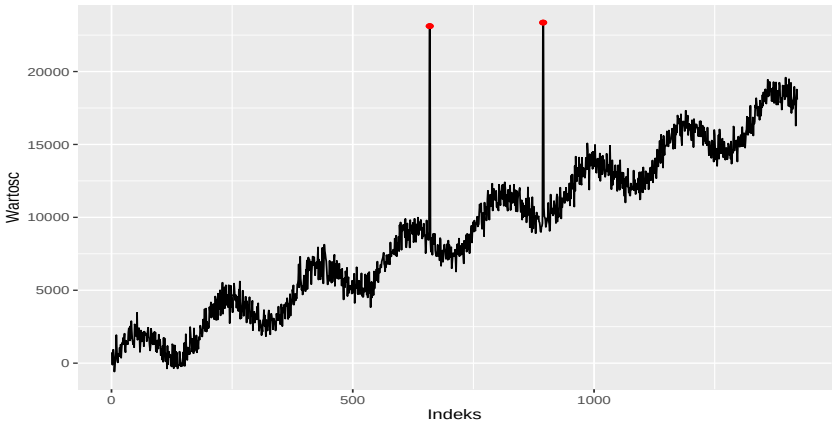
$$\widehat{Y}_t = e.func(\{Y_{t-k}, \ldots, Y_{t-1}, Y_{t+1}, \ldots Y_{t+k}\}).$$

Wartość progowa to $\tau = \alpha \cdot t.func(Y_t)$. Domyślnie funkcje $p.func$ oraz $e.func$ to średnie próbkowe, natomiast $t.func$ odchylenie standardowe próbkowe.

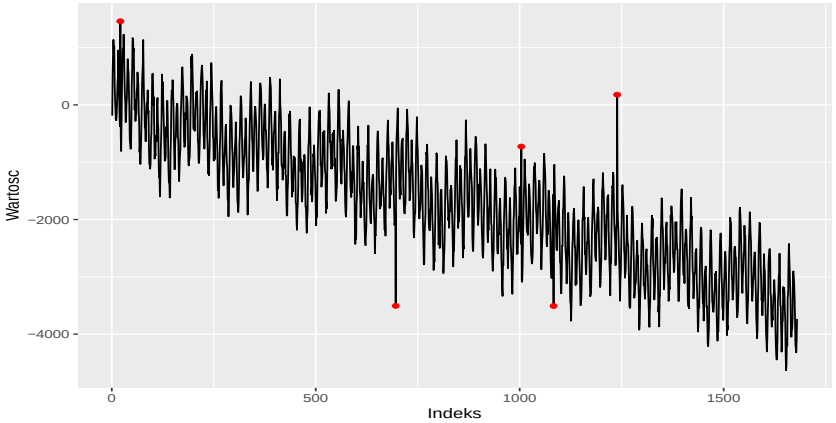
Dane Yahoo! użyte w analizie porównawczej



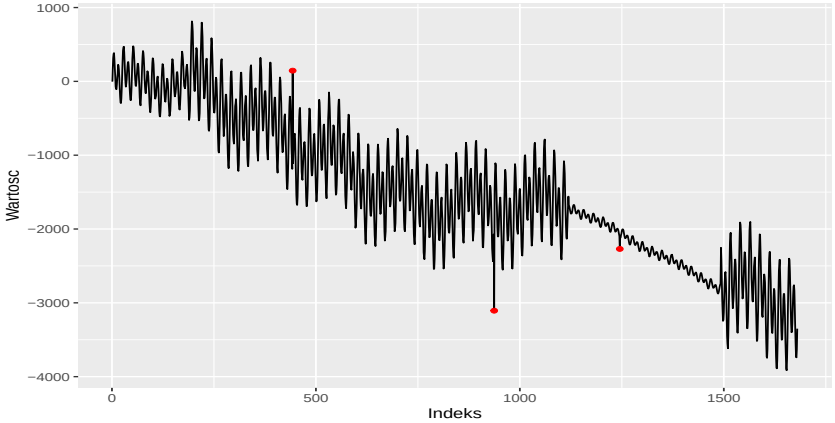
A1



A2



A3



A4

Wyniki analizy porównawczej metod detekcji punktowych obserwacji odstających

		A1		A2		A3		A4	
algorytm	implementacja	F1	czas [s]	F1	czas [s]	F1	czas [s]	F1	czas [s]
`OSWM`	implementacja własna	**0.558**	**0.016**	0.725	**0.013**	0.983	0.037	0.819	0.034
`TSWM`	implementacja własna	0.354	0.035	**0.929**	0.041	0.984	**0.03**	0.834	**0.03**
`tso`	`tsoutliers`	0.339	39.99	0.587	4.619	**0.995**	7.82	**0.865**	27.98
`locate.outliers`	`tsoutliers`	0.354	0.559	0.656	0.427	0.881	0.533	0.685	0.689
`tsoutliers`	`forecast`	0.368	0.057	0.796	0.051	0.962	0.05	0.596	0.066
`detect_outliers`	`tsrobprep`	0.439	1.457	0.683	1.816	0.795	1.744	0.726	1.9
`ad_vec`	`AnomalyDetection`	0.458	0.057	0.534	0.075	0.6874	0.067	0.566	0.076

Czym są anomalne sekwencje?

Nietrywialne dopasowanie

Sekwencja M o początku w chwili p będzie nietrywialnym dopasowaniem dla sekwencji C o początku w chwili q, gdy $|p - q| \geq m$ (przy czym zakładamy, że sekwencje są tej samej długości m).

Sekwencja anomalna

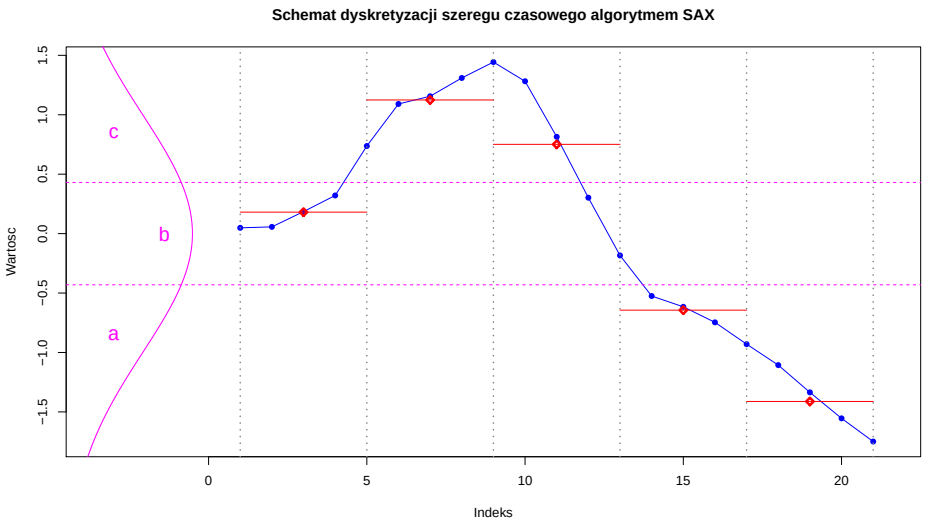
Sekwencja D, o długości m oraz początku w chwili l, szeregu czasowego Y_t jest nazywana anomalną (ang. *discord*), kiedy D ma największą odległość do najbliższego nietrywialnego dopasowania.

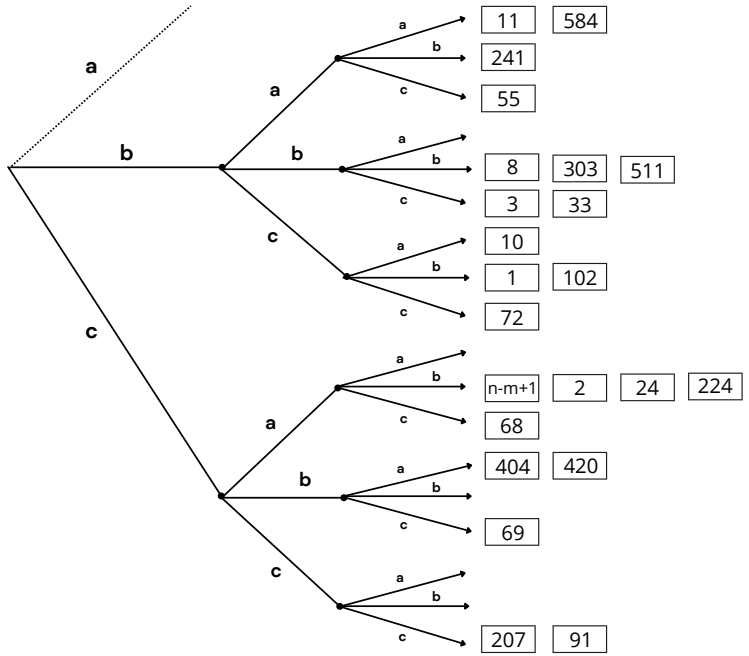
Jak ich szukać?

Algorytm 1 Brute force

best_so_far_dist $\leftarrow 0$ {największa aktualnie odległość do najbliższego sąsiada}
best_so_far_loc $\leftarrow NULL$ {odpowiadający jej początek sekwencji}
for $p = 1$ **to** $n - m + 1$ **do**
  $NN\_dist = \infty$
  for $q = 1$ **to** $n - m + 1$ **do**
    if $|p - q| \geq m$ **then**
      if $Dist(Y_p, \ldots, Y_{p+m-1}\ Y_q, \ldots, Y_{q+m-1}) < NN\_dist$ **then**
        $NN\_dist \leftarrow Dist(Y_p, \ldots, Y_{p+m-1}\ \ Y_q, \ldots, Y_{q+m-1})$
      end if
    end if
  end for
  if $NN\_dist > best\_so\_far\_dist$ **then**
    $best\_so\_far\_dist \leftarrow NN\_dist$
    $best\_so\_far\_loc \leftarrow p$
  end if
end for
return $[best\_so\_far\_loc,\ best\_so\_far\_dist]$

HOT-SAX

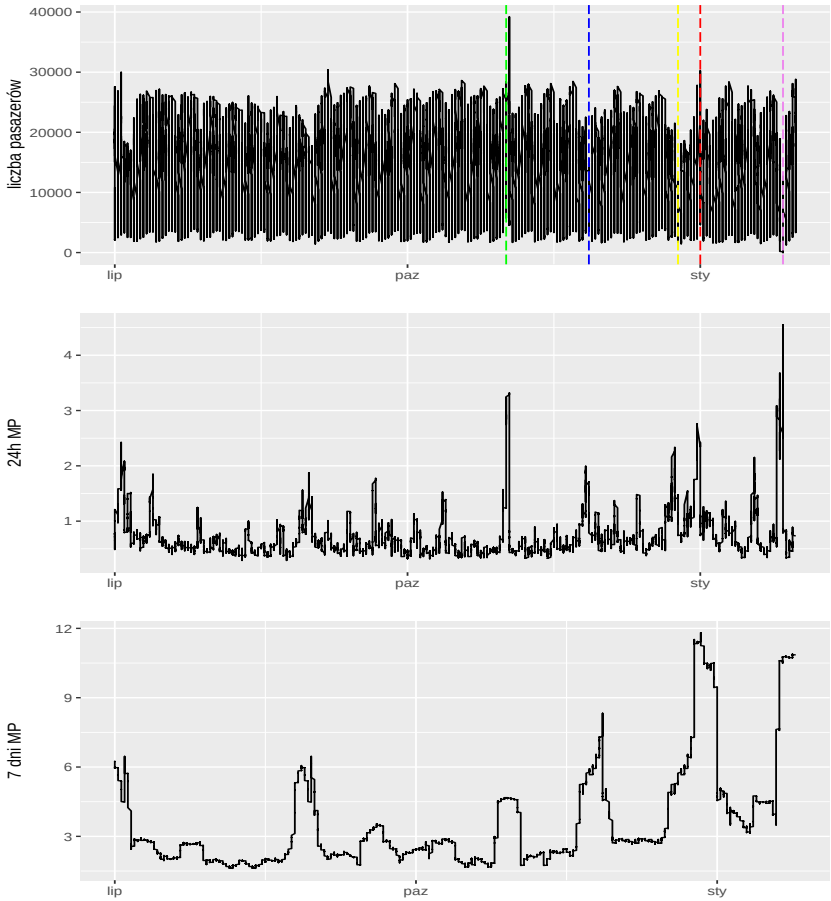

Schemat dyskretyzacji szeregu czasowego algorytmem SAX
Wartosc
Indeks
c
b
a
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
20


a
b
c
11
584
241
55
8
303
511
3
33
10
1
102
72
n-m+1
2
24
224
68
404
420
69
207
91

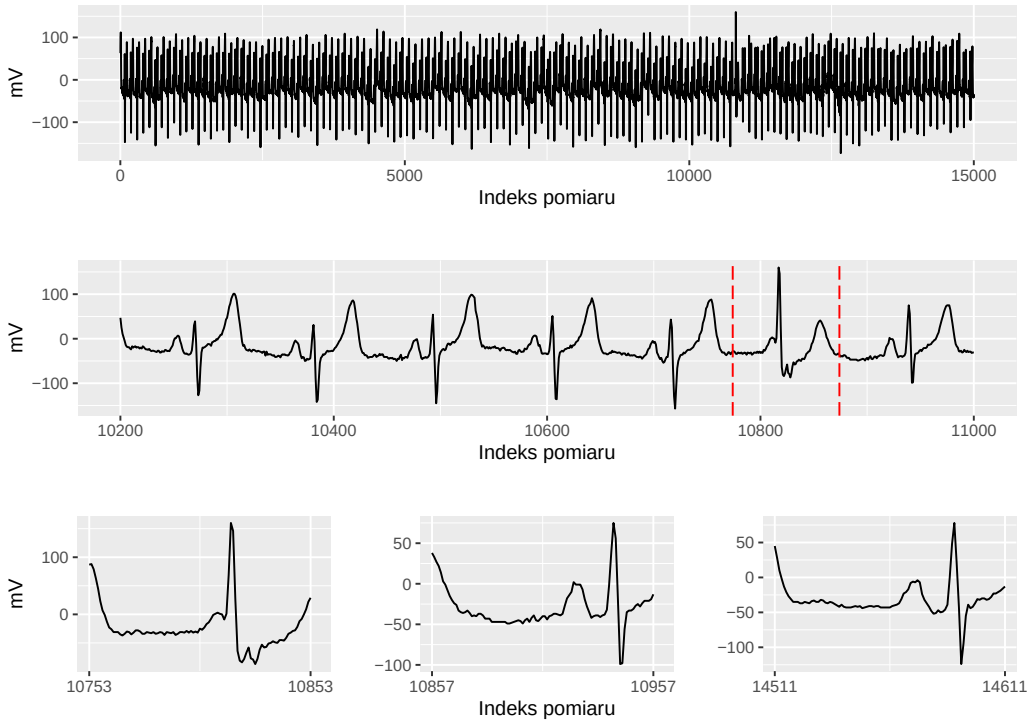
Profil macierzowy

Profil macierzowy

jest wektorem kolejnych odległości euklidesowych (standaryzowanych) sekwencji o długości m (czyli dla i-tego indeksu szeregu porównywana będzie sekwencja $(Y_i, \ldots, Y_{i+m-1})$ względem ich najbliższego nietrywialnego dopasowania.



Studium przypadku: detekcja anomalnych sekwencji w EKG



Detekcja rytmu przedsionkowego w EKG

Celem analizy było porównanie metod wykrywania anomalnych sekwencji bazując na danych EKG zawierających zmianę pracy serca wyznaczoną przez eksperta.