PRZEGLĄD STATYSTYCZNY
TOM LXV – ZESZYT 1 – 2018

Daniel KOSIOROWSKI[1]
Jerzy P. RYDLEWSKI[2]
Zygmunt ZAWADZKI[3]

Wykrywanie funkcjonalnych obserwacji odstających na przykładzie monitorowania jakości powietrza[4]

1. WSTĘP

Istnieje szereg zjawisk i obiektów ekonomicznych, które w naturalny sposób dają opisywać się za pomocą funkcji pewnego kontinuum. Mamy tutaj na uwadze między innymi dobową aktywność uczestników rynku finansowego, trajektorie rozwoju przedsiębiorstw, krzywe użyteczności konsumentów bądź funkcje obrazujące rozkład dochodów gospodarstw domowych w pewnym państwie. W ostatnich latach w literaturze statystycznej zaproponowano nowatorską metodologię statystyczną, która pozwala analizować dane funkcjonalne (patrz Bosq, 2000; Ramsay i inni, 2009; Horváth, Kokoszka, 2012; Górecki i inni, 2014). Wspomnianą metodologię określa się mianem funkcjonalnej analizy danych (ang. *Functional Data Analysis, FDA*). FDA to obszar współczesnej statystyki, w ramach którego bada się obiekty wielowymiarowe, przy czym obiekty te rozpatruje się ze względu na cechę, którą w naturalny sposób można indeksować pewną niezależną zmienną, związaną na przykład z czasem bądź przestrzenią. FDA jest zasadniczo odmienna od jedno- i wielowymiarowej analizy danych statystycznych zarówno z punktu widzenia matematycznego jak i empirycznego. Najważniejsza różnica polega na tym, że na obserwacje patrzymy jak na realizacje funkcjonalnej zmiennej losowej albo jak na trajektorie odpowiednio zdefiniowanego procesu stochastycznego. Za wymienionymi wyżej autorami rozważamy krzywą losową jako funkcję rzeczywistą, której dziedziną jest przedział

[1] Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Statystyki, ul. Rakowicka 27, 31–510 Kraków, Polska, autor prowadzący korespondencję – e-mail: daniel.kosiorowski@ @uek.krakow.pl.

[2] AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie, Wydział Matematyki Stosowanej, Katedra Równań Różniczkowych, al. Mickiewicza 30, 30–059 Kraków, Polska.

[3] Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Statystyki, ul. Rakowicka 27, 31–510 Kraków, Polska (członek zespołu badawczego w Katedrze Statystyki UEK w Krakowie).

[4] Daniel Kosiorowski uprzejmie dziękuje za wsparcie finansowe ze strony UEK w Krakowie w postaci środków na utrzymanie potencjału badawczego przyznanych Wydziałowi Zarządzania w latach 2017 i 2018, Jerzy P. Rydlewski uprzejmie dziękuje za wsparcie finansowe ze strony AGH w Krakowie, dotacja statutowa dla WMS grant numer 11.11.420.004.

[0, T], gdzie T jest znane. Takie funkcje traktujemy jako elementy ośrodkowej przestrzeni Hilberta L^2[0,T] funkcji całkowalnych z kwadratem z naturalnym dla tej przestrzeni iloczynem skalarnym. W monografii Bosqa (2000) dowiedziono, że istnieją rozkłady prawdopodobieństwa dla tak określonych obiektów funkcjonalnych o wartościach w przestrzeni Hilberta.

Funkcjonalne szeregi czasowe są ciągami funkcji indeksowanymi kolejną chwilą, w której czyniona jest obserwacja. Próby obserwacji są realizacjami losowych funkcji, tj. losowych elementów pewnych przestrzeni funkcyjnych, które na ogół są nieskończeniewymiarowymi, rzeczywistymi, ośrodkowymi przestrzeniami Banacha lub Hilberta. Ośrodkowość jest kluczowym założeniem, bowiem zapewnia, że liniowa kombinacja elementów losowych jest takim elementem losowym. Załóżmy, że rozważamy losowe funkcje postaci $X: (\Omega, \mathrm{B}, \mathrm{P}) \to \mathrm{V}$, gdzie $(\Omega, \mathrm{B}, \mathrm{P},)$ jest przestrzenią probabilistyczną oraz V oznacza rzeczywistą i ośrodkową przestrzeń Banacha lub Hilberta wyposażoną w normę $\|\cdot\|$, przy czym w przypadku przestrzeni Hilberta norma indukowana jest przez iloczyn skalarny. Dla wszystkich $\omega \in \Omega$ mamy $X_\omega: t \to X(\omega. t) \in \mathrm{V}$. Rzecz jasna, zwykle dysponujemy danymi dyskretnymi, które następnie przekształcamy do postaci funkcji. W tym celu stosuje się interpolację bądź częściej wygładzanie za pomocą skończonej liniowej kombinacji wybranych funkcji bazowych. Zwykle są to baza Fouriera albo baza złożona ze sklejek funkcji (ang. *B-spline*). W przypadku zastosowań ekonomicznych (nietypowa okresowość zjawiska bądź brak okresowości) rozsądniej jest zastosować bazę złożoną ze sklejek. Baza Fouriera wybierana jest dla funkcji okresowych lub prawie okresowych. Konieczność ograniczenia się do bazy skończonej powoduje redukcję liczby wymiarów i pewne wygładzenie obiektów funkcjonalnych. Wybór liczby funkcji bazowych ma wpływ na wszystkie przeprowadzane dalej obliczenia. Gdy dane są już wyrażone jako funkcje, to kolejnym krokiem jest dokonanie ich transformacji za pomocą analizy składowych głównych (por. Ramsay i inni, 2009; Horváth, Kokoszka, 2012; dla wielowymiarowych danych funkcjonalnych patrz Górecki i inni, 2018).

W ramach FDA rozważa się szereg metod wnioskowania statystycznego. Niektóre z nich stanowią funkcjonalne uogólnienia znanych metod statystyki jedno i wielowymiarowej. Mamy tu na uwadze, przykładowo, funkcjonalną analizę wariancji, funkcjonalny test Wilcoxona bądź funkcjonalne regresje. Podobnie jak ma to miejsce w przypadku jedno- i wielowymiarowego wnioskowania, tak i w przypadku FDA jakość wnioskowania krytycznie zależy od występowania pośród danych tzw. obserwacji odstających. O ile jednak metodykę wykrywania obserwacji odstających w przypadku wielowymiarowym cechuje znaczący stopień zaawansowania, to nie jest tak w przypadku FDA. W przypadku funkcjonalnym wnioskowanie bazuje często na niewielkiej liczbie obserwacji w porównaniu do liczby stopni swobody, ponadto obserwacje funkcjonalne stosunkowo łatwo można zanieczyścić. Podczas wykonywania analizy statystycznej można dosyć

łatwo pomylić obserwacje odstające z błędem. Tymczasem mogą one nieść sporo istotnych informacji o badanym zjawisku. Nieuwzględnienie tej wiedzy może prowadzić do błędnego wyboru modelu, estymacji obciążonej albo wręcz do otrzymania niepoprawnych rezultatów. Dlatego też ważne jest zidentyfikowanie obserwacji odstających przed wykonaniem właściwego modelowania i analizy danych. Trzeba dodać w tym miejscu, że jak na razie, pomimo podejmowania pewnych prób (patrz Hubert i inni, 2015), nie istnieje powszechnie akceptowana definicja funkcjonalnych obserwacji odstających. Niemniej, istnieją dwa podejścia do problemu radzenia sobie z obserwacjami odstającymi. Pierwsze podejście zakłada wykrycie obserwacji odstających, a następnie ich usunięcie ze zbioru analizowanych danych i zastosowanie znanych metod do oczyszczonego zbioru danych. Drugie podejście zakłada stosowanie metod odpornych od samego początku, prowadzenie wnioskowania statystycznego na podstawie naszej wiedzy o ich rozkładach. Zwróćmy uwagę, że w pierwszym podejściu, obserwacje odstające usuwa się ze zbioru danych na podstawie arbitralnej decyzji statystyka, jakiejkolwiek zaawansowanej metody by nie użył.

W celu wykrycia funkcjonalnych obserwacji odstających stosuje się, zaproponowany przez Suna, Gentona (2011), funkcjonalny wykres pudełkowy, który umożliwia zobrazowanie rozkładu danych funkcjonalnych oraz wykrycie krzywych nietypowych dla analizowanego zbioru danych. Jak na razie, pomimo braku formalnych definicji, istnieje zgoda co do tego, by obserwacje odstające dzielić na dwa typy: obserwacje odstające ze względu na amplitudę (ang. *magnitude outliers*) oraz na kształt (ang. *phase outliers, shape outliers*). Odzwierciedla to zmienność funkcji ze względu na różnicę skali oraz położenia. Do wykrycia „shape outliers" używa się zaproponowanego przez Arribas-Gil, Romo (2014) wykresu wartości odstających (ang. outliergram) a następnie usuwa się je z analizowanego zbioru danych. Następnie, za pomocą funkcjonalnego wykresu pudełkowego można wykryć obserwacje odstające ze względu na amplitudę (patrz: Tarabelloni, 2017). Do podstawowych pakietów statystycznych służących analizie danych funkcjonalnych należą pakiety środowiska R *fda* (Ramsay i inni, 2009) i *fda.usc* (patrz Febrero-Bande, de la Fuente, 2012). Wykrywanie funkcjonalnych obserwacji odstających umożliwiają m. in. pakiety R *roahd* (patrz Tarabelloni, 2017) oraz *DepthProc* (patrz Kosiorowski, 2012).

Celem niniejszej pracy jest wykorzystanie przedstawionej wyżej metodyki do wykrycia nietypowych obserwacji funkcjonalnych pokazujących zanieczyszczenie powietrza pyłem PM10 w Katowicach oraz w Krakowie. Umożliwia to decydentom wykrycie odcinków czasu, kiedy poziom zanieczyszczenia powietrza jest szczególnie dokuczliwy. Pozwala to w konsekwencji odpowiednio optymalizować miejską i regionalną politykę w zakresie ochrony środowiska, minimalizować negatywny wpływ zanieczyszczenia powietrza na zdrowie mieszkańców i atrakcyjność turystyczną regionu.

2. GŁĘBIA DLA DANYCH FUNKCJONALNYCH

Jako pierwszy pojęcie głębi zdefiniował Tukey (1975) chcąc umożliwić porządkowanie danych wielowymiarowych, aby następnie umożliwić korzystanie z wypracowanych dla przypadku jednowymiarowego metod statystycznych. Wprowadzona w ten sposób głębia Tukeya (głębia domkniętej półprzestrzeni) znajduje odtąd różnorakie zastosowania. Kolejny rodzaj głębi, głębię symplicjalną, wprowadziła Liu (1990). Od tego czasu w literaturze i w zastosowaniach pojawiło się wiele innych rodzajów głębi (głębia przestrzenna, głębia Mahalanobisa). Wprowadzenie pojęcia głębi umożliwia zdefiniowanie w naturalny sposób kwantyli dla danych wielowymiarowych.

Pojęcie głębi poddano wnikliwym badaniom teoretycznym i empirycznym w pracach Liu, Singh (1993), Liu i inni (1999), Zuo, Serfling (2000a, 2000b) oraz Mosler (2013), Kosiorowski (2012), Kosiorowski (2016), Kosiorowski i inni (2017).

Głębię w przypadku wielowymiarowym można zdefiniować następująco. Rozważmy borelowską σ – algebrę B podzbiorów $\Re^{\rho}$ oraz rodzinę rozkładów prawdopodobieństwa P zdefiniowanych na B. Miara głębi jest funkcją postaci

$$D\colon \Re^{\rho} \times \mathrm{P}\colon (z, P) \to \mathrm{D}\,(\mathrm{z}|P) \in\ [0{,}1]. \tag{1}$$

Dysponując wektorem losowym $X\colon (\Omega, \mathrm{B}) \to \Re^{\rho}$ z indukowanym rozkładem prawdopodobieństwa P_x powiemy, że $D(\cdot\,|P_x)\colon \Re^{\rho} \to [0{,}1]$ jest statystyczną funkcją głębi. Stwierdzimy też, że $D(z|X) \in [0{,}1]$ jest głębią punktu z względem rozkładu prawdopodobieństwa wektora X. Dla próby $X^n = \{x_1, \dots, x_n\}$ z X, zdefiniujemy empiryczną funkcję głębi jako wielkość $D(\cdot\,|X^n) = D\big(\cdot\,\big|\tilde{P}_n\big)$, gdzie P_x zostaje zastąpiony przez rozkład empiryczny $\tilde{P}_n$.

Dla danego wektora losowego X, miary głębi D oraz wartości $\alpha \in [0{,}1]$ definiujemy α – obszar centralny jako $C_{\alpha}(X) = \{z \in \Re^{\rho}\colon D(z|X) \geq \alpha\}$. Umożliwia nam to porządkowanie danych względem pewnego centrum, wyznaczonego przez dane oraz wybraną głębię.

Z czasem pojawiły się definicje głębi dla danych funkcjonalnych (por. np. Fraiman, Muniz, 2001; Cuevas i inni, 2006). Najbardziej chyba powszechną definicję głębi funkcjonalnej, głębię pasma (ang. *band depth, BD*) oraz zmodyfikowaną głębię pasma (ang. *modified band depth, MBD*) wprowadzili López-Pintado, Romo (2007, 2009). Głębia tego rodzaju oparta jest na pomiarze jak często dana funkcja znajduje się w paśmie tworzonym przez inne funkcje pochodzące z tej samej próbki danych funkcyjnych. Jednocześnie López-Pintado, Jörnsten (2007) podały definicje uogólnionej głębi pasma (ang. *generalized band depth, GBD*) oraz skorygowanej uogólnionej głębi pasma (ang. *corrected generalized band depth, cGBD*), które wydają się rozsądnym rozszerzeniem definicji głębi pasma w tym sensie, że pozwalają uwzględniać krzywe

różniące się kształtem czy amplitudą. Pozwala to na wykrywanie obserwacji odstających oraz na prowadzenie bardziej odpornej analizy dla danych funkcjonalnych. Dlatego też w naszych dalszych rozważaniach skupiamy się na tym rodzaju głębi funkcjonalnej.

Podstawowe własności głębi pasma oraz innych głębi funkcjonalnych badali Mosler, Polyakova (2016) oraz Nieto-Reyes, Battey (2016). Najpowszechniejsze zastosowanie głębi wynika z możliwości definiowania opartych na głębiach kwantyli. Dobre statystyczne własności takich obiektów (mierzalność w odpowiednim sensie oraz zgodność) udowodnili Gijbels, Nagy (2015) dla niektórych klas głębi funkcjonalnych niecałkowalnych oraz Nagy i inni (2016) dla klasy głębi funkcjonalnych całkowalnych. Głębia należy do klasy głębi funkcjonalnych (por. Def. 2.3 Nagy i inni, 2016), jeżeli daje się wyrazić jako „zwykła" całka Lebesgue'a ze zwykłej funkcji głębi jednowymiarowej liczonej dla krzywej względem brzegowego rozkładu prawdopodobieństwa. Do klasy głębi funkcjonalnych całkowalnych należy MBD oraz niektóre rodzaje uogólnionych głębi pasma. Które dokładnie, pozostaje wciąż otwartym problemem.

Celem dalszych rozważań podamy za López-Pintado, Romo (2009) definicję MBD. Dla każdej funkcji x pochodzącej z próby funkcyjnej $X^n = \{x_1, \dots, x_n\}$ oraz dla dowolnego j=1,2,…,n niech

$$A_j(x) \equiv A\left(x; x_{i_1}, x_{i_2}, \dots, x_{i_j}\right) \equiv \left\{t \in I: \min_{r=i_1,\dots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\dots,i_j} x_r(t)\right\}. \quad (2)$$

będzie podzbiorem przedziału I (na którym określona jest funkcja x), na którym funkcja x zawiera się w paśmie wyznaczanym przez obserwacje $x_{i_1}, x_{i_2}, \dots, x_{i_j}$. Niech λ oznacza miarę Lebesgue'a, wtedy dla j=2,3,…,n definiujemy wielkość

$$MBD_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < \dots < i_j \leq n} \frac{\lambda\left(A_j(x)\right)}{\lambda(I)} \quad (3)$$

mierzącą "jak często" (czyli przy założeniu, że I jest przedziałem czasu) dana obserwacja x zawiera się w paśmie. Jeżeli ustalimy J=2,3,…,n, to zmodyfikowana głębia pasma funkcji x względem próby X^n równa się

$$MBD_{n,J}(x) = \sum_{j=2}^{J} MBD_n^{(j)}(x). \quad (4)$$

W zastosowaniach przyjmujemy zwykle J=2, czyli rozważamy wyłącznie pasma generowane przez każdą z par funkcji. Teoretyczna wersja MBD ma postać

$$MBD_J(x) = \sum_{j=2}^{J} MBD^{(j)}(x), \tag{5}$$

gdzie $MBD^{(j)}(x) = E\sum_{1\le i_1<\cdots<i_j\le n} \frac{\lambda\left(A_j(x;X_1,X_2,\ldots,X_j)\right)}{\lambda(I)}$.

Warto zwrócić uwagę na fakt, że w przypadku skończenie-wymiarowym wartość $MBD_n^{(j)}(X)$ definiowana jest jako proporcja współrzędnych x zawartych w przedziale tworzonym przez j różnych punktów z próbki (por. López-Pintado, Romo, 2009). W przypadku jedno-wymiarowym BD, która oparta jest o obliczenie proporcji, jak często dana krzywa jest w *całym przedziale I* zawarta w kolejnym paśmie generowanym przez pozostałe obserwacje, to po prostu MBD. Ponadto BD w większym stopniu niż MBD uwzględnia kształt krzywych. Tymczasem MBD uwzględnia w większym stopniu amplitudę lub ilość krzywych niż ich kształt. Warto zanotować, że krzywe pozostające prawie zawsze w centrum i przyjmujące wartości ekstremalne w niewielkich przedziałach będą miały sporą wartość MBD, a niedużą wartość BD. Użycie takiego czy innego rodzaju głębi zależy od rodzaju analizowanych funkcji oraz od celu przeprowadzanego wnioskowania. Gdy krzywe są bardzo nieregularne, zastosowanie MBD umożliwi otrzymanie mniejszej liczby danych powiązanych (ang. *ties*), niemniej nie pozwoli na analizowanie różnego kształtu tych krzywych.

3. WYKRES WARTOŚCI ODSTAJĄCYCH

Wykrywanie obserwacji odstających typu „phase" dla jednowymiarowych danych funkcjonalnych zostało zaproponowane w pracy Arribas-Gil, Romo (2014). Podejście to opiera się na zależnościach pomiędzy zaproponowaną przez López-Pintado, Romo (2011) zmodyfikowaną głębią pasma (ang. *Modified Band Depth, MBD*) oraz zmodyfikowanym indeksem nadwykresu funkcji (ang. *Modified Epigraph Index, MEI*) wprowadzonym przez Martin-Barregana i innych (2015). Przypadek wielowymiarowy rozważali Ieva, Paganoni (2016).

Rozważamy losową funkcję X o wartościach w $\mathrm{V} = C^0(I)$, $I \subset \Re$, wtedy dla elementu $Z \in V$ definiujemy indeks nadwykresu (ang. *ephigraphic index, EI*) oraz indeks podwykresu (ang. *hypographic index, HI*) wzorami:

$$EI(Z|X) = P(\{Z(t) \le X(t), \forall t \in \Re\}) = P(\{G(X) \subseteq epi(Z)\}), \tag{6}$$

$$HI(Z|X) = P(\{Z(t) \geq X(t), \forall t \in \Re\}) = P(\{G(X) \subseteq hyp(Z)\}), \tag{7}$$

gdzie *epi* oznacza nadwykres funkcji, a *hyp* oznacza podwykres funkcji.

Interpretacja obu obiektów jest czytelna, mianowicie EI mierzy prawdopodobieństwo, że obserwacja znajduje się całkowicie powyżej wykresu Z, natomiast HI mierzy prawdopodobieństwo, że obserwacja znajduje się całkowicie poniżej wykresu Z.

Korzystając z analogii do definicji zmodyfikowanej głębi pasma, można zdefiniować zmodyfikowane wersje obu indeksów:

$$MEI(Z|X) = E\frac{\lambda(\{t \in I: Z(t) \leq X(t)\})}{\lambda(I)} = \frac{1}{\lambda(I)}\int_I P\big(Z(t) \leq X(t)\big)dt \tag{8}$$

$$MHI(Z|X) = E\frac{\lambda(\{t \in I: Z(t) \leq X(t)\})}{\lambda(I)} = \frac{1}{\lambda(I)}\int_I P\big(Z(t) \geq X(t)\big)dt \tag{9}$$

gdzie λ oznacza miarę Lebesgue'a.

Wersje próbkowe MEI oraz MHI dla losowej próby $X_1, \ldots, X_N$ przyjmują postać

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

oraz (10)

$$MHI(Z|X^N) = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda(\{t \in I: Z(t) \geq X_i(t)\})}{\lambda(I)}.$$

W pracy Arribas-Gil, Romo (2014) pokazano zaskakująco łatwą do zobrazowania zależność pomiędzy MEI oraz MBD. Relacja ta może zostać wykorzystana do wykrywania obserwacji odstających ze względu na kształt. Prawdziwa jest mianowicie nierówność:

$$MBD(Z|X^N) \leq \alpha_0 + \alpha_1 MEI(Z|X^N) + \alpha_2 N^2 MEI^2(Z|X^N), \tag{11}$$

gdzie $\alpha_0 = \alpha_2 = -2/N\big(N-1\big)$ oraz $\alpha_1 = 2(N+1)/\big(N-1\big)$.

„Nietypowe" obserwacje funkcjonalne ze względu na kształt cechuje mała wartość MBD w porównaniu do parabolicznej zależności w równaniu powyżej, a więc duża różnica pomiędzy lewą i prawą stroną nierówności jest charakterystyczna dla obserwacji odstających. Tymczasem obserwacje „typowe" będą się koncentrowały wokół opisanej powyższym równaniem paraboli. Wykres wartości

odstających pozwala wyrazić graficznie zmienność MBD ze względu na MEI, co czyni możliwą identyfikację obserwacji odstających ze względu na kształt. Ostatnio Ieva, Paganoni (2016) uogólnili wykres wartości odstających dla przypadku wielowymiarowych danych funkcjonalnych. Przypomnijmy, że po zastosowaniu wykresu wartości odstających pozostaną obserwacje odstające ze względu na amplitudę. W celu ich wykrycia stosuje się odpowiedni funkcjonalny wykres pudełkowy.

4. FUNKCJONALNY WYKRES PUDEŁKOWY

Klasyczny funkcjonalny wykres pudełkowy wprowadzili Sun, Genton (2011). Podobnie jak dla przypadku jednowymiarowych danych rzeczywistych, funkcjonalny wykres pudełkowy służy do zobrazowania rozkładów prawdopodobieństwa analizowanych danych i tym samym służy do określenia obserwacji centralnych oraz może posłużyć do wykrycia obserwacji odstających. W celu zdefiniowania danych typowych i nietypowych można posłużyć się pojęciem głębi funkcjonalnej. Rozważamy funkcje losowe, takie jak zdefiniowano we wstępie. Jeżeli mamy próbę losową N funkcjonalnych obserwacji $\{X_1, X_2, \dots, X_N\}$, to na początku porządkujemy je w porządku malejącym, bazując na wybranej głębi funkcjonalnej otrzymując zbiór postaci $\{X_{(1)}, X_{(2)}, \dots, X_{(N)}\}$ W następnym kroku Sun, Genton (2011) definiują próbkowy obszar α-centralny, zawierający α % najbardziej centralnych obserwacji z próby:

$$C_\alpha(X) = \left\{\left(t, z(t)\right): \min_{I=1,2,\dots,[\alpha N]} X_I(t) \le z(t) \le \max_{r=1,2,\dots,[\alpha N]} X_r(t)\right\}. \tag{12}$$

Następnie, podobnie jak w przypadku klasycznego wykresu ramka-wąsy, obliczamy $C_{0.5}$, czyli obszar zawierający 50% najbardziej centralnych funkcji. Potem wybieramy pewien czynnik F≥1 celem powiększenia tego obszaru, by zawierał kolejne obserwacje. Ograniczeniem obszaru będą obwiednie funkcji całkowicie zawierające α % najbardziej centralnych obserwacji z próby (oczywiście teraz $\alpha < 0{,}5$). W ten sposób konstruujemy odpowiednik jednowymiarowego pudełka. Obserwacje, które przekraczają choćby dla jakiegoś podzbioru swojej dziedziny funkcje ograniczające, uznajemy za obserwacje odstające. W przypadku zwykłego jednowymiarowego wykresu ramka-wąsy, zwykle przyjmujemy F=1,5, co oznacza, że frakcja danych uznanych za obserwacje odstające wynosi około 69,8%.

Będziemy dalej rozważali ulepszony funkcjonalny wykres pudełkowy zaproponowany przez Suna, Gentona (2012), który umożliwia rozwiązanie problemu uodpornienia danych funkcjonalnych. W podejściu tym można kontrolować prawdopodobieństwo odrzucenia obserwacji nietypowych dla zbiorów danych mających rozkład gaussowski. Takie dopasowanie parametrów pozwala na traktowanie funkcjonalnego wykresu pudełkowego, nie tylko jako narzędzia służą-

cego do zobrazowania danych, ale także umożliwia wykrycie obserwacji odstających. Takie podejście ma jednak pewne wady, wypunktowane przez Tarabelloniego (2017), który zaproponował udoskonalenie dotychczas stosowanej procedury. Zaproponował on mianowicie połączenie procedury dopasowywania parametrów wraz z wyznaczaniem naturalnego odpornego funkcjonalnego estymatora położenia i skali, co pomoże uzyskać odpowiednie narzędzie służące do uodpornienia danych funkcjonalnych. Przechodząc dalej do szczegółów, w artykule Suna, Gentona (2012) rozważano wybór odpowiedniej wartości F dla przypadku funkcjonalnego wykresu pudełkowego tak, aby tylko z góry wybrana frakcja najbardziej odstających krzywych zostaje uznana za obserwacje odstające albo odrzucona, wszystko to dla przypadku, kiedy dane generowane są przez odpowiedni proces gaussowski. Na tym kończą się analogie z przypadkiem jednowymiarowym, bowiem nie istnieje sposób na otrzymanie teoretycznej wartości F. Ponieważ dane funkcjonalne pochodzące z procesu gaussowskiego są bardziej złożone niż jednowymiarowe zmienne losowe o rozkładzie normalnym, to także procedura wyznaczania wartości F musi uwzględniać empiryczną funkcjonalną wartość oczekiwaną i funkcjonalną kowariancję badanego procesu. Innymi słowy będzie ona zależała od każdego rozpatrywanego zbioru danych funkcjonalnych. Podejście z pracy Suna, Gentona (2012) wymaga estymowania parametrów położenia i rozrzutu ze zbioru danych, by następnie przy ich pomocy utworzyć sztuczną populację funkcji o rozkładzie gaussowskim, która nie zawiera obserwacji odstających tak, aby optymalną wartość F można obliczyć numerycznie poprzez narzucenie odpowiedniego empirycznego rozkładu prawdopodobieństwa. Niemniej trudno te metodę zastosować w praktyce (patrz Tarabelloni, 2017).

Przechodząc do szczegółów, przed przeprowadzeniem symulacji rozkładu gaussowskiego, trzeba odpowiednio uodpornić zbiór danych funkcjonalnych. Sztuczna populacja powinna posiadać taką samą funkcjonalną wartość oczekiwaną i funkcjonalną kowariancje, jak początkowy zbiór danych. Są one jednak przecież nieznane i także trzeba je wyestymować. Zwykle początkowo stosuje się dekompozycję Karhunena-Loèvego (patrz Loève, 1978):

$$X = \mu + \frac{1}{N}\sum_{i=1}^{\infty} \xi_i \sqrt{\lambda_i}\, \psi_i, \tag{13}$$

gdzie ψ_i to wektory własne, a λ_i wartości własne C – operatora kowariancji X, natomiast $\xi_i\, \xi$ to nieskorelowane, rzeczywiste zmienne losowe o zerowej wartości oczekiwanej i wariancji równej jeden. Rzecz jasna, w zastosowaniach ograniczamy się do pierwszych K wartości własnych estymowanych dla rozpatrywanych danych funkcjonalnych i stąd odporny estymator rozrzutu powinien pozwolić na otrzymanie wartości własnych i wektorów własnych operatora kowariancji. Tarabelloni (2017) zauważył, że skorygowany wykres pudełkowy nie gwarantuje, że te estymatory wartości własnych i wektorów własnych operatora kowariancji X są wystarczająco

dobre. Ponadto, są one zależne od rozkładu i mogą nie być operatorami dodatnio określonymi. Są estymatorami zgodnymi operatora kowariancji C tylko dla danych pochodzących z procesu gaussowskiego (patrz Sun, Genton, 2012). W zamian Tarabelloni (2017) proponuje dwojakie podejście do tego zagadnienia. Pierwsze podejście oznacza wyznaczenie empirycznego estymatora sferycznej kowariancji (ang. *spherical covariance estimator*, C_S). Można go interpretować jako rzutowanie X na sferę jednostkową, której środek stanowi mediana przestrzenna. C_S jest wtedy kowariancją danych zrzutowanych na sferę jednostkową. Za pracą Gerviniego (2008) można użyć koncepcji sferycznej kowariancji w odpornej analizie składowych głównych dla danych funkcjonalnych, wykorzystując fakt posiadania przez C i C_S takich samych wektorów własnych oraz korzystając ze wzorów na zależności pomiędzy ich wartościami własnymi. Dlatego też Tarabelloni (2017) uważa, że wyznaczanie C_S stanowi sensowną, bardziej odporną alternatywę dla C.

Drugie podejście Tarabelloniego (2017) bazuje na koncepcji z pracy Krausa, Panaretosa (2012), którzy zaproponowali estymowanie rozrzutu poprzez estymator kowariancji mediany przestrzennej C_M. Przy odpowiednich założeniach C_M ma takie same wektory własne jak C, podobnie też tak jak w przypadku pierwszego podejścia, wartości własne C i C_M są powiązane skomplikowanymi zależnościami. Niemniej, ze względu na obiecujące własności spektralne i ze względu na odporność, zarówno sferyczna kowariancja jak i kowariancja mediany są używane do otrzymania skorygowanego funkcjonalnego wykresu pudełkowego. Używa się tych wielkości, tzn. C_S i C_M, by znaleźć wartości i wektory własne operatora C i by następnie utworzyć populację o rozkładzie gaussowskim służącą do wyznaczenia wartości F. W praktyce obliczenia wartości własnych sprowadzają się do rozwiązania wielu jednowymiarowych problemów estymowania parametrów skali. Do tego celu użyć można mediany wartości bezwzględnej odchyleń od mediany (ang. *Median Absolute Deviation, MAD*) albo wprowadzonych w pracy Rousseeuw, Croux (1993) estymatorów S_N i Q_N. Zdefiniowane są one następująco:

$$MAD(x_1, x_2, \dots, x_N) = b \cdot med\{|x_i - med(x_1, x_2, \dots, x_N)|\}, \tag{14}$$

$$S_N = c \cdot med_{i=1,\dots,N}\{med_{j=1,\dots,N}|x_i - x_j|\}, \tag{15}$$

gdzie stałe parametry b>0 albo c>0 wybiera się tak, by uzyskać zgodny w sensie Fishera estymator rozrzutu. Innym, odpornym estymatorem rozrzutu, jest

$$Q_N = d \cdot \{|x_i - x_j|, i < j\}_{(k)}, \text{ gdzie } k = \binom{\left[\frac{N}{2}\right] + 1}{2}, \tag{16}$$

równe w przybliżeniu kwantylowi rzędu 0,25 różnic par obserwacji, pomnożonych przez pewną stałą d>0, wybraną tak, by uzyskać zgodny w sensie Fishera estymator rozrzutu.

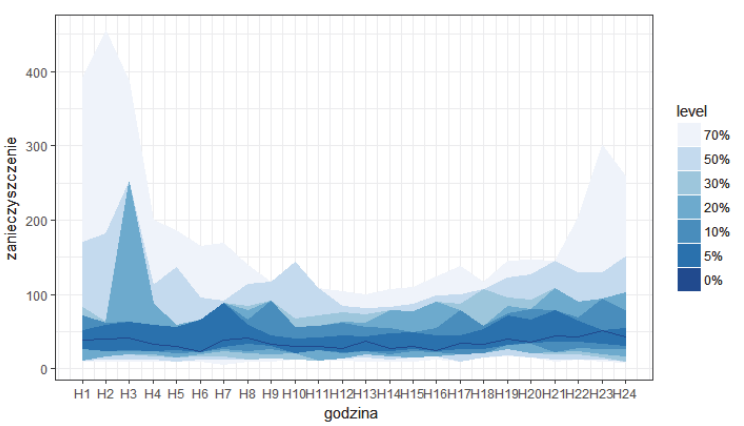
Rousseeuw, Croux (1993) doradzali zastosowanie Q_N jako, że jego punkt załamania próby skończonej wynosi 50% i ma asymptotyczną efektywność równą około 88,27% dla rozkładów gaussowskich. Wybranie w procedurze obliczania *F*, błędnej wartości parametru *d* skutkuje wyborem błędnego operatora kowariancji i nieprawidłowym wykresem pudełkowym. Wyjściem z sytuacji jest wybór, w celu utworzenia rang obserwacji, głębi funkcjonalnej niezmienniczej ze względu na parametr przesunięcia i skali. Wybrana przez nas MBD spełnia ten warunek (por. np. Nagy i inni, 2016). Ostatecznie Tarabelloni (2017) używa modelu

$$Y_i^* = \frac{Y_i - \mu_x}{\sqrt{\lambda_1}}\mu = \sum_{j=1}^{\infty}\sqrt{\frac{\lambda_j}{\lambda_1}}\xi_{i,j}\psi, \tag{17}$$

gdzie $\xi_{i,j}$ to niezależne zmienne losowe o standardowym rozkładzie normalnym. Zamiast średniej μ_X rozważa przestrzenną medianę X jako parametr położenia. W taki sposób otrzymujemy wartość F. Ponadto iloraz $\sqrt{\frac{\lambda_j}{\lambda_1}}$ można w prosty sposób estymować korzystając z własności Q_N. W rezultacie Tarabelloni (2017) otrzymuje estymatory wektorów własnych i wartości własnych macierzy kowariancji, które następnie służą do wygenerowania sztucznej gaussowskiej populacji, która z kolei posłuży do wyznaczenia spójnej i niezależnej od rozkładu wartości F. Zastosowanie sferycznej kowariancji bądź kowariancji mediany wraz z Q_N powoduje, że estymator jest bardziej odporny.

Na rysunku 1 przedstawiamy przykładowy funkcjonalny wykres pudełkowy utworzony z wykorzystaniem zmodyfikowanej głębi pasma (MBD). Pokazuje on zanieczyszczenie powietrza pyłem PM10 w Katowicach w analizowanym przez nas okresie czasu, tj. od 1 września 2016 r. do 28 lutego 2017 r.

Rysunek 1. Funkcjonalny wykres pudełkowy utworzony z wykorzystaniem zmodyfikowanej głębi pasma (MBD) obrazujący zanieczyszczenie powietrza pyłem PM10 w Katowicach w μg/m^3 w analizowanym okresie



Źródło: opracowanie własne. Użyto pakietu *DepthProc*.

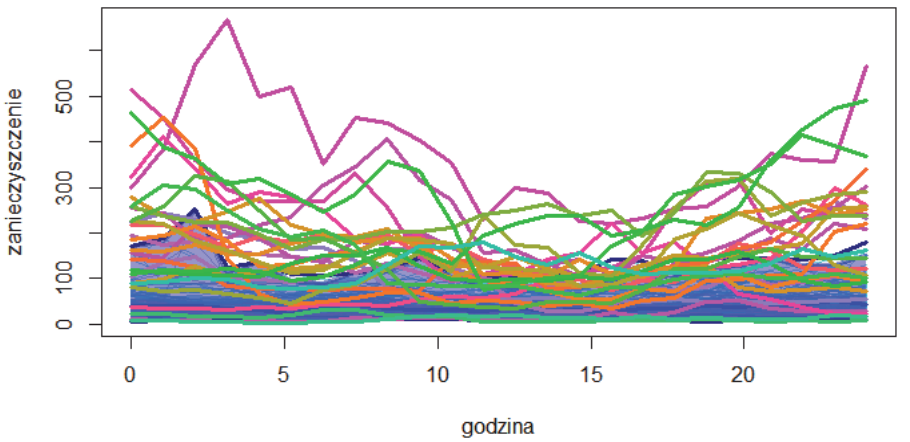
5. MONITOROWANIE JAKOŚCI POWIETRZA W KATOWICACH I KRAKOWIE

Na zanieczyszczenie powietrza składają się emisje różnego rodzaju związków, są to min. dwutlenek siarki, dwutlenek azotu, ozon, tlenek węgla, benzen, tlenki azotu, pył zawieszony PM2,5 oraz pył zawieszony PM10 – czyli wszystkie cząstki o wielkości 10 mikrometrów lub mniejszej. Smog ma ujemy wpływ na zdrowie mieszkańców, innymi słowy generuje koszty społeczne i zdrowotne. W celu monitorowania zanieczyszczenia powietrza dokonuje się pomiarów stężenia pewnych związków chemicznych w powietrzu. Pomiarów dokonuje się wyłącznie w pewnej liczbie stacji rozmieszczonych, zwykle nierównomiernie, na pewnym obszarze. W analizowanym przez nas przypadku są to województwo śląskie i małopolskie, gdzie Wojewódzkie Inspektoraty Ochrony Środowiska w Katowicach i Krakowie (WIOŚ) umieszczają na stronach internetowych http://powietrze.katowice.wios.gov.pl oraz http://www.krakow.pios.gov.pl dane ze stacji pomiarowych z terenu swojego województwa. Prezentowane są tam wyniki automatycznych pomiarów jakości powietrza, są one przekazywane bezpośrednio ze stacji pomiarowych i nie są zweryfikowane. Zaznaczmy za WIOŚ, że wyniki te są poddawane okresowej weryfikacji i mogą ulec zmianie. Analizowaliśmy stężenie w powietrzu pyłu zawieszonego PM10 w Katowicach (stacja przy ul. Kossutha). Dane pochodzą z okresu 181 dni od 1 września 2016 r. do 28 lutego 2017 r. Analizujemy 181 krzywych pokazujących, jak zmienia się stężenie w powietrzu pyłu zawieszonego PM10. Analizowaliśmy też dane pokazujące zanieczyszczenie powietrza w Krakowa z grudnia 2016 r. czterema rodzajami substancji: tlenkiem azotu (NO), tlenkami azotu (NOx), pyłem zawieszonym PM 2,5 oraz pyłem zawieszonym PM10.

Patrzymy na funkcje losowe obserwowane w regularnych odstępach czasu jak na funkcjonalne szeregi czasowe. Można też konstruować funkcjonalne szeregi czasowe poprzez rozdzielenie ciągłego przedziału czasowego na naturalne części tzn. godziny, dni, tygodnie, miesiące, lata. Na kolejnych rysunkach znajdują się funkcjonalne obserwacje pokazujące zanieczyszczenie powietrza pyłem PM10 w Katowicach. Rysunek 2 przedstawia skorygowany funkcjonalny wykres pudełkowy pokazujący zanieczyszczenie powietrza pyłem PM10 w Katowicach w $\mu g/m^3$. Wyróżniono obserwacje odstające ze względu na amplitudę. Rysunek 3 przedstawia po lewej obserwacje funkcjonalne obrazujące zanieczyszczenie powietrza pyłem PM10 w Katowicach w $\mu g/m^3$. Po prawej wykres wartości odstających, na którym wyróżniono obserwacje funkcjonalne odstające ze względu na kształt, ponadto wskazano numery takich obserwacji. Funkcjonalne obserwacje odstające ze względu na kształt zostały zestawione na rysunku 4, gdzie prezentujemy je na tle wszystkich obserwacji. Funkcjonalny wykres pudełkowy umożliwił nam wykrycie obserwacji odstających ze względu na amplitudę (wielkość). Zostały one zestawione na rysunku 5, gdzie prezentujemy je na tle wszystkich obserwacji pokazujących zanieczyszczenie powietrza pyłem PM10 w Katowicach. Rysunek 6 przedstawia zanieczyszczenie powietrza

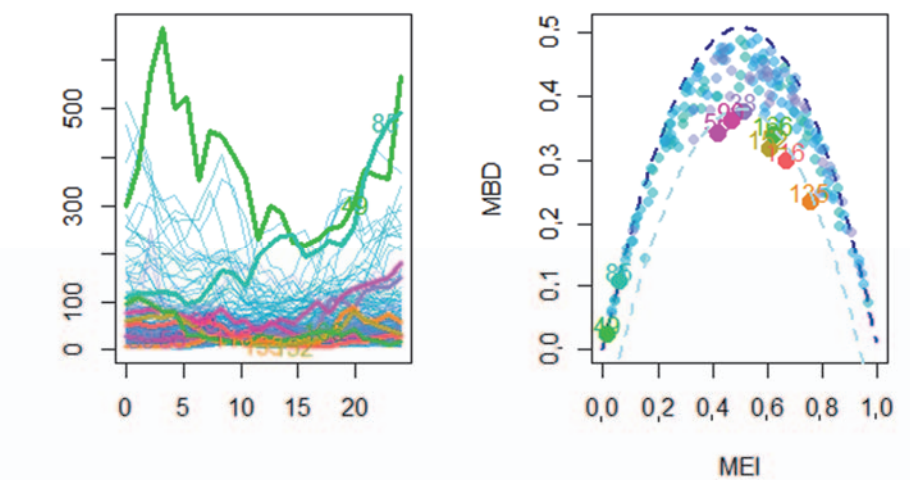
w Krakowie czterema wybranymi rodzajami substancji w μg/m^3. Natomiast rysunek 7 przedstawia funkcjonalne wykresy pudełkowe utworzone z wykorzystaniem zmodyfikowanej głębi pasma (MBD) obrazujące zanieczyszczenie w Krakowie w μg/m^3 w grudniu 2016 r. czterema substancjami (NO, NOx, PM2,5 i PM10).

Rysunek 2. Skorygowany funkcjonalny wykres pudełkowy pokazujący zanieczyszczenie powietrza pyłem PM10 w Katowicach w μg/m^3



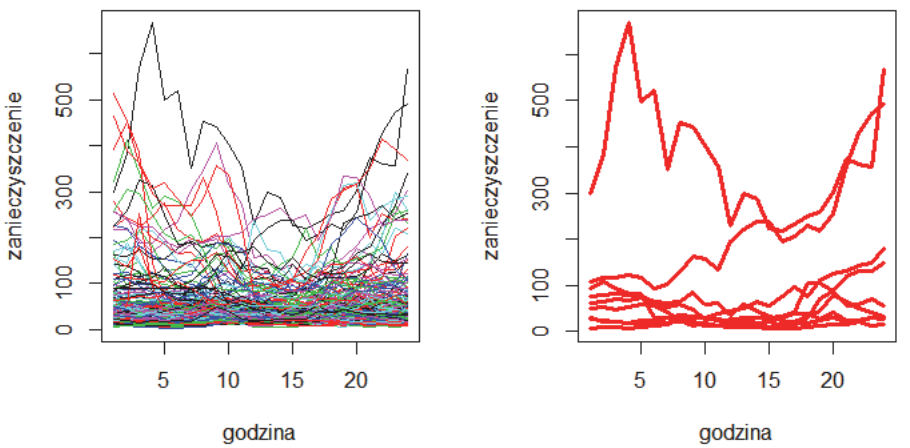
Wyróżniono obserwacje odstające ze względu na amplitudę.
Źródło: opracowanie własne. Użyto pakietu *DepthProc*

Rysunek 3. Po lewej obserwacje funkcjonalne obrazujące dobowe zanieczyszczenie powietrza pyłem PM10 w Katowicach w μg/m^3. Po prawej wykres wartości odstających, na którym wyróżniono obserwacje odstające ze względu na kształt



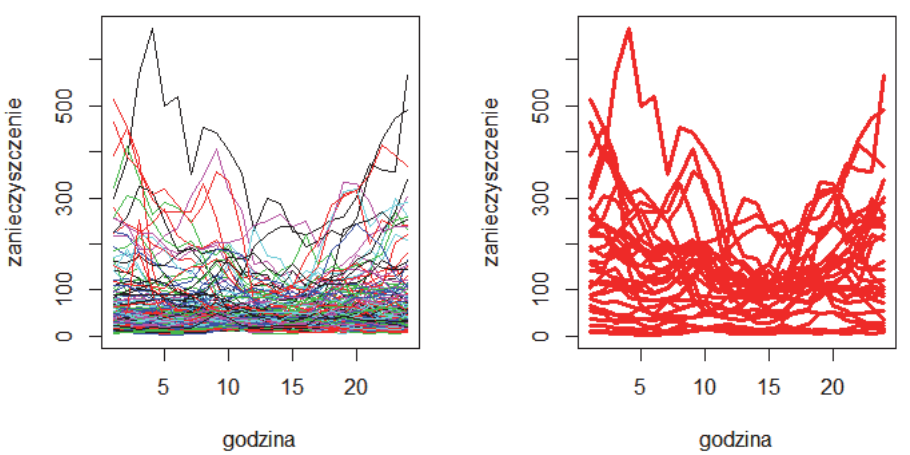
Wskazano numery obserwacji odstających ze względu na kształt.
Źródło: opracowanie własne. Użyto pakietu *roahd*.

Rysunek 4. Po lewej obserwacje funkcjonalne obrazujące zanieczyszczenie powietrza pyłem PM10 w Katowicach w μg/m^3. Po prawej obserwacje funkcjonalne odstające co do kształtu



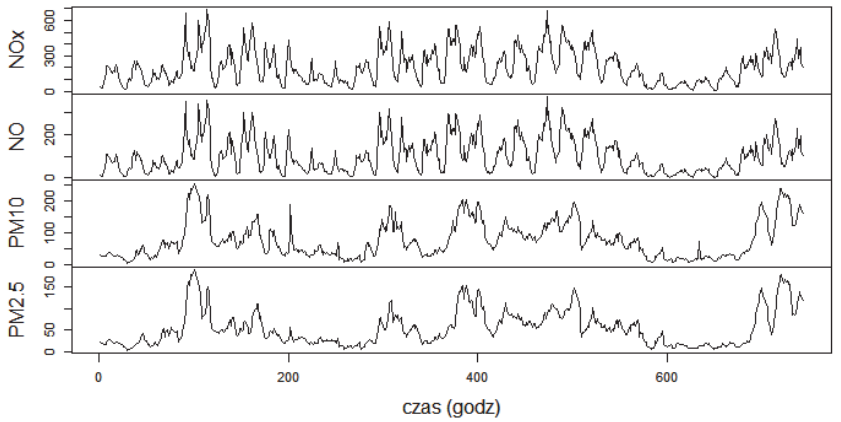
Źródło: opracowanie własne. Użyto pakietu *DepthProc*.

Rysunek 5. Po lewej obserwacje funkcjonalne obrazujące zanieczyszczenie powietrza pyłem PM10 w Katowicach w μg/m^3. Po prawej obserwacje funkcjonalne odstające co do amplitudy (wielkości)



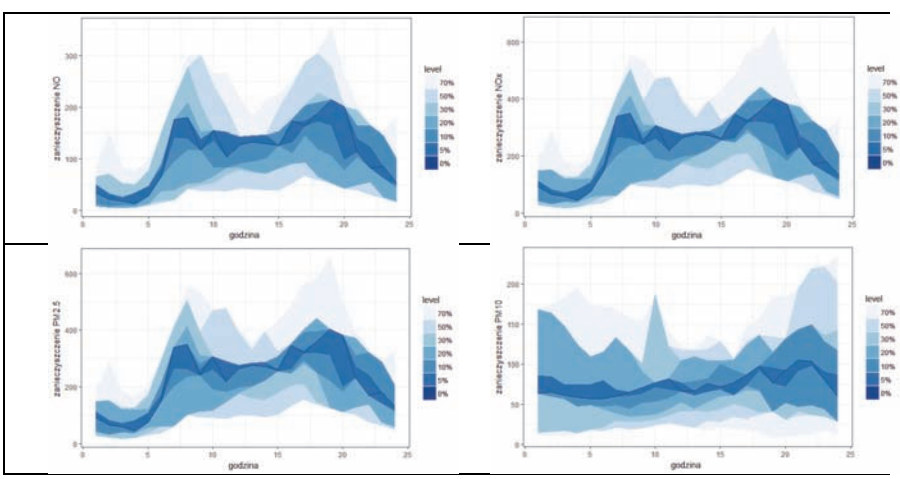
Źródło: opracowanie własne. Użyto pakietu *DepthProc*.

Rysunek 6. Zanieczyszczenie powietrza w Krakowie czterema wybranymi substancjami w $\mu g/m^3$ w dniach od 1.12.2016 do 31.12.2016 r."



Źródło: opracowanie własne. Użyto pakietu *DepthProc*.

Rysunek 7. Funkcjonalne wykresy pudełkowe utworzone z wykorzystaniem zmodyfikowanej głębi pasma (MBD) obrazujące zanieczyszczenie powietrza tlenkiem azotu (NO), tlenkami azotu (NOx), pyłem PM2,5 i pyłem PM10 w Krakowie w $\mu g/m^3$ w grudniu 2016 r.



Ź r ó d ł o: opracowanie własne. Użyto pakietu *DepthProc*.

Podsumowując, w pracy zaprezentowano metodykę wykrywania funkcjonalnych obserwacji odstających ze względu na kształt oraz na amplitudę oraz przedstawiono jej użyteczność empiryczną na przykładach dotyczących monitorowania zanieczyszczenia powietrza w Katowicach i Krakowie. Należy podkreślić, że znane metody wnioskowania statystycznego stosowane w obrębie FDA na ogół krytycznie zależą od występowania pośród danych obserwacji odstają-

cych. Zaprezentowane metody wraz z ich implementacjami w darmowych pakietach środowiska R „uodparniają" metody wnioskowania i co za tym idzie prowadzą do bardziej miarodajnych wniosków merytorycznych. Taka wiedza merytoryczna umożliwia decydentom optymalizację polityki miejskiej i regionalnej w zakresie ochrony powietrza. Wykorzystywane w pracy metody i zbiory danych dostępne są w pakiecie *DepthProc.*

LITERATURA

Arribas-Gil A., Romo J., (2014), Shape Outlier Detection and Visualization for Functional Data: the Outliergram, *Biostatistics*, 15 (4), 603–619.

Cuevas A., Febrero M., Fraiman R., (2006), On the Use of the Bootstrap for Estimating Functions with Functional Data, *Computational Statistics & Data Analysis*, 51 (2), 1063–1074.

Febrero-Bande M. O., de la Fuente M., (2012), Statistical Computing in Functional Data Analysis: The R Package fda.usc, *Journal of Statistical Software*, 51 (4), 1–28.

Fraiman R., Muniz G., (2001), Trimmed Means for Functional Data, *Test*, 10 (2), 419–440.

Gervini D., (2008), Robust Functional Estimation Using the Median and Spherical Principal Components, *Biometrika*, 95 (3), 587-600.

Gijbels I., Nagy S., (2015), Consistency of Non-Integrated Depths for Functional Data, *Journal of Multivariate Analysis,* 140, 259–282.

Górecki T., Krzyśko M., Waszak Ł., Wołyński W., (2014), Methods of Reducing Dimension for Functional Data, *Statistics in Transition,* 15 (2), 231–242.

Górecki T., Krzyśko M., Waszak Ł., Wołyński W., (2018), Selected Statistical Methods of Data Analysis for Multivariate Functional Data, *Statistical Papers,*59 (1), 153–182.

Horváth L., Kokoszka P., (2012), *Inference for Functional Data with Applications*, Springer-Verlag, New York.

Hubert M., Rousseeuw P., Segaert, P., (2015), Multivariate Functional Outlier Detection, *Statistical Methods and Applications,* 24 (2), 177–202.

Ieva F., Paganoni A. M., (2016), A Taxonomy of Outlier Detection Methods for Robust Classification in Multivariate Functional Data. Technical Report 15/2016, MOX – Modeling and Scientific Computing Laboratory.

Kosiorowski D., (2012), *Statystyczne Funkcje Głębi w Odpornej Analizie Ekonomicznej*, Wydawnictwo UEK w Krakowie, Kraków.

Kosiorowski D., (2016), Dilemmas of robust analysis of economic data streams, *Journal of Mathematical Sciences* (Springer), 218 (2), 167–181.

Kosiorowski D., Rydlewski, J. P., Snarska M., (2017) Detecting a Structural Change in Functional Time Series Using Local Wilcoxon Statistic, *Statistical Papers*, DOI 10.1007/s00362-017-0891-y.

Kosiorowski D., Zawadzki, Z. (2014) DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena, arXiv preprint arXiv:1408.4542.

Kraus D., Panaretos V. M., (2012), Dispersion Operators and Resistant Second-Order Functional Data Analysis, *Biometrika*, 99 (4), 813–832.

Liu R. Y, (1990), On a Notion of Data Depth Based on Random Simplices, *The Annals of Statistics*, 18 (1), 405–414.

Liu R. Y., Parelius J., Singh K., (1999), Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. *The Annals of Statistics,* 27 (3), 783–858.

Liu R. Y., Singh K., (1993), A Quality Index Based on Data Depth and Multivariate Rank Tests, *Journal of the American Statistical Association,* 88 (421), 252–260.

Loève M., (1978), *Probability Theory.* Springer-Verlag, New York.

López-Pintado S., Jörnsten R., (2007), Functional Analysis via Extensions of the Band Depth, w: Liu R., Strawderman W., Zhang C. H., (red.), *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, 54,103–120, Institute of Mathematical Statistics, IMS Lecture Notes – Monograph Series.

López-Pintado S., Romo J., (2007), Depth-Based Inference for Functional Data, *Computational Statistics & Data Analysis*, 51 (10), 4957–4968.

López-Pintado S., Romo J., (2009), On the Concept of Depth for Functional Data, *Journal of the American Statistical Association*, 104 (486), 718–734.

Martin-Barragan B., Lillo R. E., Romo J., (2015)., Functional Boxplots Based on Epigraphs and Hypographs, *Journal of Applied Statistics*, 43 (6), 1088–1103.

Mosler K., (2013), Depth Statistics, w: Becker C., Fried R., Kuhnt S., (red.), *Robustness and Complex Data Structures,* Springer-Verlag Berlin Heidelberg, 17–34.

Mosler K., Polyakova Y., (2016), General Notions of Depth for Functional Data, *arXiv:* 1208.1981v2.

Nagy S., Gijbels I., Omelka M., Hlubinka D., (2016), Integrated Depth for Functional Data: Statistical Properties and Consistency, *ESIAM Probability and Statistics,* 20, 95–130.

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

Ramsay J. O., Hooker G., Graves S., (2009), *Functional Data Analysis with R and Matlab*, Springer – Verlag, New York.

Rousseeuw P. J., Croux C., (1993), Alternatives to the Median Absolute Deviation, *Journal of the American Statistical Association*, 88 (424), 1273–1283.

Sun Y., Genton M., (2011), Functional Boxplots, *Journal of Computational and Graphical Statistics*, 20 (2), 316–334.

Sun Y., Genton M., (2012), Adjusted Functional Boxplots for Spatio-Temporal Data Visualization and Outlier Detection, *Environmetrics*, 23 (1), 53–64.

Szlachtowska E., (2017), Odporna analiza skupisk w badaniach nowej ekonomii, Rozprawa doktorska, Uniwersytet Ekonomiczny w Krakowie.

Tarabelloni N., (2017), *Robust Statistical Methods in Functional Data Analysis,* Rozprawa doktorska, Politecnico di Milano.

Tukey J., (1975), Mathematics and the Picturing of Data, *Proceedings of the International Congress of Mathematicians, Vancouver*, 2, 523–531.

Zuo Y., Serfling R., (2000a), General Notions of Statistical Depth Function, *The Annals of Statistics*, 28 (2), 461–482.

Zuo Y., Serfling R., (2000b), Structural Properties and Convergence Results for Contours of Sample Statistical Depth Functions, *The Annals of Statistics*, 28 (2), 483–499.

Źródła internetowe:
http://powietrze.katowice.wios.gov.pl (data dostępu: 24 marca 2017 r.)
http://www.krakow.pios.gov.pl (data dostępu: 25 marca 2017 r.)

WYKRYWANIE FUNKCJONALNYCH OBSERWACJI ODSTAJĄCYCH NA PRZYKŁADZIE MONITOROWANIA JAKOŚCI POWIETRZA

Streszczenie

W pracy omówiono sposoby wykrywania obserwacji odstających w zbiorach danych funkcjonalnych. Omówiono mianowicie funkcjonalne obserwacje odstające ze względu na kształt i ze względu na amplitudę. Zdefiniowano wykres wartości odstających, służący do wykrywania funkcjonalnych obserwacji odstających ze względu na kształt. Omówiono też skorygowany funkcjonalny wykres pudełkowy służący do wykrywania funkcjonalnych obserwacji odstających ze względu na amplitudę. Elementy statystycznej analizy służącej do wykrywania obserwacji odstających zobrazowano na przykładzie danych pokazujących zanieczyszczenie powietrza w Katowicach oraz w Krakowie wybranymi czterema rodzajami substancji.

Słowa kluczowe: funkcjonalne obserwacje odstające, wykrywanie funkcjonalnych obserwacji odstających, statystyka odporna, głębia funkcjonalna, analiza zanieczyszczenia powietrza

FUNCTIONAL OUTLIERS DETECTION BY THE EXAMPLE OF AIR QUALITY MONITORING

Abstract

Methods of functional outliers detection in functional setting have been discussed, i.e. shape outliers and magnitude outliers. Outliergram has been discussed, a tool for functional shape outliers detection. Robust adjusted functional boxplot has been discussed as well, a tool for functional magnitude outliers detection. „The elements of functional outliers analysis have been applied to air pollution data for Katowice and Kraków."

Keywords: functional outliers, functional outliers detection, robust statistics, functional depth, air pollution analysis