Wyzwania dla zastosowań odpornej funkcjonalnej analizy danych w ekonomii

Daniel Kosiorowski[1] Dominik Mielczarek[2] Jerzy P. Rydlewski[2] Ewa Szlachtowska[1,2]

[1]Cracow University of Economics, Department of Statistics
[2]AGH University of Science and Technology, Faculty of Applied Mathematics

Spotkanie w ramach cyklu wykładów otwartych **"Wielowymiarowe Metody Statystyczne"** organizowanych przez poznański odział PTS

13 Czerwiec 2018, Poznań

Plan referatu

1. Zastosowania funkcjonalnej analizy danych w ekonomii
   - Odporna analiza ekonomiczna
   - Odporność procedury statystycznej
2. Odporna funkcjonalna analiza danych
   - Statystyczne funkcje głębi
   - Metody wykrywania funkcjonalnych outlierów
3. Wyzwania dla zastosowań odpornej FDA w ekonomii
   - Odporna analiza funkcjonalnych szeregów czasowych (FTS)
   - Hierarchiczne funkcjonalne szeregi czasowe (HFTS)
   - Propozycje odpornej predykcji HFTS
   - Popularne algorytmy analizy skupisk (AS) w FDA
   - Algorytm k-średnich w FDA
   - Odporna AS w FDA
   - Odporna analiza dyskryminacyjna
4. Przykłady empiryczne
   - Funkcjonalne outliery w dobowym zanieczyszczenie powietrza w Katowicach
   - Wykrywanie nietypowych zachowań internautów
   - Odporna predykcja HFTS opisującego zanieczyszczenie powietrza na Śląsku
   - Odporna AS dobowego zanieczyszczenia powietrza w Krakowie

Wprowadzenie

- Funkcjonalna analiza danych (FDA) to obszar współczesnej statystyki, w ramach którego bada się obiekty o wysokich wymiarach, gdzie każdy obiekt rozpatruje się ze względu na cechę, którą w naturalny sposób można indeksować pewną niezależną zmienną np. związaną z czasem bądź przestrzenią (por. prace prof. Krzyśko wraz z zespołem 2006–2018).
- FDA jest zasadniczo odmienna od jedno i wielowymiarowej statystyki zarówno z punktu widzenia matematycznego jak i empirycznego. Najważniejsza różnica wiąże się z tym, że na obserwacje patrzymy jak na realizacje funkcjonalnej zmiennej losowej albo jak na trajektorie odpowiedniego procesu stochastycznego.
- O ile metodologię wykrywania obserwacji odstających w przypadku wielowymiarowym cechuje znaczący stopień zaawansowania, to nie jest tak w przypadku FDA, gdzie wnioskowanie zależy często od niewielkiej liczby obserwacji w porównaniu do liczby stopni swobody, obserwacje funkcjonalne stosunkowo łatwo zanieczyścić.

Przykłady zastosowań FDA mające wymiar ekonomiczny

- Modelowanie umieralności na nowotwory płuc (Erbas i in., 2007)
- Predykcja obrotów na rynku opcji (Shen and Huang, 2008)
- Przewidywanie zmian klimatycznych (Shang and Hyndman, 2011)
- Modelowanie zjawisk demograficznych (Hyndman i Shang, 2009)
- Przewidywanie popytu na elektryczność (Antoch et al., 2008)
- Modelowanie transakcji dokonywanych kartami płatniczymi (Horvath i Kokoszka, 2011), Kargin i Onatski (2008)
- Monitorowanie zachowań internautów Kosiorowski i in (2017, 2018a)
- Modelowanie zanieczyszczenia powietrza Kosiorowski i in (2017, 2018b)
- Wykrywanie nietypowych cykli gospodarczych Kosiorowski i in (2018c)

Pierwsza wątpliwość–przejście danych do ciągłych funkcji

Na ogół zakładamy, że obserwujemy obiekt i w $j-tej$ chwili gdzie $i = 1, ..., K$, $j = 1, ..., J_i$, ze względu na zmienną X, tzn, x_{ij} . Warto podkreślić, że dopuszczamy sytuację obserwacji pojawiających się w nierównych odstępach czasu, $t_2 - t_1 \neq t_3 - t_2$. Punktem wyjścia dla dalszych analiz stanowią dane postaci $\{t_{ij}, x_{ij}\}$, gdzie $i = 1, ..., K$, $j = 1, ..., J_i$.

W ramach FDA na takie dane patrzymy jak na ciągłe funkcje czasu $x(t)$, $t \in [0, T]$.

Przejście od danych empirycznych do ciągłych funkcji

Po pierwsze musimy przekształcić dyskretne dane $\{t_{ij}, x_{ij}\}$ do danych funkcjonalnych $\{x_i(t), i = 1, ..., K; t \in [0, T]\}$. Możemy w tym celu wykorzystać interpolację bądź wygładzanie.

Najczęściej pierwszym krokiem w FDA jest wybór układu funkcji bazowych tzn. funkcji ϕ_k, $1, ..., L$ wykorzystywanych w celu wyrażenia $x(t)$ jako ich liniowej kombinacji.

$$x(t) = \sum_{k=1}^{L} c_k \phi_k(t) = \mathbf{c}^T \phi(t),$$

$t \in [0, T]$, gdzie $c_1, c_2, ..., c_k$ oznaczają współczynniki tej kombinacji.
Przykład układu bazowego to szeregi Fouriera: $\phi_1(t) = 1$, $\phi_2(t) = \sin(\omega t)$, $\phi_3(t) = \cos(\omega t), \ldots; \omega = 2\pi/T$.

Najczęściej współczynniki $c_1, c_2, ..., c_l$ reprezentacji $x(t) = \sum_{k=1}^{L} c_k \phi_k(t) = \mathbf{c}^T \phi(t)$, optymalizuje się *stosując zwykłą MNK* dla każdej funkcji oddzielnie.
(por. Ramsay i B. W. Silverman, 2005 Krzyśko i in., 2013)

Spojrzenia na obserwacje funkcjonalne

OBSERWACJE FUNKCJONALNE

Traktujemy obserwacje jak losowe krzywe $X = \{x(t), t \in [0, T]\}$, gdzie T jest ustalone, tzn. jako losowe elementy ośrodkowej przestrzeni Hilberta $L^2([0, T])$ wyposażonej w zwykły iloczyn skalarny $\langle x, y \rangle = \int x(t)y(t)dt$ (por. Horváth & Kokoszka, 2012).

Przestrzeń probabilistyczna wyposażona jest w Borelowską $\sigma-$algebrę zdarzeń.

FUNKCJONALNE SZEREGI CZASOWE

Funkcjonalny szereg czasowy to szereg funkcji indeksowanych czasem.

Próba z modelu funkcjonalnej autoregresji FAR(1)

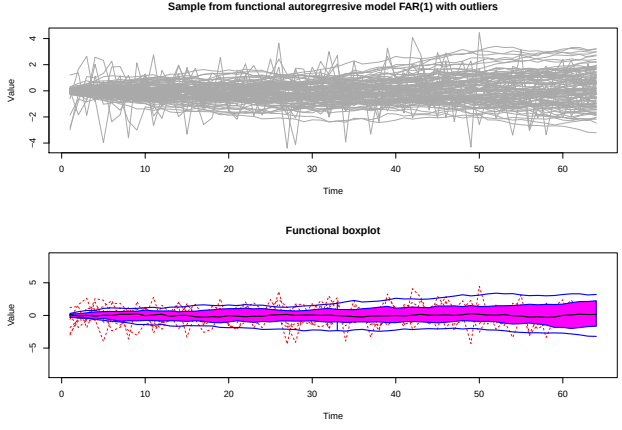
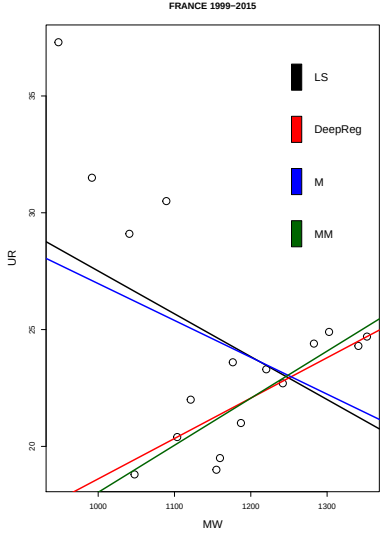


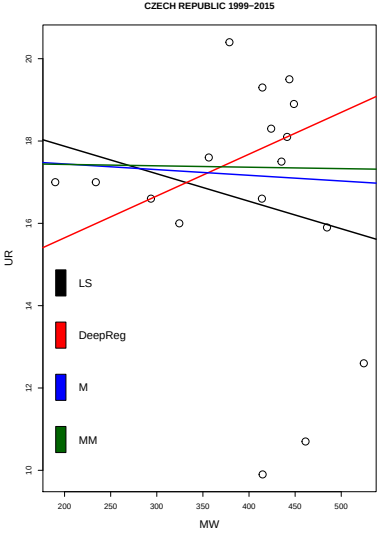
Figure 1: Próba 100 obs. z procesu FAR(1) z jądrem wykładniczym

Odporność procedury statystycznej

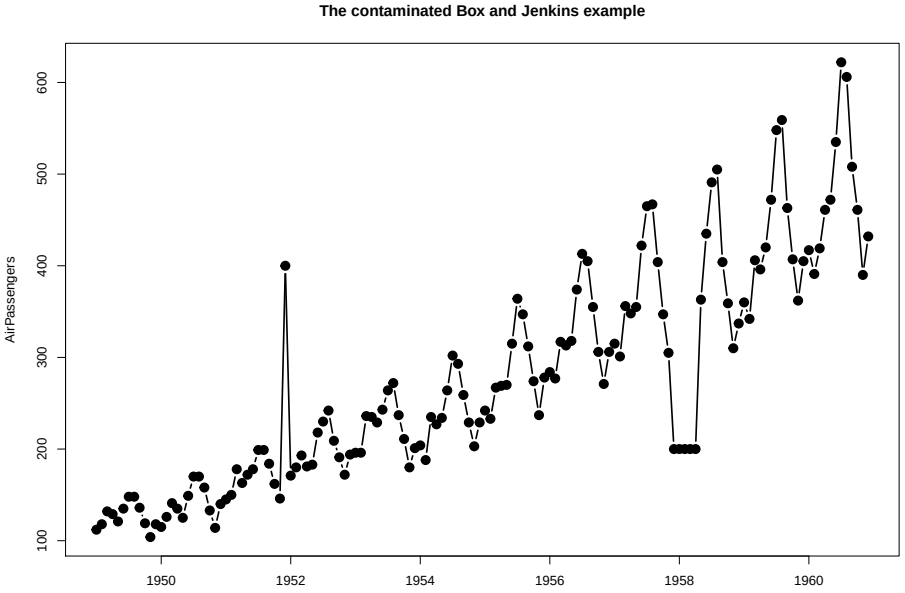
- Zasadniczy cel statystyki odpornej (ang. *robust statistics*) sprowadza się do wskazania na podstawie próby wzorca, charakterystyki populacji, które odzwierciedlają większość masy probabilistycznej. Dysponując takim wzorcem, definiuje się jednostki odstające (ang. outliers) jako odbiegające od niego w sensie pewnej miary bliskości.
- W statystyce odpornej staramy się wskazać procedury statystyczne posiadające dobre własności zarówno w sytuacji, gdy dane generują zakładany model, jak i wtedy, gdy prawdziwy model znajduje się w pewnym sąsiedztwie zakładanego modelu. Na poziomie próby staramy się odkryć, co wskazuje większość danych.
- Odporna procedura decyzyjna, to taka, która pomimo niedoskonałości danych, nie prowadzi do wysokich strat decydenta, który z niej korzysta.
- O ile w ogólności można przyjąć, że procedura odporna, to taka, w przypadku której niewielkie zmiany na wejściu (dane) prowadzą do niewielkich zmian na wyjściu (decyzja), to w konkretnych zagadnieniach statystycznych koncepcja i metoda pomiaru odporności mogą się istotnie różnić.

Zależność pomiędzy stopą bezrobocia (UR) a płacą minimalną (MW) we Francji i Republice Czeskiej w okresie 1999-2016





Przykład szeregu czasowego 1D zanieczyszczonego outlierami



Formalne koncepcje odporności

Niech pewna zmienna losowa X przyjmuje wartości z pewnej przestrzeni próbkowej $(\mathbf{X}, \mathbb{B})$. Oznaczmy przez P jej rozkład prawdopodobieństwa (zob. Bosq 2000). Rozważmy odporność funkcjonału $T(P) = T(X)$. Funkcjonał ten jest szacowany na podstawie próby $X^n = \{X_1, ..., X_n\}$ złożonej z n niezależnych kopii X. Ściślej, szacujemy T za pomocą funkcjonału empirycznego $T_n(P_n) = T_n(X_1, ...X_n)$, opierającego się o rozkład empiryczny P_n policzony z próby $X^n = \{X_1, ..., X_n\}$.

Można zaproponować następującą modyfikację definicji F. Hampela jakościowej odporności funkcjonału T, która wykorzystuje pewną ustaloną metrykę d_P na rodzinie miar probabilistycznych $\mathbb{P}$:

Definicja 1: Powiemy, że ciąg funkcjonałów $\{T_n\}$ jest **jakościowo odporny** w $P \in \mathbb{P}$, wtedy i tylko wtedy gdy dla dowolnego $\varepsilon > 0$, istnieje $\delta > 0$ oraz dodatnia liczba całkowita n_0 taka że, dla wszystkich $Q \in \mathbb{P}$ oraz $n > n_0$

$$d_P(P, Q) < \delta \Rightarrow d_T(L_P(T_n), L_Q(T_n)) < \varepsilon,$$

gdzie $P, Q \in \mathbb{P}$ oznaczają dwa rozkłady w przestrzeni Hilberta L^2 oraz L_P, L_Q oznaczają charakterystyki rozkładów P, Q. L_{P_n}, L_{Q_m} mogą oznaczać wartości miar jakości wyników analizy skupisk, wartości błędów klasyfikacji itp.

Mniej formalnie – punkty załamania procedury statystycznej

DEFINICJA 2: **Punkt załamania próby skończonej** (FSBP) procedury to maksymalna frakcja "złych" obserwacji w próbie, którymi zastępujemy dobre obserwacje, która nie prowadzi do załamania jakości procedury mierzonej np. za pomocą jej obciążenia bądź efektywności

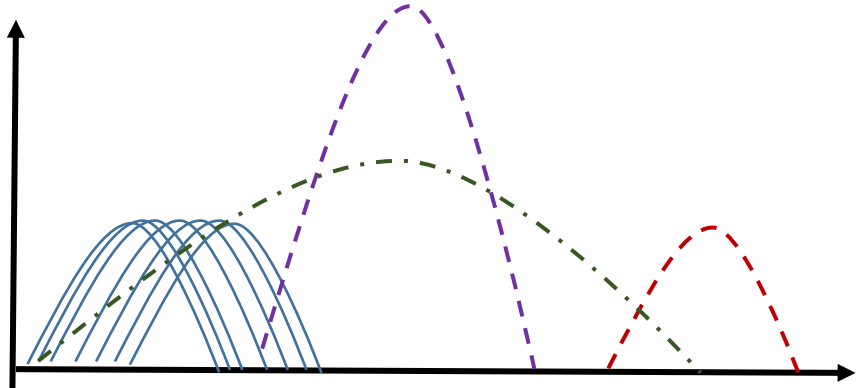
DEFINICJA 3: **Punkt załamania Gentona i Lucasa** (GLBP) to maksymalna frakcję obserwacji w próbie prowadzącą do utraty następującej własności funkcjonału statystycznego (procedury AS, klasyfikatora): *kontinuum możliwych prób prowadzi do kontinuum możliwych decyzji zależnych od wartości funkcjonału (wyniku AS, wartości klasyfikatora* tzn. obecność outlierów w próbie powoduje, że wartość funkcjonału prowadzi do decyzji z brzegu przestrzeni decyzji.

Procedura odporna to procedura o niskim punkcie załamania

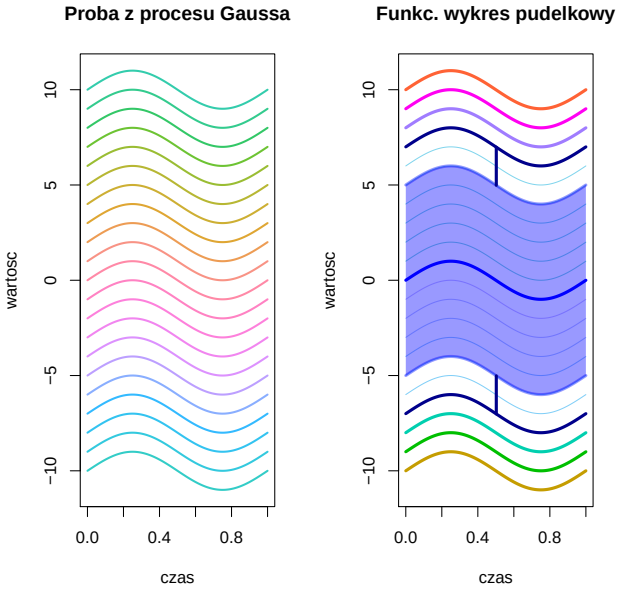
Obserwacje odstające pośród danych funkcjonalnych

- Funkcjonalne outliery (**FO**) mogą niekorzystnie wpływać na zachowanie się procedur statystycznych wykorzystywanych w ramach FDA takich jak funkcjonalna analiza wariancji i składowych głównych, funkcjonalna regresja bądź analiza dyskryminacyjna.
- Z drugiej strony FO mogą sygnalizować nieprawidłowości, oszustwa w wymianie handlowej, "systematyczne niegwałtowne " włamania do systemów transakcyjnych i informacyjnych.
- Outliery mogą sygnalizować pojawienie się nowego zjawiska.
- Występowanie zależności w czasie negatywnie wpływa na własności narzędzi do wykrywania outlierów opierające się na popularnych funkcjach głębi
- O ile metodologia wykrywania obserwacji odstających w przypadku wielowymiarowym cechuje znaczący stopień zaawansowania, to nie jest tak w przypadku FDA, gdzie wnioskowanie zależy często od niewielkiej liczby obserwacji w porównaniu do liczby stopni swobody, obserwacje funkcjonalne stosunkowo łatwo zanieczyścić.

Funkcjonalne outliery – wprowadzenie



Funkcjonalne outliery – wprowadzenie



Funkcjonalne outliery cd.

Zasadniczy problem każdej procedury FDA – właściwa analiza zmienności danych. Próby obserwacji – realizacji losowych funkcji – losowych elementów pewnych przestrzeni funkcyjnych na ogół nieskończenie wymiarowych, rzeczywistych, ośrodkowych przestrzeni Banacha lub Hilberta. Zupełność i ośrodkowość przestrzeni są kluczowym założeniem – gwarantują, że sumowalna liniowa kombinacja elementów losowych dalej jest elementem losowym rozważanej przestrzeni a jednocześnie każdy element losowy jest kombinacją liniową co najwyżej przeliczalnej liczby elementów z pewnej bazy (ośrodka). Rozważamy losowe funkcje postaci

$$X : (\Omega, \mathcal{B}, \mathcal{P}) \to \mathcal{V},$$

gdzie $(\Omega, \mathcal{B}, \mathcal{P})$ jest przestrzenią probabilistyczną oraz $\mathcal{V}$ oznacza ośrodkową przestrzeń Banacha lub Hilberta wyposażoną w normę $\|\cdot\|$, w przypadku przestrzeni Hilberta, norma indukowana jest przez iloczyn skalarny. Dla wszystkich $\omega \in \Omega$ mamy

$$X_\omega : t \to X(\omega, t) \in \mathcal{V}$$

Statystyczne funkcje głębi – podstawowe narzędzia odpornej FDA

Statystyczna funkcja głębi – przypadek wielowymiarowy: rozważmy Borelowską σ algebrę $\mathcal{B}$ podzbiorów $\mathbb{R}^p$ oraz rodzinę rozkładów prawdopodobieństwa $\mathcal{P}$ zdefiniowanych na $\mathcal{B}$. **Miara głębi** jest funkcją

$$D : \mathbb{R}^p \times \mathcal{P} : (z, P) \to D(z|P) \in [0, 1],$$

spełniajacą pewien zestaw warunków (patrz Zuo i Serfling, 2000).
Dysponując wektorem losowym $X : (\Omega, \mathcal{B}) \to \mathbb{R}^p$ z indukowanym rozkładem prawdopodobieństwa P_X powiemy, że

$$D(\cdot|P_X) : \mathbb{R}^p \to [0, 1],$$

jest **statystyczną funkcją głębi**. Powiemy też, że $D(z|X) \in [0, 1]$ jest głębią punktu z względem rozkładu prawdopodobieństwa wektora X .

Dla próby $X^n = \{x_1, ..., x_n\}$ z X, zdefiniujemy empiryczną funkcję głębi – wielkość $D(\cdot|X^n) = D(\cdot|\tilde{P}_n)$, gdzie P_X zostaje zastąpiony przez rozkład empiryczny $\tilde{P}_n$. Dla danego wektora losowego X oraz miary głębi D oraz wartości $\alpha \in [0, 1]$ definiujemy $\alpha-$ obszar centralny jako

$$C_\alpha(X) = \{z \in \mathbb{R}^p : D(z|X) \geq \alpha\}.$$

Głębię obserwacji można interpretować jako miarę jej odstawania od centrum – za pomocą głębi wykrywać outliery.

Popularnym przykładem jest głębia **Tukey'a**. Dla $x \in \mathbb{R}^d$

$$D(x; F) = \inf_H \left\{ Prob(H) : x \in H \subset \mathbb{R}^d, \text{ H jest domkniętą podprzestrzenią} \right\},$$

gdzie *Prob* oznacza prawdopodobieństwo. W przypadku próby $X^n = \{x_1, ..., x_n\}$ w miejsce F podstawiamy jego empiryczny odpowiednik F_n.

Funkcja głębi z próby–obliczenia pakiet DepthProc

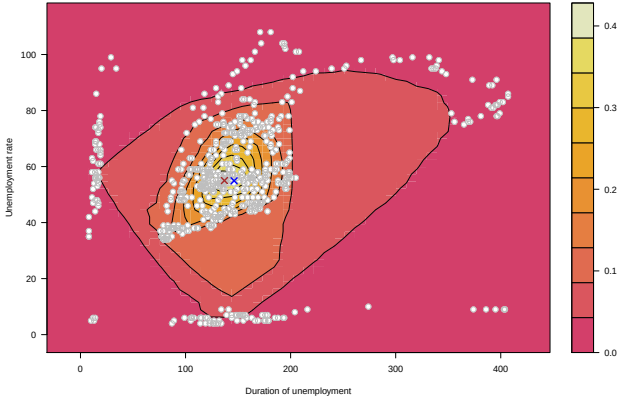
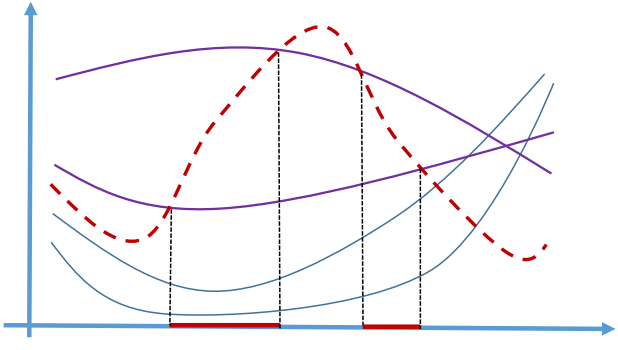


Figure 5: Stopa bezrobocia vs. czas trwania bezrobocia w USA w okresie 1957–2011. Wykres konturowy sporządzono z wykorzystaniem głębi Tukey'a.

MBD - głębia dla danych funkcjonalnych



Przykłady funkcji głębi dla danych funkcjonalnych

Generalized band depth (GBD, por. López-Pintado and Jörnsten, 2007) oraz modified band depth (MBD, por. López-Pintado and Romo, 2009) funkcji x względem próby funkcji X^N mierzy częstość znajdowania się krzywej w centrum. Dysponujemy próbą n funkcji $X^n = \{x_1, ..., x_n\}$. Zdefiniujmy:

$$GBD(x|X^n) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \frac{CI(x; x_{i_1}, x_{i_2})}{\lambda([0,T])}, \tag{1}$$

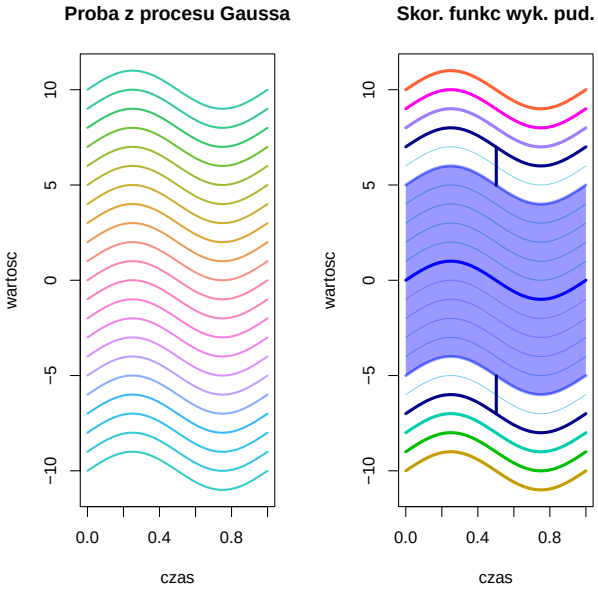
gdzie $CI(x; x_{i_1}, x_{i_2}) = \max_{t_S}\{\lambda(t_S) : \min_{r=i_1,i_2} x_r \le x(t) \le \max_{r=i_1,i_2} x_r, \forall t \in t_S\}$, gdzie $t_S \subset [0,T]$ oznacza przedział zwarty. Zatem $CI(x; x_{i_1}, x_{i_2})$ jest najdłuższym przedziałem zwartym, na ktorym x jest w paśmie wyznaczonym przez x_{i_1}, x_{i_2}.

Przykłady funkcji głębi dla danych funkcjonalnych cd.

Niech tym razem $A(x; x_{i_1}, x_{i_2}) = \{t \in [0, T] : \min_{r=i_1,i_2} x_r \leq x(t) \leq \max_{r=i_1,i_2} x_r\}$.
Modified band depth (MBD) bierze pod uwagę "frakcję czasu", gdy x znajduje się w odpowiednim paśmie wyznaczonym przez elementy próby, a mianowicie

$$MBD(x|X^n) = \frac{2}{n(n-1)} \sum_{1 \leq i_1 < i_2 \leq n} \frac{\lambda(A(x; x_{i_1}, x_{i_2}))}{\lambda([0, T])}. \tag{2}$$

Wykrywanie funcjonalnych outlierów typu "magnitude"



Wykrywanie funcjonalnych outlierów typu "magnitude"

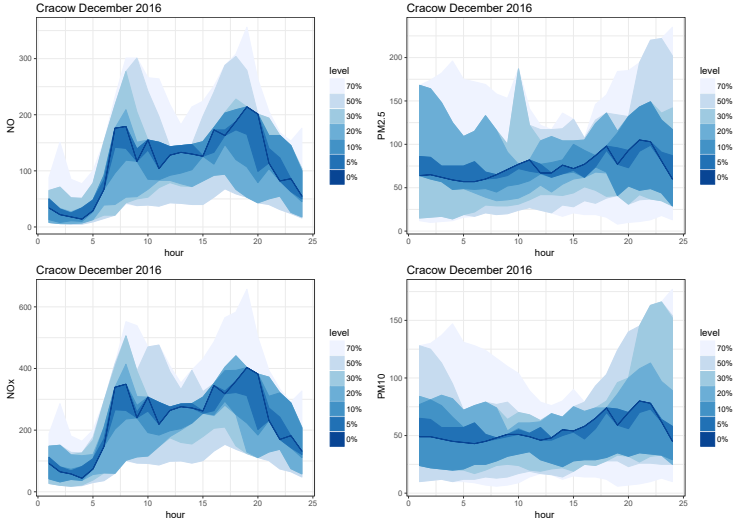


Figure 8: funkcjonalny boxplot

Wykrywanie outlierów w FTS–aktualny stan wiedzy

- Febrero i in. (2007, 2008): Pierwszy bootstrapowy test wykrywania do outlierów dla przypadku niezależnych obserwacji o tym samym rozkładzie.
- Funkcjonalny boxplot i skorygowany boxplot Sun i Genton (2011, 2012), Tarabelloni (2017).
- Outliergram zaproponowany przez Arribas-Gil i Romo (2014).
- Rana i in. (2015): W przypadku FTS setup, lokalne trendy wynikające ze struktury zależności mogą maskować obecność outlierów–zaproponowali modyfikację procedury Febrero i in.(2008).
- Nagy i in. (2017): Oryginalne podejście do klasyfikacji i wykrywania outlierów wykorzystujące pochodne funkcji i tzw. głębie całkowe.

Outlierogram–wykrywanie outlierów typu "shape"

Dysponujemy próbą $X^n = \{X_1, ..., X_n\}$ funkcji określonych na przedziale $I \subset \mathbb{R}$. Modified Epigraphic Index (MEI) funkcji Z względem próby X^n:

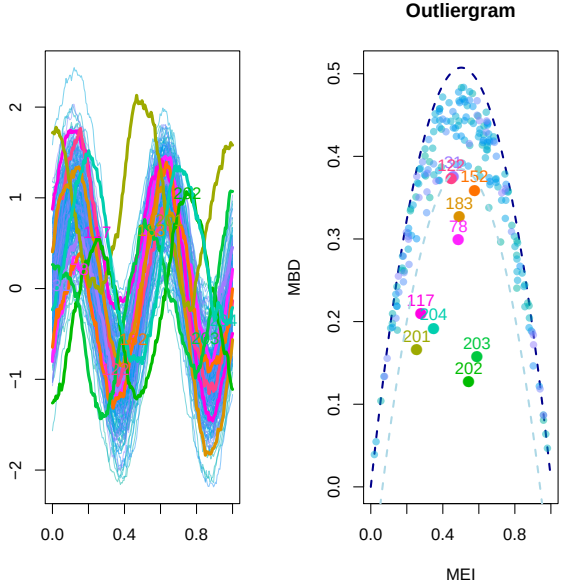
$$MEI(Z|X^n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda(\{t \in I : Z(t) \leq X_i(t)\})}{\lambda(I)},$$

gdzie λ oznacza miarę Lebesgue'a. Arribas- Gil i Romo (2014) pokazali, że MEI i MBD (zmodyfikowana głębia pasma) pozostają w relacji, którą można wykorzystać do wykrywania outlierów typu "shape":

$$MBD(Z, X^n) \leq a_0 + a_1 MEI(Z|X^n) + a_2 n^2 MEI^2(Z|X^n),$$

gdzie $a_0 = a_2 = -2/n(n-1)$, $a_1 = 2(n+1)/(n-1)$.
Outliery typu "shape" mają niewielką wartość MBD w porównaniu do tej opisanej za pomocą parabolicznego trendu, rozbieżność pomiędzy prawą i lewą stroną nierówności pozwala je zidentyfikować.

Wykrywanie funcjonalnych outlierów typu "shape"



FUNKCJONALNE SZEREGI CZASOWE (FTS)

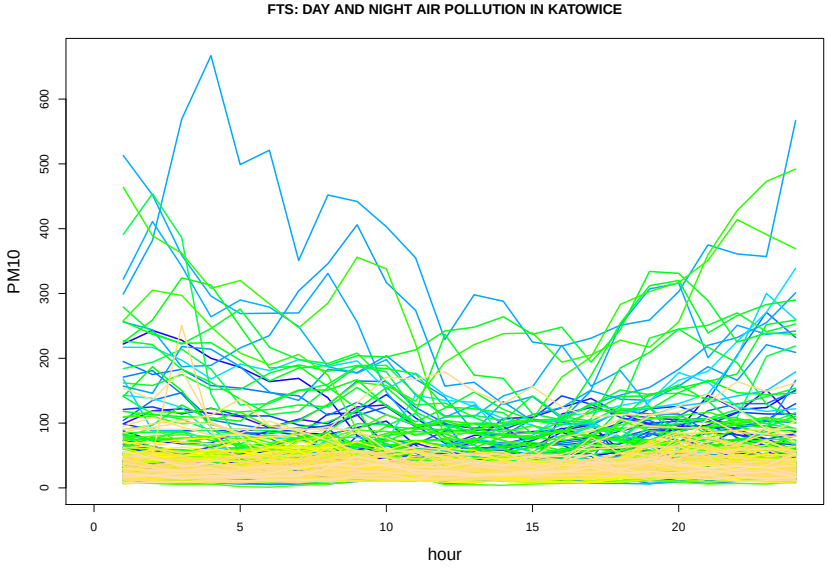
OBSERWACJE FUNKCJONALNE

Traktujemy obserwacje jak losowe krzywe $X = \{x(t), t \in [0, T]\}$, gdzie T jest ustalone, tzn. jako losowe elementy ośrodkowej przestrzeni Hilberta $L^2([0, T])$ wyposażonej w zwykły iloczyn skalarny $\langle x, y \rangle = \int x(t)y(t)dt$ (por. Horváth & Kokoszka, 2012).

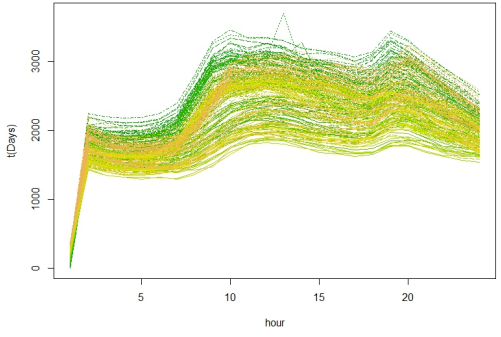
FUNKCJONALNE SZEREGI CZASOWE

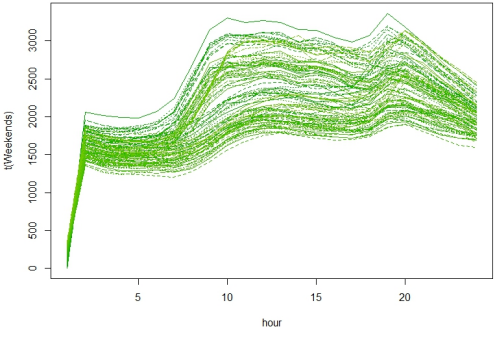
Funkcjonalny szereg czasowy to szereg funkcji indeksowanych czasem.

Przykład funkcjonalnego szeregu czasowego



Obserwacje funkcjonalne pokazujące zużycie energii elektrycznej w dni robocze (po lewej) i w weekendy (po prawej) w 2016 w Danii





Hierarchiczne funkcjonalne szeregi czasowe (HFTS)

W ekonomii często stykamy się z danymi, w których występuje pewna forma hierarchii, tzn. np., zanieczyszczenie powietrza bądź zapotrzebowanie na energię elektryczną w ujęciu gminy, powiatu, województwa, regionu, państwa.

Hierarchiczny szereg czasowy to ciąg obserwacji, w przypadku HFTS funkcji, zgrupowanych względem ustalonej struktury hierarchii (np. gospodarstwo domowe, miasto, region, stan).

Naszym celem jest uzyskania prognozy zjawiska z **uzgodnionymi (ang. reconciliated)** prognozami uzyskiwanymi na różnych poziomach hierarchii.

Hierarchiczne funkcjonalne szeregi czasowe (HFTS)

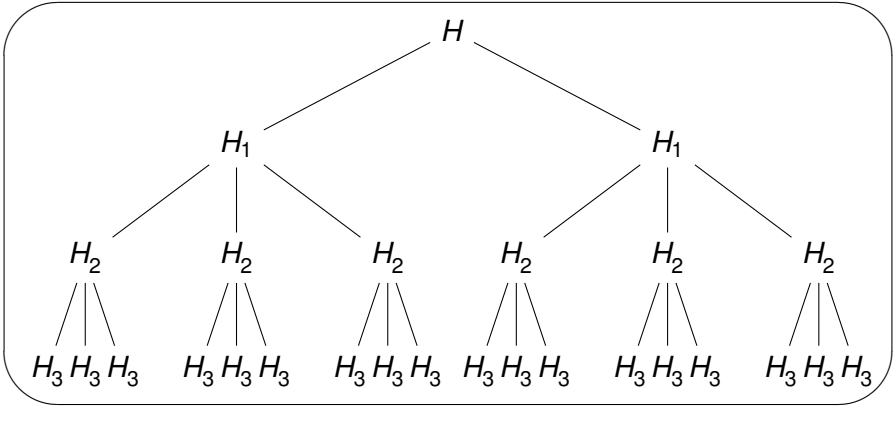
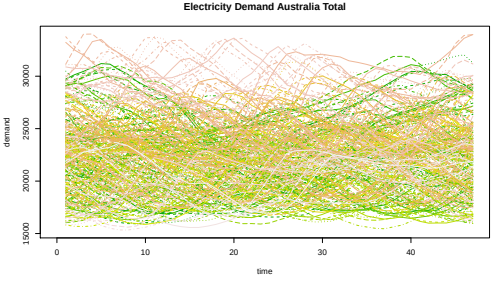
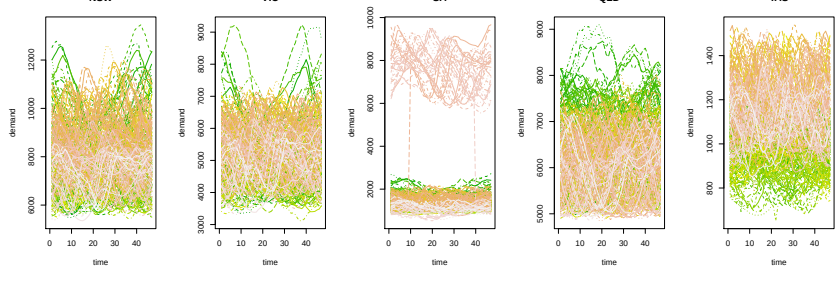


Figure 11: Przykładowa struktura hierarchicznego szeregu czasowego

Przykład hierarchicznego szeregu czasowego (HFTS)





Znane metody prognozowania HFTS

METODY:
Propozycja Hyndmana i Shanga (2017), rodzina naszych propozycji na bazie ruchomej funkcjonalnej mediany (2017a,b,2018), ruchoma funkcjonalna średnia, metoda naiwna

KRYTERIA OCENY PREDYKTORA HFTS:

* trafność predykcji
* zgodność (statystyczna)
* zgodność agregacyjna (ang. aggregate consistency termin merytoryczny, H&S)
* nieobciążoność
* odporność
* złożoność obliczeniowa
* sposób oceny niepewności prognozy

Predyktor referencyjny–propozycja Hyndmana i Shanga (2017)

Podejście wywodzi się z wcześniejszych prac autorów dotyczących szeregów hierarchicznych 1D. Uzgodniona prognoza ma postać

$$\hat{X}_{n+1}(t) = F(\hat{x}^{level_1}, \hat{x}^{level_2}, ..., , \hat{x}^{level_L}),$$

gdzie $\hat{x}^{level_i}$ oznacza prognozy uzyskane dla FTS na poziomie i oraz F jest pewną postacią estymatora uogólnionej metody najmniejszych kwadratów (GMNK).

Propozycja Hyndma i Shanga część 2

Struktura HFTS w chwili t:

$$X_t = S_t b_t,$$

gdzie wektor X_t zawiera wszystkie szeregi na wszystkich poziomach hierarchii;
b_t jest wektorem najbardziej zdezagregowanych szeregów (tzn., na najniższym poziomie hierarchii);
S_t jest macierzą zależności pomiędzy wektorami.
Prognoza przybiera postać

$$\hat{X}_{n+1} = S_{n+1}\beta_{n+1} + \epsilon_{n+1},$$

gdzie $\hat{X}_{n+1}$ jest macierzą prognoz dla wszystkich szeregów na wszystkich poziomach hierarchii. $\beta_{n+1} = E[b_{n+1}|X_1, ..., X_n]$ jest nieznaną wartością oczekiwaną rozkładu prognozy dla szeregów na najniższym poziomie oraz ϵ_{n+1} oznacza błędy związane z uzgadnianiem prognoz.

Propozycja Hyndmana i Shanga część 3

Prognozy dla poszczególnych poziomów uzyskuje się stosując wysoce efektywną jednak nieodporną metodę, w której FTS są przekształcane do rodzin jednowymiarowych szeregów czasowych współrzędnych funkcjonalnych składowych głównych (ang. functional principal component scores) (por. Kosiorowski, 2014), β_{n+1} w podejściu Shanga i Hyndmana (2017) szacowane są za pomocą uogólnionej metody najmniejszych kwadratów GMNK:

$$\hat{\beta}_{n+1} = \left(S_{n+1}^T W^{-1} S_{n+1} \right)^{-1} S_{n+1}^T W^{-1} \hat{X}_{n+1},$$

gdzie W jest macierzą diagonalną, z wariancjami prognoz dla szeregów. Finalnie, prognozę uzyskujemy z równania

$$\overline{X}_{n+1} = S_{n+1} \hat{\beta}_{n+1}.$$

Zalety i wady metody Hyndmana i Shanga

Prognozy metodą H&S są agregacyjnie zgodne (spełniają ograniczenia narzucone przez sposób agregacji szeregów) oraz są nieobciążone w sensie wartości oczekiwanej.

Metoda jest bardzo wymagająca pod względem obliczeniowym i programistycznym (optymalizacja MNNK przy rzadkich macierzach eksperymentu, uogólnione odwrotności wielkich macierzy).
Metodę da się jednak "uodpornić" (por Kosiorowski i in., 2017b, 2018b).

Metoda Shanga i Hyndmana jest wyposażona w bezpośredni wewnętrzny mechanizm uzgadniania prognoz, dla porównania w naszej metodzie podwójnej mediany, uzgodnienie jest produktem ubocznym własności konkretnej mediany funkcjonalnej, nieprzechodniości głębi, która ją indukuje.

Metoda S&H redukuje zagadnienie predykcji FTS do zagadnienia pewnej formy funkcjonalnej regresji (ang. functional principal component regression – funkcje przedstawia się w bazie empirycznych funkcjonalnych składowych a następnie stosuje się metody prognozowania stacjonarnych szeregów czasowych 1D do współrzędnych tej reprezentacji (por. Kosiorowski, 2014)).

Rodzina propozycji na bazie ruchomej funkcjonalnej mediany

Mamy próbę N funkcji, tzn., $X^N = \{x_i(t), i = 1, 2, ..., N\}$ oraz $t \in [0, T]$. Niech $FD(y|X^N)$ oznacza funkcjonalną głębię $y(t)$ z próby X^N. Medianę z próby definiujemy jako

$$MED_{FD}(X^N) = \arg\max_{i=1,...,N} FD(x_i|X^N).$$

Uwaga 1: jeżli maksimum przyjmuje więcej niż jedna funkcja, to jako medianę przyjmujemy ich punktową średnią.
Uwaga 2: Różne funkcjonalne głębie mogą prowadzić do różnych median, uwypuklających różne aspekty kształtu/położenia większości danych.

Nasze propozycje odpornych predyktorów

- Uogólnione wygładzanie wykładnicze dla FTS z parametrem lokalności Kosiorowski i in (2016, 2018a SiT).
- Metoda podwójnej funkcjonalnej mediany Kosiorowski, Mielczarek, Rydlewski (2018b CEJEME).
- Metoda zagregowanej funkcjonalnej mediany Kosiorowski, Mielczarek, Rydlewski (2018c) submitted.

Propozycja metody podwójnej mediany

W propozycji korzystamy z ruchomej funkcjonalnej mediany:

$$\hat{x}_{n+1}(t) = MED_{FD}(W_{n,k})$$

gdzie $W_{n,k}$ jest ruchomym oknem o długości k z końcem w chwili n, tzn.,

$$W_{n,k} = \{x_{n-k+1}, ..., x_n\}.$$

Uwaga: można zastosować układ wag, w którym najbliższe obserwacje "ważą więcej" niż obserwacje bardziej odległe w czasie.

Propozycja metody podwójnej mediany– szczegóły

Pierwszy krok: obliczamy ruchomą funkcjonalną medianę indukowaną przez głębię MBD dla każdego węzła na najniższym poziomie hierarchii tzn.,

$$MED_{MBD}(W_{n,k}).$$

W przykładzie empirycznym dla każdego miasta w chwili t, liczymy funkcjonalną medianę indukowaną przez *MBD* na podstawie ruchomego okna o długości 10:

$$\hat{x}_{t+1}^{miasto} = MED_{MBD}\{x_t^{miasto}, x_{t-1}^{miasto}, ..., x_{t-9}^{miasto}\}.$$

Drugi krok: dla poziomu hierarchi wyższego o jeden liczymy funkcjonalną medianę z median policzonych w kroku pierwszym.
Powtarzamy drugi krok do chwili, gdy dotrzemy do najwyższego poziomu hierarchii.
W przykładzie empirycznym drugi krok jest krokiem ostatnim, otrzymujemy prognozę dla $t = 10, ..., 181$:

$$\hat{x}_{t+1} = MED_{MBD}\{\hat{x}_{t+1}^{miasto_1}, ..., \hat{x}_{t+1}^{miasto_5}\}.$$

Propozycja metody podwójnej mediany– szczegóły

Uwaga: Otrzymana mediana dla pewnego węzła nie musi równać się medianie wszystkich funkcji, które biorą udział w algorytmie obliczniowym, który do niego prowadzi. Jest to bezpośrednia konsekwencja faktu, że operacja brania mediany nie jest przechodnia. Nawet dla dwóch równolicznych zbiorów może zachodzić

$$MED\{x_1, ..., x_m, y_1, ..., y_m\} \neq MED\{MED\{x_1, ..., x_m\}, MED\{y_1, ..., y_m\}\}.$$

Inaczej niż w przypadku średnich funkcjonalnych gdzie zachodzi równość.

Zatem w propozycji *hierarchiczna struktura danych zostaje uwzględniona w procesie liczenia median funkcjonalnych z median z niższych poziomow hierarchii*.

Ocena niepewności prognozy

W ujęciu ogólnym, zagadnienie oceny niepewności predyktora HFTS jest zagadnieniem otwartym.
S&H stosują częsty zabieg w FDA –przekształcają FTS w rodzinę jednowymiarowych szeregów czasowych współrzędnych empirycznych składowych głównych a następnie stosują pewną metodę bootstrap do tych szeregów czasowych (bootstrap maksymalnej entropii Vinod i de Lacalle (2009)

Jakkolwiek, ich zabieg jest sprytny i użyteczny, stosując go tracimy z oczu bogactwo zachowań FTS w porównaniu do zachowń szeregów 1D.
W naszych propozycjach koszystamy z funkcjonalnych boxplotów i skorygowanych boxplotów (w szczególności objętości pudełek i $\alpha-$centralnych obszarów), które można wykorzystać do oceny niepewności prognoz FTS (por López-Pintado i in., 2010).

Typowe podejścia do AS w FDA

- Obserwacje funkcjonalne wyrażamy w ustalonym układzie funkcji bazowych a następnie stosujemy klasyczne metody AS do współrzędnych tych reprezentacji – Abraham i in. (2003), Serban i Wasserman (2005), James i Sugar (2003) and Luan and Li (2003) – wyniki AS są wrażliwe na wybór bazy, liczbę jej elementów, wybór węzłów).
- Dane rzutowane są na skończenie wymiarową przestrzeń rozpiętą przez kilka szczególnie ważnych empirycznych składowych głównych a następnie stosuje się klasyczne metody AS do współrzędnych obserwacji w układzie współrzędnych tych składowych Song i in (2007).
- Praca bezpośrednio z surowymi danymi empirycznymi przy nałożeniu pewnych ograniczeń co do gładkości funkcji Ma i in. (2006), Rocci i Gattone - zredukowane funkcjonalne k-średnich.

Algorytm k-średnich dla danych funkcjonalnych

- Niech $\{x_1(t), x_2(t), \ldots, x_I(t)\}$ będzie zbiorem funkcjonalnych obserwacji, gdzie $t \in \Gamma$, a Γ jest to przedział na prostej $\mathbb{R}$. Funkcje należą do ośrodkowej przestrzeni Hilberta.
- Załóżmy, że funkcjonalne obserwacje $x_i(t)$ dane są wzorem

$$x_i(t) = \sum_{g=1}^{G} u_{ig} m_g(t) + \varepsilon_i(t), \quad i = 1, \ldots, I, \tag{3}$$

gdzie

- m_g – gładkie nieznane funkcje (centroidy),
- $\varepsilon_i(t)$ – nieobserwowalne błędy losowe o średniej zero,
- $u_{ig} \in \{0, 1\}$, $\sum_g u_{ig} = 1$ dla każdego i oraz $u_{ig} = 1$, jeżeli x_i należy do g-tego skupiska.

Parametry u_{ig} i m_g są szacowane minimalizując wyrażenie

$$\sum_{i,g} u_{ig} \int_{\Gamma} \left[x_i(t) - m_g(t)\right]^2 dt = \sum_{i,g} u_{ig} \|x_i(t) - m_g(t)\|^2 \to \min . \tag{4}$$

Odporna analiza skupisk w FDA

- Zasadnicza kwestia – jak rozumieć odporność procedury AS?
- Odporna wersja algorytmu Warda.
- Metoda k-lokalnych median.

Algorytm Warda dla danych funkcjonalnych

Metoda Warda należy do metod grupowania aglomeracyjnego. Algorytm ten łączy grupy w taki sposób, aby miara niejednorodności wewnątrz skupienia nie wzrosła „zbyt wiele".

Niech A i B będą niepustymi podzbiorami przestrzeni L_2. Wówczas odległość pomiędzy zbiorami A i B zdefiniowana jest następująco

$$d(A,B) = \sum_{i \in A \cup B} \|f_i - \bar{f}_{A\cup B}\|^2 - \sum_{i \in A} \|f_i - \bar{f}_A\|^2 - \sum_{i \in B} \|f_i - \bar{f}_B\|^2,$$

gdzie $\|\cdot\|$ oznacza normę w przestrzeni L_2, a $\bar{f}_A$ jest centroidem skupienia A.

- Zauważmy, że odległość Warda określa o ile wzrośnie rozproszenie wewnątrz nowego skupienia po połączaniu skupień A i B.
- Odległość $d(A, B)$ nazywamy kosztem połączenia skupień A i B (ang. *merging cost*).
- jeżeli $A = \{f_1, \ldots, f_{n_A}\}$ oraz $B = \{g_1, \ldots, g_{n_B}\}$, to odległość pomiędzy dwoma skupieniami można zapisać jako

$$d(A, B) = \frac{n_A n_B}{n_A + n_B} \|\overline{f} - \overline{g}\|^2$$

- Odległość w metodzie Warda jest zawsze dodatnia oraz $d(A, B) = 0$ wtedy i tylko wtedy, gdy $\overline{f}_A = \overline{f}_B$.

Kryterium grupowania – metoda Warda

Szukamy podziału zbioru danych $\mathbf{C} \in \Pi_k$ minimalizującego wyrażenie

$$\sum_{C \in \mathbf{C}} \sum_{f \in C} \|f - \bar{f}_C\|^2,$$

gdzie

$$\Pi_k = \{\mathbf{C} : \mathbf{C} \text{ – podział zbioru danych i } \#\mathbf{C} = k\}$$

Rozszerzenie metody Warda – Szlachtowska (2018)

Metoda Warda jest oparta na odległości pomiędzy skupieniami lub równoważnie na odległości pomiędzy centroidami skupień. Niestety, metoda oparta na odległości pomiędzy centroidami nie jest efektywna dla skupień o równych średnich. W przestrzeni euklidesowej próbuje się modyfikować algorytm Warda poprzez modyfikacje odległości pomiędzy skupieniami. Nasze rozważania dotyczą przestrzeni L_2 oraz danych funkcjonalnych określonych na tej przestrzeni. Wyniki zawarte w pracy Székely i Rizzo przenieśliśmy dla danych funkcjonalnych.

Székely G.J., Rizzo M.L. (2005), ”Hierarchical Clustering Via Joint Between-Within Distances: Extending Ward’s Minimum Variance Method”, Journal of Classification, 22(2), 151–183.

Odległość między skupieniami

Niech $A = \{f_1, \ldots, f_{n_A}\}$ i $B = \{g_1, \ldots, g_{n_B}\}$ będą niepustymi podzbiorami przestrzeni L_2. Wówczas odległość pomiędzy zbiorami A i B definiujemy następująco

$$d(A,B) = \frac{n_A n_B}{n_A + n_B} \left(\frac{2}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \|f_i - g_j\|^\alpha - \frac{1}{n_A^2} \sum_{i=1}^{n_A} \sum_{j=1}^{n_A} \|f_i - f_j\|^\alpha - \frac{1}{n_B^2} \sum_{i=1}^{n_B} \sum_{j=1}^{n_B} \|g_i - g_j\|^\alpha \right), \tag{5}$$

gdzie $\|\cdot\|$ oznacza odległość w przestrzeni L_2, a $\alpha \in (0, 2]$.

Można wykazać, że dla każdego $\alpha \in (0, 2]$ w zmodyfikowanym algorytmie Warda odległość między skupieniami może być obliczana za pomocą rekurencyjnej formuły Lance'a-Williamsa.

Twierdzenie

Niech A, B, C będą rozłącznymi, niepustymi skończonymi podzbiorami przestrzeni L_2 takimi, że

$$d(A, B) \leq \min\{d(A, C), d(B, C)\}.$$

Wówczas rekurencyjna formuła dla odległości $d(A \cup B, C)$ wyraża się wzorem

$$\begin{aligned} d(A \cup B, C) = & \frac{n_A + n_C}{n_A + n_B + n_C} d(A, C) \\ & + \frac{n_B + n_C}{n_A + n_B + n_C} d(B, C) - \frac{n_C}{n_A + n_B + n_C} d(A, B) \end{aligned} \tag{6}$$

- W przypadku $\alpha = 2$ jest proporcjonalna do kwadratu odległości w przestrzeni L_2. W rezultacie szukamy takiego grupowania, które minimalizuje wariancję w skupieniach, czyli minimalizuje funkcję celu w metodzie Warda. Należy się zatem spodziewać, że odległość $d(A, B)$ jest wyrażona jako ważony kwadrat odległości między centroidami (średnimi) skupień A i B.

Twierdzenie

Niech $A = \{f_1, \ldots, f_{n_A}\}$ *i* $B = \{g_1, \ldots, g_{n_B}\}$ *będą podzbiorami przestrzeni* L^2*. Wówczas*

$$d(A, B) = \frac{2n_A n_B}{n_A + n_B} \|\overline{f} - \overline{g}\|^2,$$

gdzie $\overline{f}$ *jest centroidem skupienia A, a* $\overline{g}$ *jest centroidem skupienia B.*

Metoda k-lokalnych median Szlachtowska i Kosiorowski (2017)

- Globalna vs. lokalna funkcja głębi – Paindaveine i Van Bever (2013)
- Lokalność głębi funkcjonalnej – Squera et i in. (2016); nasze propozycje na bazie podeścia Paindaveine i Van Bever (2013)
- Odporna analiza skupisk z wykorzystaniem głębi lokalnych

Algorytm k-lokalnych median dla danych funkcjonalnych

Wybór parametrów:

1. Wybieramy parametr k, gdzie oznacza wymaganą liczbę skupień.
2. Wybieramy parametr β (domyślna wartość β to 0,2). Za pomocą parametru β możemy rozpatrzyć problem na różnych „poziomach rozdzielczości”, tj. zmieniając wartość parametru β możemy kontrolować dokładność podziału. Z drugiej strony obliczone wartości lokalnej głębi dla wszystkich punktów są pomocne przy wyborze odpowiedniej liczby skupień.

Algorytm k-lokalnych median dla danych funkcjonalnych cd.

W rozważanym algorytmie grupowania rozważamy dwie miary podobieństwa/bliskości. Jako odporną miarę podobieństwa użyjemy lokalną zmodyfikowaną głębię pasma. Kryterium oparte na lokalnej głębi stanowi pierwszą część funkcji celu, w której mierzymy jakość grupowania ze względu na odporność

$$F = \sum_{i=1}^{k} \sum_{j \in C_i} LD^{\beta}(f_j, P_{n,i}),$$

gdzie $P_{n,i}$ oznacza rozkład empiryczny w i-tym skupieniu. Natomiast jako miarę przydziału do skupień będziemy używać odległości L_2.

W pierwszym kroku wybieramy wartości parametrów k i β, a następnie obliczamy wartości lokalnej głębi MBD dla wszystkich obserwacji w zbiorze danych. Następnie wybieramy k centroidów $c_1, \ldots, c_k$ spełniających warunki

$$\sum_{i=1}^{k} LD^{\beta}(c_i, P_n) \to \max$$

$$\sum_{i,j=1}^{k} \|c_i - c_j\|_2 \to \max$$

W drugim kroku tworzymy nowe skupienia w taki sposób, że dla każdej krzywej f obliczamy odległości w normie L_2 od wszystkich centroidów. Następnie na podstawie wartości $clus(f)$ przypisujemy obserwację do najbliższego jej centroidu, tj.

$$clus(f) = \{i : \ d(f, c_i) = \min\{d(f, c_1), d(f, c_2), \ldots, d(f, c_k)\}\},$$

gdzie d oznacza odległość w przestrzeni L_2. Jeśli do zbiorze $clus(f)$ należą co najmniej dwie indeksy, tj. istnieją co najmniej dwa centroidy, do których obserwacja f jest równo odległa, wówczas przypisujemy krzywą funkcjonalną do skupienia o niższej wartości indeksu.

W trzecim kroku dla nowo powstałych skupień obliczamy lokalne mediany MBD względem empirycznych rozkładów poszczególnych skupień. Mediana ta traktowana jest jako nowy centroid. Powtarzamy krok drugi i trzeci, aż centroidy nie zmienią się lub dopóki tylko 1% punktów zmieni skupienie.

Przycięty algorytm k-lokalnych median dla danych funkcjonalnych

Inspiracja algorytm tclust (zob. algorytm tclust, Fritz,Garcia-Escudero, Mayo-Iscar, *J Stat Soft* 2006, Szlachtowska et. al. 2016)
Niech $d(f,g) = \|f-g\|_2$ oznacza odległość L_2 między funkcjami f i g. W każdej iteracji algorytmu, dla każdej funkcji f obliczamy jej odległość do wszystkich centroidów. W rezultacie otrzymujemy ciąg odległości, który możemy posortować rosnąco

$$d(f, c_{(1)}) \leq d(f, c_{(2)}) \leq \cdots \leq d(f, c_{(k)}).$$

Niech $clus(f)$ oznacza indeks skupienia, do którego przypisano krzywą funkcjonalną f. Wówczas

$$clus(f) = \{i : \ d(f, c_i) = d(f, c_{(1)})\}.$$

Użytkownik wybiera parametr γ, tj. wielkość przycinania. Dla każdej obserwacji g obliczamy stopień przynależności do skupienia, do którego została dana obserwacja przypisana. W tym celu obliczamy odległość obserwacji od centroidu danego skupienia. Obliczone wartości ustawiamy malejąco

$$d(f, c_f) \geq d(g, c_g) \geq \cdots \geq d(h, c_h).$$

Następnie odrzucamy γ część obserwacji o najwyższych wartościach stopnia afiliacji, tj. najbardziej odległe od centroidów skupień, do których zostały przypisane.

Ocena jakości grupowania

Dla tej metody podobnie jak w algorytmie tclust, można obliczyć współczynniki dyskryminacyjne dla każdej obserwacji (przyciętej i nieprzyciętej).

Jakość decyzji przypisania do skupienia dla nieprzyciętej obserwacji f do skupienia można ocenić porównując stopień przynależności $d(f, c_f) = d(f, c_{(1)})$ to drugiego najlepszego przypisania $d(f, c_{(2)})$. Stąd

$$DF(f) = \log\left(\frac{d(f, c_{(2)})}{d(f, c_f)}\right) = \log\left(\frac{d(f, c_{(2)})}{d(f, c_{(1)})}\right)$$

dla nieprzyciętej obserwacji f.

Aby obliczyć wartości współczynników dyskryminacyjnych dla krzywych przyciętych ustawiamy obserwacje $f_{(1)}, \ldots, f_{(n)}$ w kolejności malejącej ze względu na ich wartości $d(f_{(i)}, c_{f_{(i)}})$.

Nie jest trudno zauważyć, że $f_{(1)}, \ldots, f_{(\lceil \gamma n \rceil)}$ to przycięte obserwacje, które nie są przypisane do żadnego skupienia.

Niemniej jednak dla przyciętej obserwacji f można obliczyć stopień przynależności $d(f, c_f)$ do najbliższego jej skupienia.

Zatem jakość decyzji o przycięciu tej obserwacji można ocenić porównując $d(f, c_f)$ oraz

$$d\left(f_{(\lceil \gamma n \rceil + 1)}, c_{f_{(\lceil \gamma n \rceil + 1)}}\right),$$

gdzie $f_{(\lceil \gamma n \rceil + 1)}$ oznacza nieprzyciętą obserwację o największej wartości $d(\cdot, c_{(\cdot)})$. Stąd

$$DF(f) = \log\left(\frac{d(f, c_f)}{d\left(f_{(\lceil \gamma n \rceil + 1)}, c_{f_{(\lceil \gamma n \rceil + 1)}}\right)}\right)$$

dla przyciętej obserwacji f.

W rezultacie czynniki dyskryminacyjne $DF(f)$ są uzyskiwane dla każdej obserwacji w zbiorze danych, czy jest to obserwacja przycięta, czy nie. Obserwacje o małych wartości $DF(f)$ (czyli o wartościach zbliżonych do zera) wskazują na wątpliwe przypisanie do skupienia lub na wątpliwą decyzję dotyczące przycinania.

Odporne klasyfikatory w przypadku FDA

Rozumienie odporności klasyfikatora w dalszym ciągu stanowi otwarty problem. Spośród ciekawych propozycji warto zauważyć Cuevas i Romo (1993), którzy badali jakościową odporność boostrapowej aproksymacji estymatora o postaci funkcjonału statystycznego T. Pokazali, że jednostajna ciągłość funkcjonału statystycznego T jest wystarczającym warunkiem na jakościową odporność jego boostrapowej aproksymacji. Niech $L_n(F) = L_n(T; F)$ będzie rozkładem próbkowym statystyki $T_n(X_1, ..., X_n)$ gdzie próba pochodzi z F, oraz $L[L_n(F_n)]$ jest rozkładem próbkowym uogólnionej statystyki $L_n(F_n)$ w przestrzeni odpowiednich miar probabilistycznych. W definicji Cuevasa i Romo (1993) "dla danego ciągu T_n statystyk generowanych przez funkcjonał statystyczny T, ciąg jego bootstrapowych oszacowań $\{L_n(F_n)\}$ jest jakościowo odporny w F wtedy i tylko wtedy gdy, ciąg transformacji $\{G \to L[L_n(G_n)]\}$ jest asymtotycznie równociągły." Definicja ta została wykorzystana m. in. przez Christmanna i in. oraz przez nas (Kosiorowski i in. 2018)

Definicja 3: *Przez* ***jakościową odporność*** rozumiemy równociągłość rozkładu statystyki, gdy zwiększa się wielkość próby.

Zatem jakościowa odporność klasyfikatora wiąże się z ciągłością względem słabej topologii stosownej przestrzeni, pewnej charakterystyki jego rozkładu (por. Rudin, 1991).

Klasyfikatory dla danych funkcjonalnych

- W obrębie FDA stosuje się m. in. metody bazujące na zasadzie k- najbliższych sąsiadów. Obserwacja klasyfikowana jest do klasy obiektów których jest najwięcej w pewnym sąsiedztwie tej. Kluczową i wciąż nierozwiązaną kwestią jest wybór metryki za pomocą której definiuje się sąsiedztwa oraz liczba sąsiadów branych pod uwagę k (Ferraty & Vieu, 2006); metody przestrzeni Hilberta z jądrem reprodukującym (RKHS) (por. Schoelkopf i Smola, 2002); metody bazujące na koncepcji najbliższego centroidu. W ramach tej rodziny metod przyporządkowujemy obserwacji etykietę klasy, której centroid jest jej najbliższy. Jako centroid można przyjąć funkcjonalną średnią bądź pewną funkcjonalną medianę. Kolejna grupa metod wykorzystuje statystyczne funkcje głębi (por. Cuevas i Fraiman, 2009).
- Kosiorowski, Bocian (2015), Kosiorowski, Mielczarek, Rydlewski (2017) klasyfikator SVM via DD plot.
- Kosiorowski, Mielczarek, Rydlewski (2018c) nowy klasyfikator nie nawiązujący do metody reprodukującego jądra przestrzeni Hilberta.

Zalety naszej propozycji

- Przestrzeń L_2 nie jest przestrzenią z jądrem reprodukującym. Najczęściej w przypadku klasyfikatorów dla danych funkcjonalnych mamy do czynienia z tzw. "trickiem jądrowym" tzn. z przekształceniem danych z przestrzeni L_2 na dane z pewnej przestrzeni, najczęściej skończenie wymiarowej przestrzeni Hilberta z jądrem reprodukującym.
- Dobór tej przestrzeni z jądrem reprodukującym ma charakter arbitralny i nie jest znana autorom odporna metoda "dopasowania" jądra do danych funkcjonalnych. Klasyfikacja danych zależy od "właściwego doboru" jądra.
- Nasza metoda jest pozbawiona tych wad, ponieważ zaproponowany klasyfikator nie dokonuje żadnych transformacji danych funkcjonalnych.
- Klasyfikator jest bardzo łatwy w implementacji i stabilny numerycznie.

Niech $X_1, X_2, \ldots, X_m$ będą dowolnymi danymi funkcjonalnymi z przestrzeni Hilberta $L_2(\Omega)$ a liczby $Y_1, Y_2, \ldots Y_m$ będą etykietami tzn.
$Y_i \in \{-1, 1\},\ i \in \{1, \ldots, m\}$.
Obserwacje X_i są więc funkcjami odwzorowującymi zbiór Ω w ciało liczb rzeczywistych i takimi, że

$$\int_{\Omega} |X_i(\omega)|^2 d\omega < \infty,$$

dla dowolnego $i \in \{1, 2, \ldots m\}$.

Nasz klasyfikator dla danych funkcjonalnych będzie postaci

$$g(X) = \int_{\Omega} X(\omega) W(\omega) d\omega + b,$$

gdzie b jest dowolną liczbą rzeczywistą a funkcja wagowa W jest istotnie ograniczona, tzn. $W \in L_{\infty}(\Omega)$ i tak dobrana, aby funkcjonał afiniczny g spełniał warunek zgodności tzn.

$$Y_i g(X_i) = 1,$$

dla dowolnego $i \in \{1, 2, \ldots m\}$.

Istnienie takiego funkcjonału gwarantuje twierdzenie Hahna-Banacha oraz liniowa niezależność próbki $X_1, X_2, \ldots, X_m$. Bowiem z twierdzenia Hahna-Banacha wynika, że każdy liniowy i ograniczony funkcjonał określony na podprzestrzeni pewnej przestrzeni unormowanej, można rozszerzyć z zachowaniem normy do funkcjonału określonego na całej tej przestrzeni.

Zaproponowany przez autorów klasyfikator dla danych funkcjonalnych jest liniową kombinacją funkcjonałów $g_i \colon L_2(\Omega) \to \mathbb{R}$ takich, że

$$g_i(X_j) = \delta_{ij},$$

czyli tzw. funkcjonałów normujących.
Klasyfikator g jest określony wzorem

$$g = \sum_{j=1}^{m} Y_j \left(1 - Y_j b\right) g_j,$$

spełnia on żądane własności.

Okazuję się więc, że wystarczy wskazać funkcjonały g_i. Przyjmijmy następujące oznaczenia dla dowolnego układu wektorów $\{W_1, \ldots, W_m\}$ z przestrzeni Hilberta z iloczynem skalarnym $< \cdot, \cdot >$

$$M(W_1, \ldots, W_m) = \det \left[< W_i, W_j > \right]_{i=1\ldots m,\, j=1\ldots m}.$$

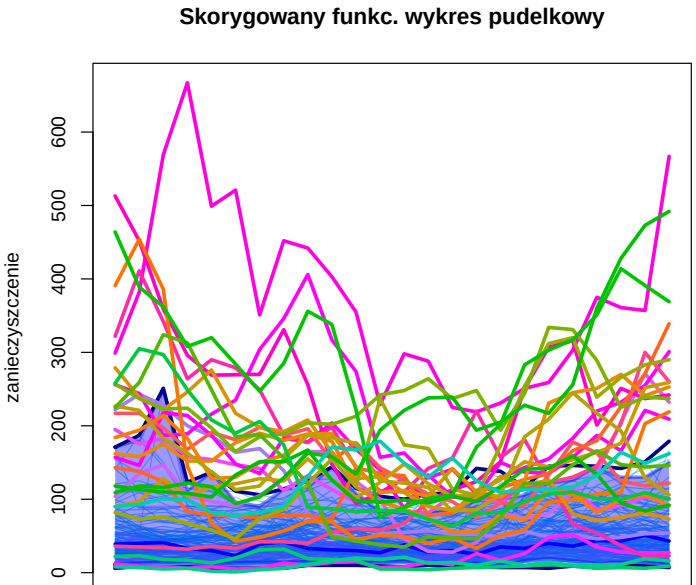
Liczba $M(W_1, \ldots, W_m)$ jest więc wyznacznikiem macierzy Gramma dla układu wektorów $\{W_1, \ldots, W_m\}$.
Przypomnijmy, że układ wektorów jest liniowy niezależny wtedy i tylko wtedy liczba $M(W_1, \ldots, W_m)$ jest większa od zera, jest natomiast równa zero jeżeli układ wektorów $M(W_1, \ldots, W_m)$ jest liniowo zależny.

Przy tak przyjętych oznaczeniach szukane funkcjonały g_i można określić następującym wzorem dla dowolnego $Y \in L_2(\Omega)$

$$g_i(Y) = \frac{M(X_1, \ldots, X_{i-1}, Y, X_{i+1}, \ldots, X_m)}{M(X_1, \ldots, X_m)}.$$

Tak więc wyznaczenie klasyfikatora dla danych funkcjonalnych sprowadza się do wyliczenia odpowiednich wyznaczników. Implementacja klasyfikatora jest więc stosunkowa łatwa.
Przeprowadzone przez autorów badania empiryczne potwierdziły większą skuteczność tego klasyfikatora od klasyfikatorów jądrowych dla danych internetowych oraz dla danych obrazujących zapotrzebowanie na energię elektryczną oraz dla danych pochodzących z pewnego serwisu internetowego podzielonego na podserwisy.
Np. dla serwisu internetowego ryzyko empiryczne naszej metody wyniosło 20%, podczas gdy dla metod RHKS wyniosło 16% (jądro gaussowskie), 18% (jądro wielomianowe) i 24% (jądro sigmoidalne), dla metod bazujacych na głębi (głębia Fraimana-Muniza) aż 46%.
Zaletą naszej metody była tu szybkość - była około 10 razy szybsza.

Przykład 1 "dobowe zanieczyszczeniu powietrza w Katowicach"



Zanieczyszczenie powietrza w Katowicach cd.

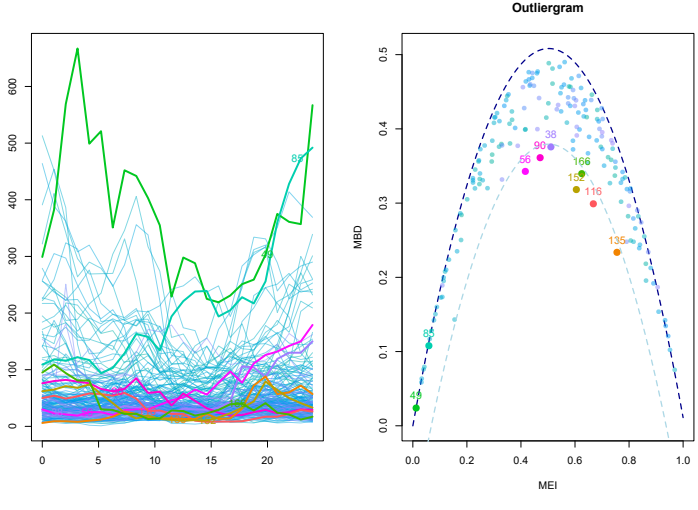


Figure 14: Outlierogram

Dobowe zanieczyszczenie powietrza w Katowicach cd.

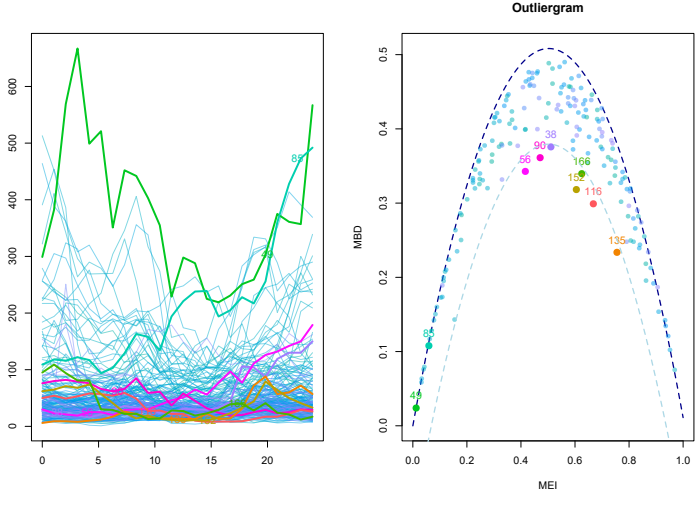


Figure 15: Outlierogram

Zanieczyszczenie powietrza w Katowicach cd.

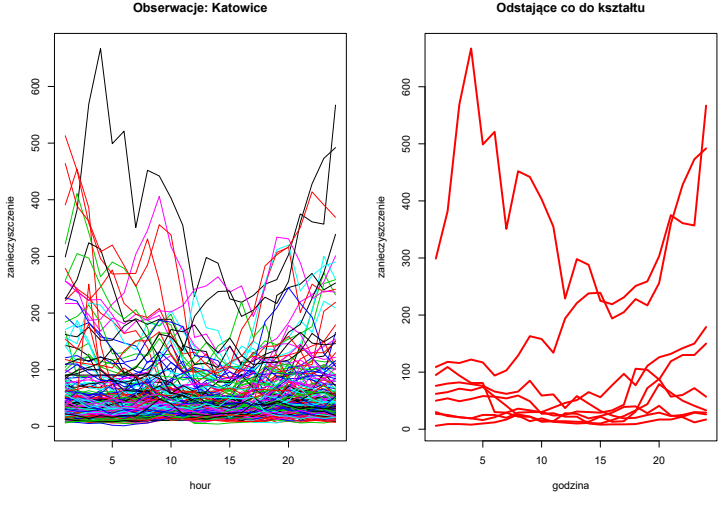


Figure 16: Outliery typu "shape"

Zanieczyszczenie powietrza w Katowicach cd.

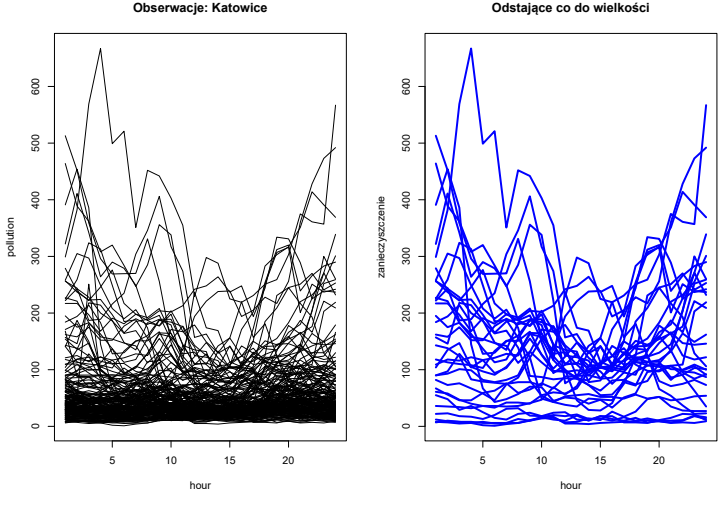
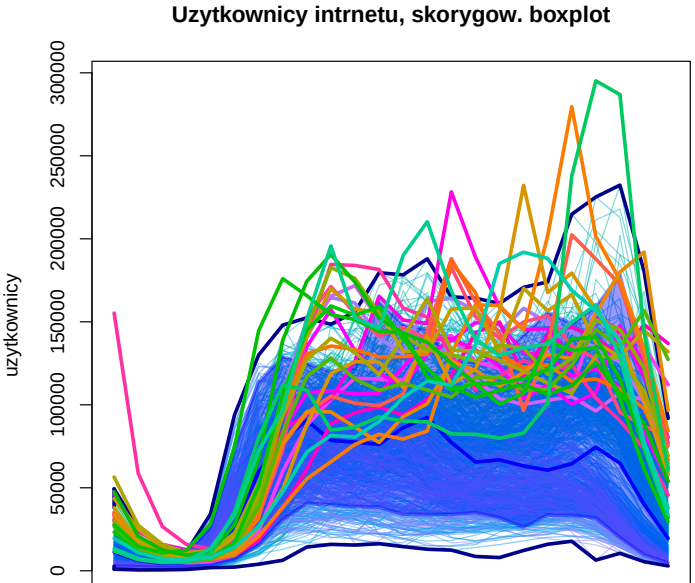


Figure 17: Outliery typu "magnitude

Monitorowanie zachowań internautów



Monitorowanie zachowań internautów

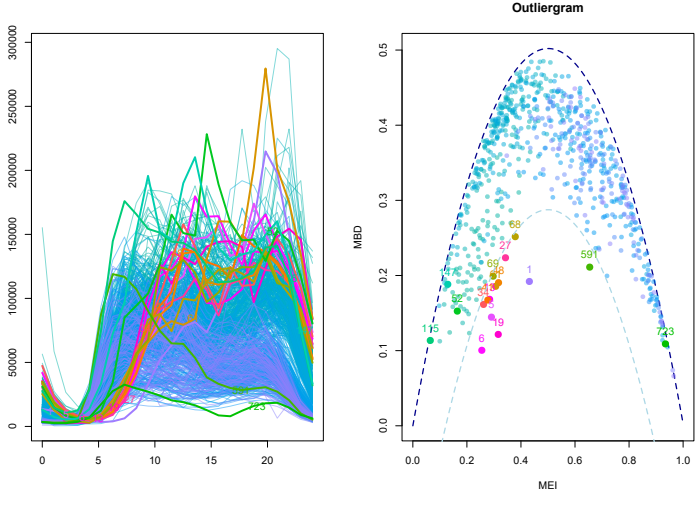


Figure 19: Outlierogram, użytkownicy Internetu

Monitorowanie zachowań internautów

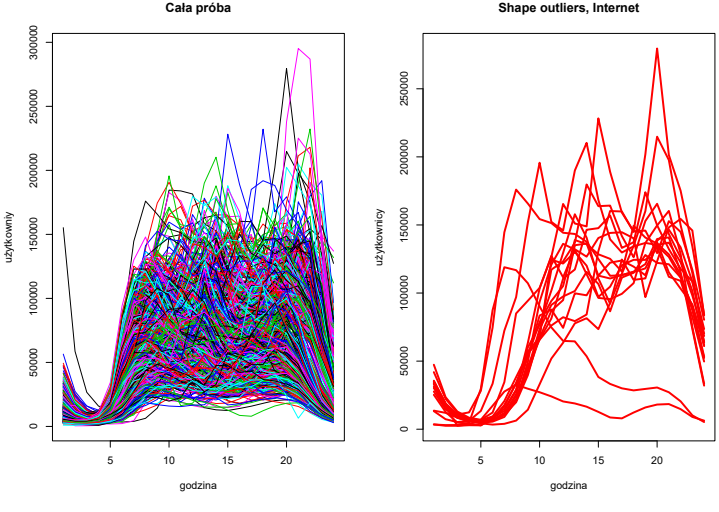


Figure 20: Outliery typu shape, użytkownicy Internetu

Monitorowanie zachowań internautów

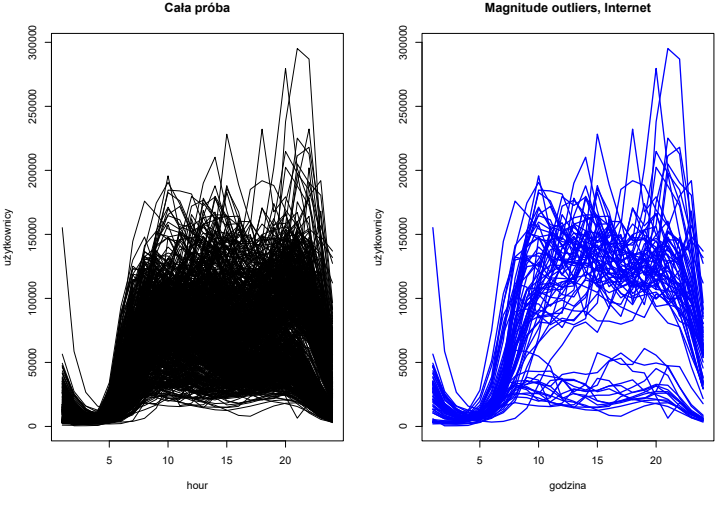


Figure 21: Outliery typu magnitude, użytkownicy Internetu

Predykcja HFTS w maksymalizacji dobrobytu społecznego

Zanieczyszczenie powietrza ma negatywny wpływ na zdrowie człowieka. Substancje niebezpieczne mogą wchodzić ze sobą w interakcje, wpływ może zależeć od grupy wiekowej i pory dnia.

NASZ CEL: **maksymalizacja "zagregowanej" użyteczności pewnej lokalnej społeczności w pewnym okresie czasu**, co symbolicznie zapiszemy jako (por. Fleurbaey i Maniquet, 2011) :

$$U_{Total} = \sum_{i=1}^{365} \int_{[0^{00},24^{00}]} U_i \left(W_{PM10}(t), C_{PM10reduc}(t) \right) dt, \tag{7}$$

gdzie i wskazuje indeks doby, W_{PM10} oznacza społeczny dobrobyt związany z redukcją emisji pyłu PM10 (dodatnie i ujemne efekty zewnętrzne wyrażone w ustalonej walucie) and C_{PM10} oznacza koszt redukcji emisji pyłu PM10 wyrażony w ustalonej walucie.
Zakładamy, że

$$W_{PM10} = F(Air_{qual}, ENV_{polit}, INF_{qual}, POP_{param}),$$

oraz

$$C_{PM10} = G(C_{fixed}, C_{var}, C_{political}, Pred_{qual}).$$

Oznacza to, że dobrobyt związany z pyłem PM10 jest funkcją jakości powietrza (wycenioną na podstawie kosztów hospitalizacji z powodu chorób płuc, wydatków związanych z alergiami (Air_{qual}), przyjazności otoczenia lokalnego (ENV_{polit}), jakości lokalnego systemu informowania o jakości powietrza i zagrożeniach (INF_{qual}) oraz socio-demograficznych parametrów społeczności (POP_{param}).

Predykcja HFTS w maksymalizacji dobrobytu społecznego

Koszty związane z redukcją pyłu PM10 wiążą się z **kosztami stałymi** obejmującymi inwestycje w nowe technologie (C_{fixed}), i **kosztami zmiennymi** związane ze zmianami warunków pogodowych, które skutkują większym bądź mniejszym popytem na energię elektryczną i cieplną (C_{var}), *koszty polityczne* związane z przechodzeniem z popularnych ststemów bazujących na węglu na "czyste" systemy opierającee się np. na energii jądrowej i **koszty związane z jakością predykcji zanieczyszczenia powietrza** ($Pred_{qual}$).

Przykład 4 HFTS w predykcji jakości powietrza na Śląsku

W badaniach empirycznych skupiliśmy się na obszarze Śląska, gdzie zlokalizowanych jest 28 stacji automatycznego pomiaru jakości powietrza nadzorowanych przez Wojewódzki Inspektorat Ochrony Środowiska (WIOŚ, w Katowicach.

Analizowaliśmy dane pochodzące z 5 stacji pomiarowych. Na surowe dane składają się po 181 trajektorii dla każdej z pięciu stacji pokazujące koncentrację pyłu PM10 w atmosferze w $\mu g/m^3$.

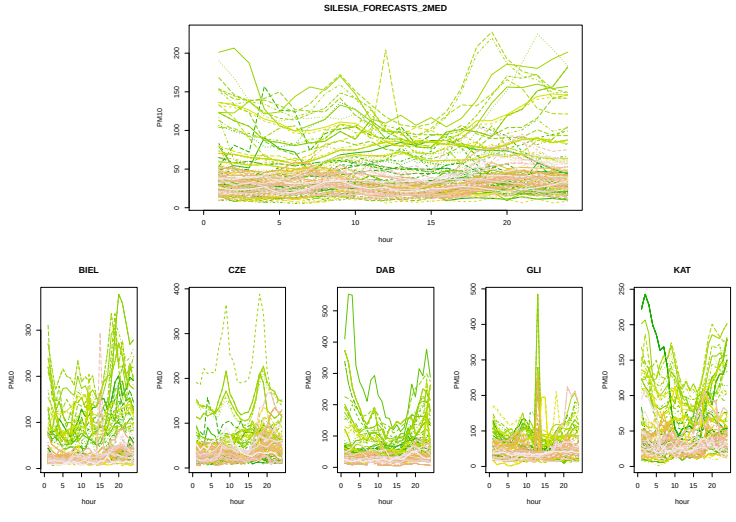


Figure 22: Prognozy koncentracji PM10 uzyskane za pomocą metody podwójnej mediany w $\mu g/m^3$ (długość okna wynosi 10 obserwacji).

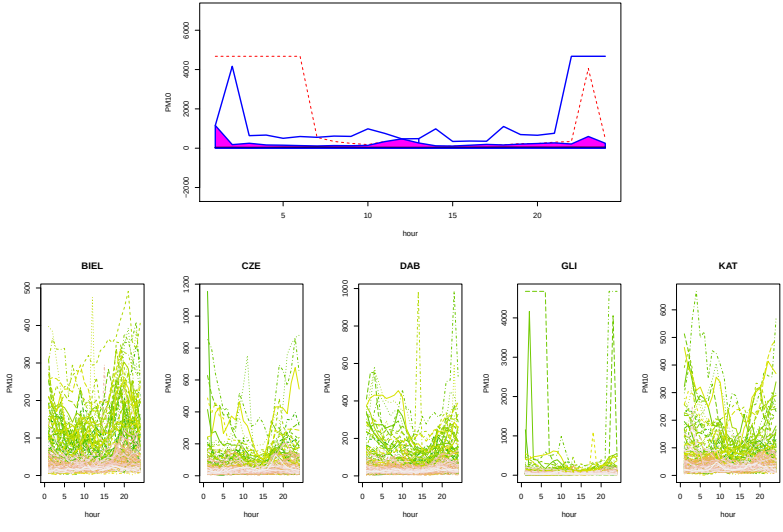


Figure 23: Surowe dane 181 trajektorii koncentracji PM10 dla pięciu stacji w atmosferze w $\mu g/m^3$. Funkcionalny boxplot przedstawia wszystkie krzywe.

Predyktor "naiwny" – ruchoma średnia punktowa

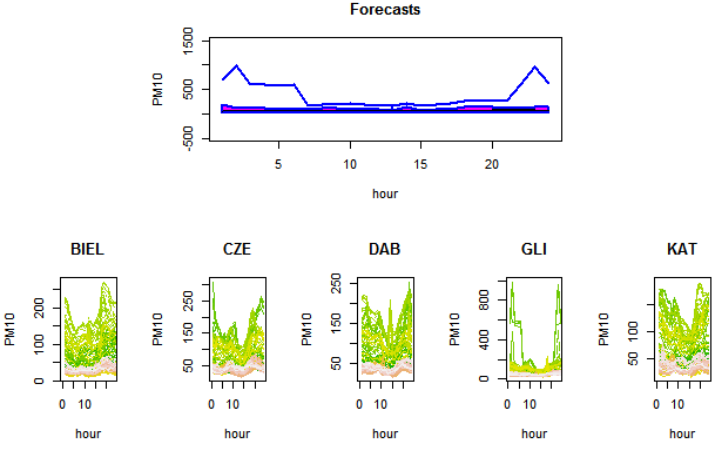


Figure 24: Przewidywania koncentracji PM10 uzyskane za pomocą metody ruchomej średniej funkcjonalnej w $\mu g/m^3$ (długość okna wynosi 10 obs.) dla pięciu stacji.

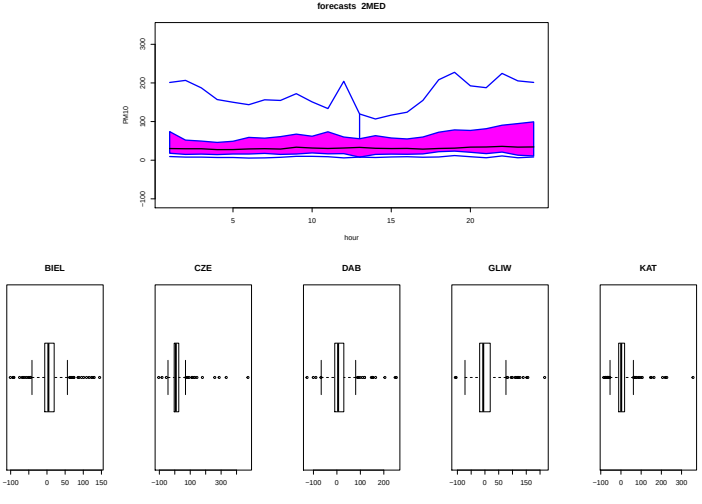


Figure 25: Prognozy koncentracji PM10 uzyskane za pomocą metody podwójnej mediany MBD w $\mu g/m^3$ (długość okna wynosi 10 obs.) dla pięciu stacji. Boxploty przedstawiają uśrednione miary różnic pomiędzy obserwowanymi i przewidywanymi trajektoriami.

Przykład 5 AS dobowego zanieczyszczenia powietrza w Krakowie

Do analizy wybrano również **dane dotyczące jakości powietrza ze stacji Aleja Krasińskiego w Krakowie** za okres od 1 stycznia do 31 grudnia 2015 r. Wyniki pomiarów siedmiu substancji są prezentowane co godzinę na stronie http://monitoring.krakow.pios.gov.pl/.
Należy podkreślić, że w/w okresie nie ma wszystkich pomiarów. Zatem pojawia się **problem brakujących danych**. Dane uzupełniono na dwa sposoby: uzupełniając średnią z danej godziny oraz uzupełniając medianą z danej godziny. W obu wariantach zastosowano algorytm k-średnich dla danych funkcjonalnych w celu zbadania jego odporności na problem brakujących danych.

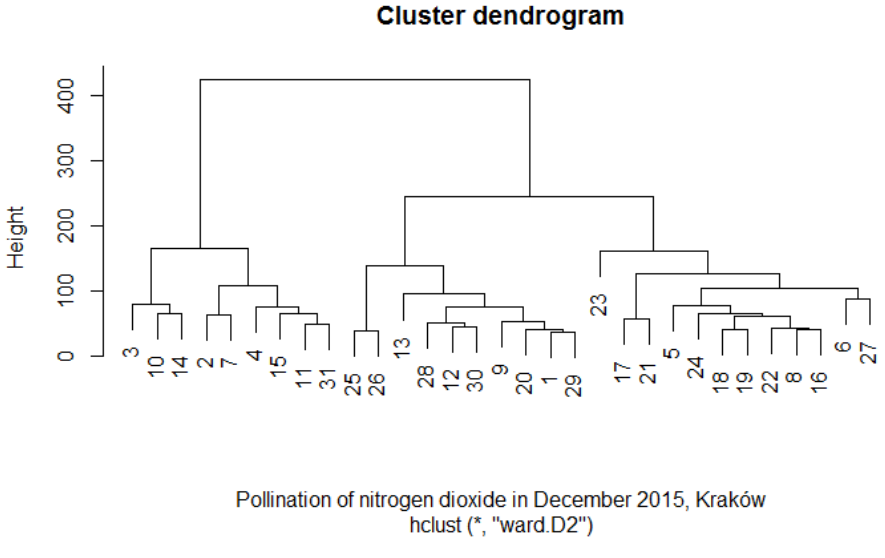
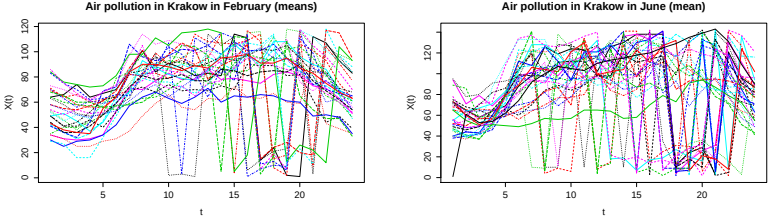


Figure 26: Zanieczyszczenie nitrogen dioxide w grudniu 2015, Kraków.

Tabela przedstawia liczbę obserwacji w skupisku w zależności od zadanej liczby skupisk k k od 2 do 5 (pomiary w grudniu 2015).

k	1	2	3	4	5
2	22	9			
3	10	9	12		
4	10	6	3	12	
5	10	6	3	11	1

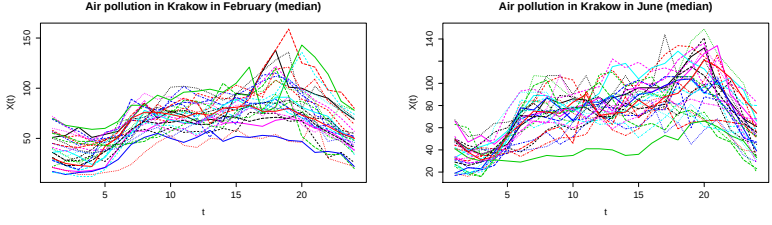
Table 1: Liczba obserwacji w skupisku .Metoda – metoda Warda „Ward.D2 ”



(a) Luty 2015 roku (b) Czerwiec 2015 roku

Figure 27: Krzywe funkcjonalne obrazujące zapylenie dla poszczególnych dni, w których brakujące dane zostały uzupełnione wartością średnią w danej godzinie.

Należy zwrócić uwagę, że przy tym sposobie uzupełnienia brakujących danych krzywe obrazujące zapylenie posiadają liczne skoki wartości, co może negatywnie wpłynąć na wynik grupowania.



(a) Luty 2015 roku (b) Czerwiec 2015 roku

Figure 28: Krzywe funkcjonalne obrazujące zapylenie dla poszczególnych dni, w których brakujące dane zostały uzupełnione medianą dla danej godziny.

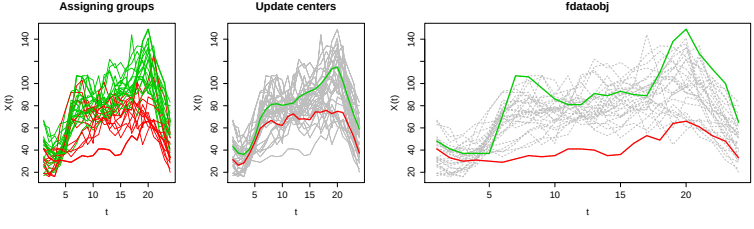


Figure 29: Środki skupień dla krzywych funkcjonalnych obrazujące zapylenie dla poszczególnych dni w czerwcu, w których brakujące dane zostały uzupełnione **medianą wartości zapylenia** w danej godzinie, parametr **k=2**.

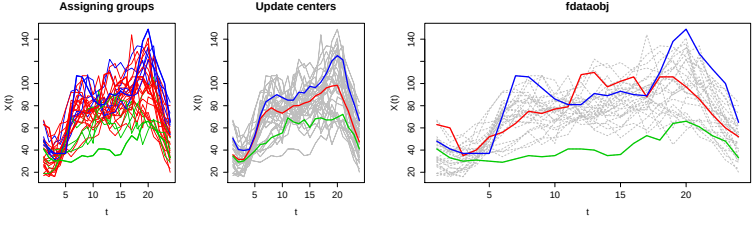


Figure 30: Środki skupień dla krzywych funkcjonalnych obrazujące zapylenie dla poszczególnych dni w czerwcu, w których brakujące dane zostały uzupełnione **medianą wartości zapylenia** w danej godzinie, parametr **k=3**.

Ustawiając $k = 5$ zauważamy, że obserwacja 23 tworzy skupisko. To Wigilia (Świąt Bożego Narodzenia). Można zauważyć największą koncentrację w południe zmniejszającą się po południu. Około godziny 18 następuje dalszy wzrost koncentracji NO2 w powietrzu, jednak istotnie mniejszy niż w godzinach rannych.

Jeżeli podzielimy obserwacje na trzy skupiska, podział przedstawia się:
Pierwsze skupisko: 1, 9, 12, 13, 20, 25, 26, 28, 29, 30 grudnia.
Drugie skupisko: 2, 3, 4, 7, 10, 11, 14, 15, 31 grudnia.
Trzecie skupisko: 5, 6, 8, 16, 17, 18, 19, 21, 22, 23, 24, 27 grudnia.

- Pierwsze skupisko to weekendy, święta i okres tuż po świętach.
- Trzecie skupisko, to dni przed świętami odznaczające się szczególnie wzmożonym ruchem związanym z wyjazdami do rodzin spoza Krakowa.

Konkluzje:

- Najmniejsze stężenie tlenków azotu obserwujemy w weekendy i święta (pierwsze skupisko).
- Największą koncentrację obsrwujemy w dni robocze (drugie skupisko).
- Tuż przed Świętami, w trakcie wyjazdów, zanieczyszczenie pozostaje na średnim poziomie.

Badanie potwierdziło hipotezę, że koncentracja tlenków azotu w atmosferze wiąże się z natężeniem ruchu na krakowskich drogach.

Pyły zawieszone – algorytm *k* lokalnych median

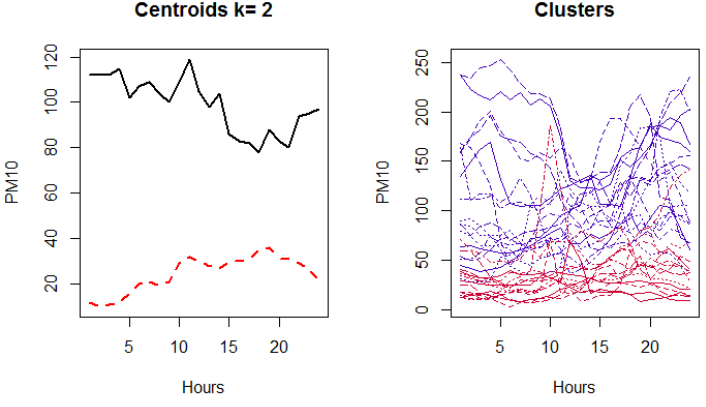


Figure 31: Medians funkcjonalne poszczególnych skupień. Metoda *k* lokalnych median dla danych funkcjonalnych, grudzień 2016 r., pył zawieszony PM10.

Podsumowanie–wyzwania dla zastosowań FDA w ekonomii

- Wyzwanie 1: Sensowne merytoryczne modelowanie zjawisk ekonomicznych za pomocą modeli FDA oraz sensowna merytorycznie interpretacja wyników procedur FDA.
- Wyzwanie 2: Postępowanie w przypadku występowania różnego rodzaju braków w danych funkcjonalnych?
- Wyzwanie 3: Znalezienie teoretycznych podstaw dla metod oceny niepewności prognozy dla FTS - np. teoria dla ruchomej mediany funkcjonalnej.
- Wyzwanie 4: Znalezienie odpowiednika outlierogramu dla innych niż MBD głębi funkcjonalnych.
- Wyzwanie 5: Prognozowanie niestacjonarnych FTS.
- Wyzwanie 6: Wyodrębnienie w klasie "shape outliers" podtypów szczególnie użytecznych w analizie cykli gospodarczych.
- Wyzwanie 7: Koncepcja odporności w przypadku AS i klasyfikacji obiektów.

Wybrane pozycje literatury

- Arribas-Gil A, Romo J (2014) Shape outlier detection and visualization for functional data: the outliergram. Biostatistics, 15(4):603–619.
- Bosq D (2000). Linear processes in function spaces. Springer.
- Górecki, T., Hörmann, S., Horváth, L. & Kokoszka, P. (2017). Testing Normality of Functional Time Series. Journal of Time Series Analysis. doi:10.1111/jtsa.12281.
- Górecki T (2015) „Wprowadzenie do analizy danych funkcjonalnych”, referat plenarny na 7th International Scientific Conference Faculty of Management Cracow University of Economics.
- Górecki T., Krzyśko M.K., Waszak Ł., Wołyński W., (2016), Selected statistical methods of data analysis for multivariate functional data, Statistical Papers, DOI10.1007/s00362-016-0757-8
- Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer, New York
- Kosiorowski D (2012). „Wstęp do statystyki odpornej: kurs z wykorzystaniem środowiska R.” Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.
- Kosiorowski D, Zawadzki Z (2017) DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena. arXiv:1408.4542
- Kosiorowski D, Rydlewski JP , Snarska M (2017a) Detecting a structural change in functional time series using local Wilcoxon statistic. Stat Papers. DOI 10.1007/s00362-017-0891-y

Wybrane pozycje literatury


Kosiorowski D, Mielczarek D, Rydlewski JP, (2018a) Forecasting of a Hierarchical Functional Time Series on Example of Macromodel for the Day and Night Air Pollution in Silesia Region — A Critical Overview. Central European Journal of Economic Modelling and Econometrics 10: 53-73.


Kosiorowski D., Rydlewski J. P., & Zawadzki Z., (2018b). Functional outliers detection by the example of air quality monitoring. Statistical Review (in Polish, to appear).


Kosiorowski D, Mielczarek D, Rydlewski JP, Snarska M, (2018c). Generalized exponential smoothing in prediction of hierarchical time series. Statistics in Transition, June 2018 Vol. 19, No. 2, pp. xxx-xxx. (to appear).


Kosiorowski D, Mielczarek D, Rydlewski JP, (2018d). Outliers in Functional Time Series – Challenges for Theory and Applications of Robust Statistics, In M. Papież & S. Śmiech (eds.), The 12 th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings, Cracow: Foundation of the Cracow University of Economics, pp. 209–218.


López-Pintado S, Romo J (2009) On the concept of depth for functional data. J. Amer. Statist. Assoc. 104: 718-734.


Nagy S, Gijbels I, Hlubinka D (2017) Depth-Based Recognition of Shape Outlying Functions. Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2017.1336445

Wybrane pozycje literatury


Nieto-Reyes A, Battey H (2016) A Topologically Valid Definition of Depth for Functional Data. Statistical Science 31(1): 61-79


Ramsay JO, Hooker G, Graves S (2009) Functional Data Analysis with R and Matlab, Springer


Sun Y, Genton MG (2011) Functional Boxplots. Journal of Computational and Graphical Statistics 20(2): 316-334.


Szlachtowska E. (2018), Odporna analiza skupisk w badaniach nowej gospodarki, Rozprawa doktorska, Uniwersytet Ekonomiczny w Krakowie.


Tarabelloni N. (2017), Robust Statistical Methods in Functional Data Analysis, Doctoral thesis and R package roahd, Politecnico di Milano.