

Analiza danych funkcjonalnych w badaniu stężenia pyłku olszy

1-3 lipca 2025

V Kongres Statystyki Polskiej

- **Agnieszka Kubik-Komar**, Uniwersytet Przyrodniczy w Lublinie, Katedra Zastosowań Matematyki i Informatyki, Lublin
- **Łukasz Smaga**, Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydział Matematyki i Informatyki, Poznań
- **Elżbieta Kubera**, Uniwersytet Przyrodniczy w Lublinie, Katedra Zastosowań Matematyki i Informatyki, Lublin
- **Małgorzata Graczyk**, Uniwersytet Przyrodniczy w Poznaniu, Katedra Metod Matematycznych i Statystycznych, Poznań
- **Krystyna Piotrowska-Weryszko**, Uniwersytet Przyrodniczy w Lublinie, Katedra Botaniki i Fizjologii Roślin, Lublin
- **Dorota Myszkowska**, Uniwersytet Jagielloński Centrum Medicum, Zakład Alergologii Klinicznej i Środowiskowej, Kraków
- **Łukasz Grewling**, Uniwersytet im. Adama Mickiewicza w Poznaniu, Katedra Botaniki Systematycznej i Środowiskowej, Poznań
- **Idalia Kasprzyk**, Uniwersytet Rzeszowski, Instytut of Biologii i Biotechnologii, Rzeszów
- **Ewa Łupikasza**, Uniwersytet Śląski, Wydział Nauk Przyrodniczych, Instytut Nauk o Ziemi, Sosnowiec
- **Barbara Majkowska-Wojciechowska**, Uniwersytet Medyczny w Łodzi, Klinika Immunologii i Alergii, Łódź
- **Małgorzata Malkiewicz**, Uniwersytet Wrocławski, Instytut Nauk Geologicznych, Wrocław
- **Małgorzata Puc**, Uniwersytet Szczeciński, Katedra Botaniki i Ochrony Środowiska, Szczecin

Dane

Funkcjonalna analiza składowych głównych

Funkcjonalna analiza skupień

Analiza danych dla wszystkich miast i lat

Regresja funkcjonalna - przewidywanie pylenia

Dane





- Dane dla olszy uzyskane zostały z pułapek typu Hirsta
 - 8 miast
 - 20 lat obserwacji
- IMGW
 - maksymalna temperatura dobową
 - minimalna temperatura dobową
 - średnia temperatura dobową
 - suma dobową opadu
 - średnia dobową prędkość wiatru
 - średnie dobowe zachmurzenie ogólne
 - średnia dobową wilgotność względną

Pułapka wolumetryczna typu Hirsta



Zbieranie danych: Pułapka wolumetryczna typu Hirsta

- Umieszczane na dachach dostępnych budynków (II, III piętro).
- Możliwość monitorowania ogólnego przepływu powietrza w otoczeniu, które zawiera mieszkankę lokalnych i bardziej odległych źródeł pyłków zebranych przez wiatr.
- Gdyby pułapka znajdowała się na poziomie gruntu, zbierałaby głównie pyłki z bezpośredniego sąsiedztwa, a wyniki między lokalizacjami nie byłyby porównywalne.
- Urządzenie to przeznaczone jest do pomiaru pyłków i zarodników grzybów w powietrzu.
- Przezroczyste powlekane pręty umieszcza się na zewnątrz, owija w małą konstrukcję chroniącą je przed żywiołami i codziennie sprawdza, ile przykleiło się do nich pyłku.
- Określa się jednostki pyłków na metr sześcienny powietrza i ustala poziom pyłków w danym obszarze.

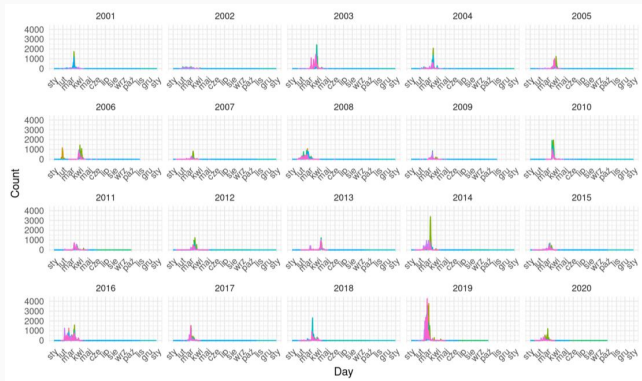
- Pobieranie próbek pyłków może być kosztowne i czasochłonne oraz wymaga dużego doświadczenia lub przeszkolenia.
- Pyłki są zbierane przez alergologów, agencje rządowe, instytucje edukacyjne, prywatnych lekarzy i komercyjne firmy badawcze.
- Ze względu na zastosowaną metodę nie można w 100% określić poziomu pyłku w każdym regionie, ale można pokazać trend dla danego obszaru.
- Na prognozę pylenia wpływa nie tylko pogoda, ale także klimat, a w szczególności zmiany klimatyczne, które jak wynika z najnowszych badań intensyfikują i wydłużają sezon pylenia.

Miasta w Polsce

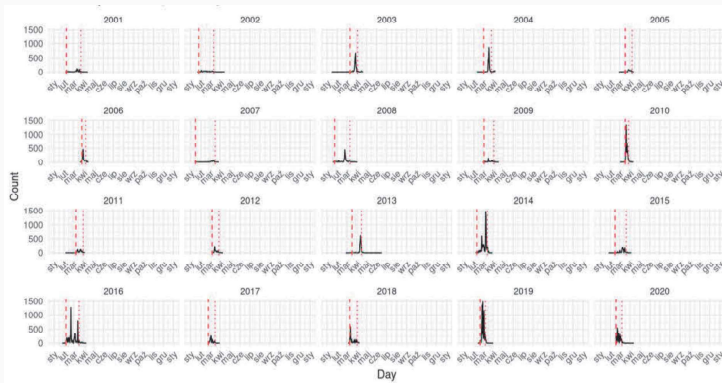
Miasta w Polsce



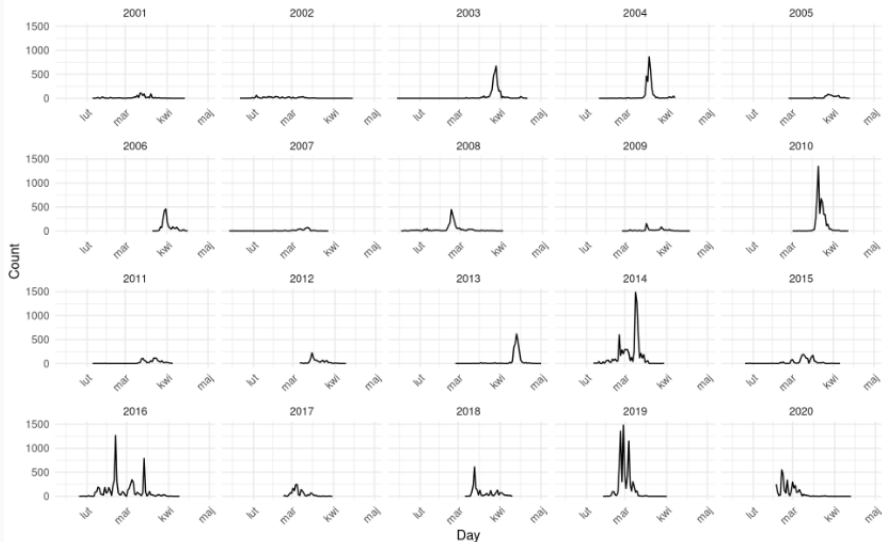
Trajektorie pylenia olszy



Trajektorie pylenia olszy - okresy pylenia



Trajektorie pylenia olszy w sezonie pyłkowym



Funkcjonalna analiza składowych głównych

Funkcjonalna analiza składowych głównych

- Niech $X(t)$ będzie obserwacją funkcyjną, $t \in \mathcal{I} = [a, b]$.
- Pierwszą wartość własną μ_1 i funkcję wagową ξ_1 definiujemy następująco:

$$\mu_1 = \sup_{\xi \in L_2[a,b]} \text{Var} \left(\int_a^b \xi(t) X(t) dt \right) = \text{Var} \left(\int_a^b \xi_1(t) X(t) dt \right)$$

$$\text{oraz } \int_a^b \xi_1^2(t) dt = 1.$$

- Analogicznie k -ta wartość własna μ_k i funkcja wagowa ξ_k :

$$\mu_k = \sup_{\xi \in L_2[a,b]} \text{Var} \left(\int_a^b \xi(t) X(t) dt \right) = \text{Var} \left(\int_a^b \xi_k(t) X(t) dt \right)$$

$$\text{oraz } \int_a^b \xi_k^2(t) dt = 1.$$

- k -ta składowa główna $\Xi_k = \int_a^b \xi_k(t) X(t) dt$ jest nieskorelowana z $k - 1$ pierwszymi składowymi głównymi.

- Niech $X_1(t), \dots, X_n(t)$ będzie próbą złożoną z niezależnych obserwacji funkcyjnych, gdzie $n \in \mathbb{N}$ oraz $t \in \mathcal{I} = [a, b]$.
- Wartości własne μ_j ($\mu_1 \geq \mu_2 \geq \dots \geq \mu_k \geq \dots$) oraz funkcje wagowe ξ_j są definiowane na podstawie zagadnienia własnego postaci:

$$\int_a^b v(s, t) \xi_j(t) dt = \mu_j \xi_j(s), \quad j = 1, 2, \dots,$$

gdzie

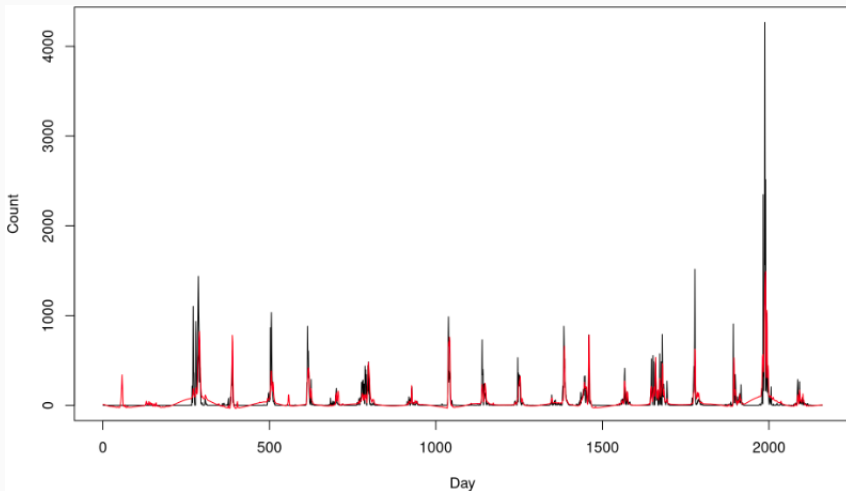
$$v(s, t) = \frac{1}{n-1} \sum_{i=1}^n (X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t)).$$

- FPCA zastosowano do danych dla olszy traktowanych jako gęste dane funkcyjne, a braki danych zamieniono na zera.
- Rozważono 8 obserwacji funkcyjnych odpowiadających rozważanym miastom w Polsce oraz latom 2001-2020.
- Otrzymano następujący skumulowany procent wyjaśnianej wariancji:

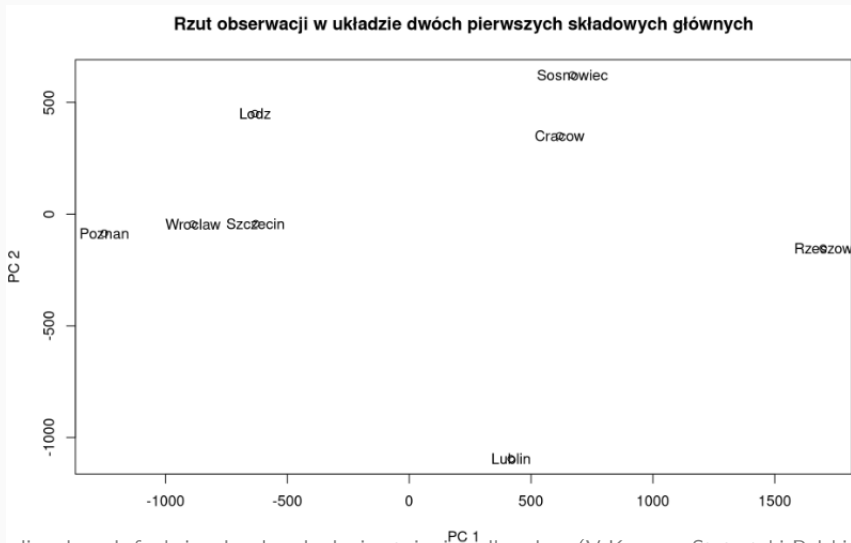
FPC1	FPC2	FPC3	FPC4	FPC5	FPC6	FPC7
0,5612	0,7382	0,8357	0,9031	0,9590	0,9869	1,0000

Funkcjonalna analiza składowych głównych - rekonstrukcje

- Rekonstrukcja danych dla Wrocławia poprzez FPCA



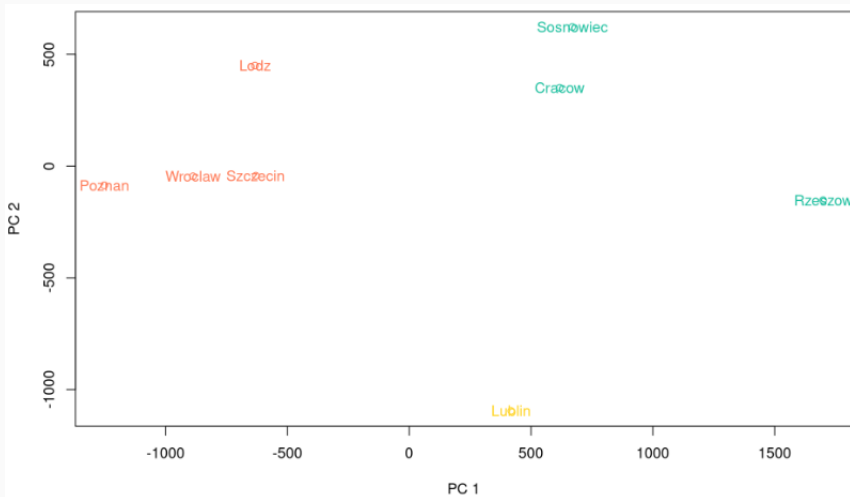
Funkcjonalna analiza składowych głównych - rzut



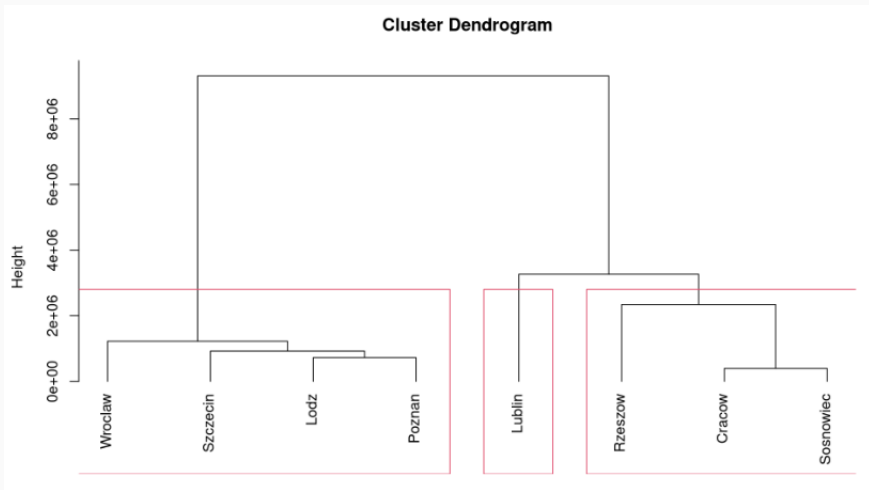
Funkcjonalna analiza skupień

- Istnieje wiele metod analizy skupień dedykowanych danym funkcjonalnym.
- Niestety dla naszych ośmiu gęstych obserwacji funkcjonalnych nie można ich wykonać ze względu na małą liczbę obserwacji.
- Dlatego zastosowano podejście rozważane w Rozdziale 9 monografii Crainiceanu i in. (2024).
- Podejście to polega na rzucie danych funkcjonalnych za pomocą FPCA na dane wielowymiarowe („współczynniki rozwinięcia”), a następnie zastosowanie znanych metod analizy skupień.
- Wzięliśmy 4 składowe główne, ponieważ wyjaśniają one ponad 90% wariancji.

Funkcjonalna analiza skupień - metoda K -średnich

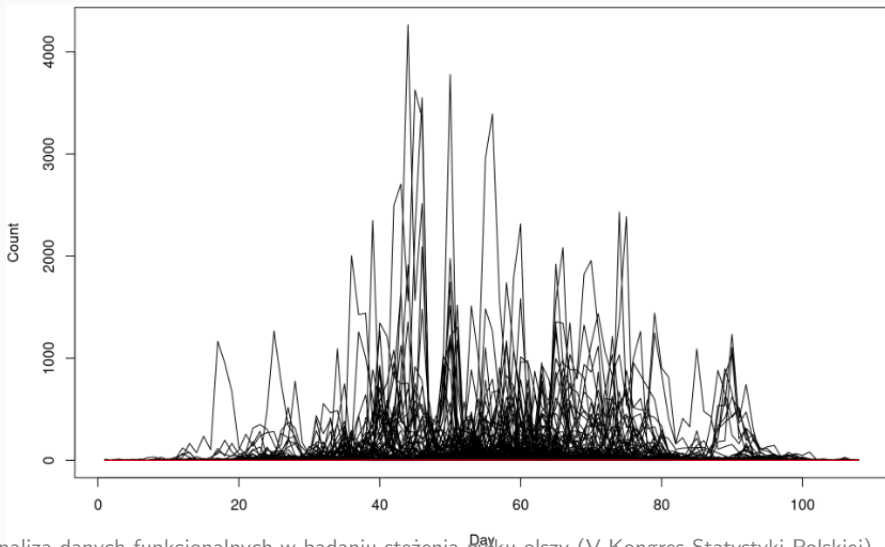


Funkcjonalna analiza skupień - dendrogram



Analiza danych dla wszystkich miast i lat

- Rozważmy teraz dane funkcjonalne dla par miasto-rok.
- Otrzymujemy w ten sposób 154 obserwacje funkcjonalne dla krótszych obcinków czasowych - 108 dni (okres od 14 stycznia do 30 kwietnia lub 1 maja danego roku: okres pylenia).



- FPCA została wykonana na danych przekształconych przez funkcję

$$f(x) = \ln(x + 1)$$

analogicznie jak zaproponowano w pracy Valderrama i in. (2010).

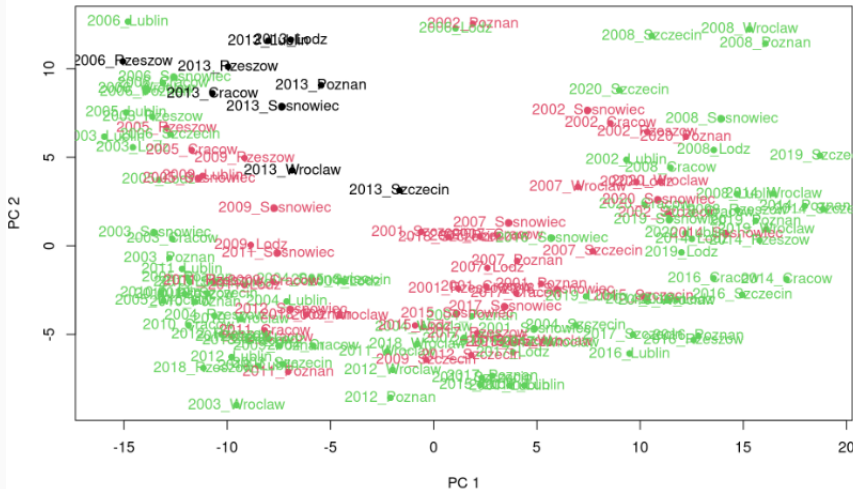
- Otrzymano następujący skumulowany procent wyjaśnianej wariancji:

FPC1	FPC2	FPC3	FPC4	FPC5	FPC6	FPC7
0,4652	0,6089	0,7072	0,7875	0,8368	0,8836	0,9101

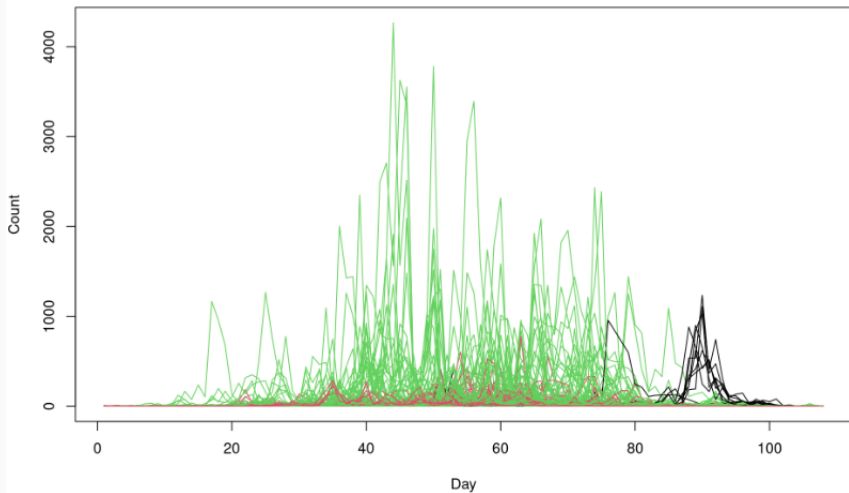
- Analizy skupień dokonano za pomocą metody analizy skupień w podprzestrzeni dyskryminacyjnej z pracy Bouveyron i in. (2015).
- Jest ona zaimplementowana w pakiecie `funFEM` języka R.
- Metoda ta bazuje na mieszanym modelu liniowym.
- Model ten opisuje zależność między współczynnikami rozwinięcia bazowego obserwacji funkcjonalnych w podprzestrzeni, do której należą te obserwacje, a ich współczynnikami w bazie podprzestrzeni ukrytej, którą uznaje się za najbardziej dyskryminującą podprzestrzeń dla K grup.

Dane miasto-rok: FPCA i funFEM

Rzut obserwacji w układzie dwóch pierwszych składowych głównych - kolory oznaczają skupienia



Dane miasto-rok: FPCA i funFEM



Regresja funkcjonalna - przewidywanie pylenia

- Celem jest dokonanie predykcji pylenia olszy w pewnym przedziale czasowym T na bazie jej pylenia oraz danych pogodowych w przedziale czasowym S .
- Za Valderram i in. (2010):
 - Przedział S nazywa się przedziałem poprzednim (ang. *past interval*), dla którego wartości zmiennych niezależnych i zależnych są znane.
 - Przedział T nazywa się przedziałem bieżącym (ang. *current interval*), dla którego wartości zmiennych nie są znane.
- W tym celu zastosowano model funkcja-funkcja regresji funkcjonalnej.

- Model funkcja-funkcja regresji funkcjonalnej został opisany w Rozdziale 6 książki Crainiceanu i in. (2024).
- ??? Strony 176-179 książki Crainiceanu, ale wydaje mi się, że wystarczy podać wzór na model, rozwijamy współczynniki w bazie i przechodzimy na standardowy model regresji.

- $W_i(s), s \in S$
- $X_{ir}(u_r), u_r \in U_r$

$$W_i(s) = f_0(s) + \int_U X_i(u)\beta(s, u)du + \epsilon_i(s)$$

$\epsilon_i(t) \sim N(0, \sigma_\epsilon^2)$ niezależne zmienne losowe

- $\beta(s, u) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \beta_{k_1 k_2} B_{k_1,1} B_{k_2,2}(u)$
- predyktor funkcjonalny
$$\int_U X_i(u) \beta(s, u) du = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \beta_{k_1 k_2} B_{k_1,1}(s) \int_U X_i(u) B_{k_2,2}(u) du$$
- $W_i(s_j) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \beta_{k_1 k_2} C_{j,k_1} D_{i,k_2} + \epsilon_i(s_j)$ where $A_{ji,k_1,k_2} = C_{j,k_1} D_{i,k_2}$ – predyktory $W_i(s_j)$
 $\beta = (\beta_{11}, \beta_{12}, \dots, \beta_{K_1 K_2})'$ – parametry regresji



$W = X\beta + \epsilon$ – standardowy model regresji

- Za poprzedni przedział S uznajemy okres pomiędzy 24 stycznia a 13-14 marca odpowiednio dla roku przestępnego i nieprzestępnego.
- Bieżący przedział T to czas pomiędzy 14-15 marca a 30 kwietnia i 1 maja odpowiednio dla roku przestępnego i nieprzestępnego.
- Zatem zbiór S uwzględnia 50 dni, a T 48 dni, czyli prawie po równo.
- Dane miasto-rok podzielono na
 - zbiór uczący - od 2001 do 2016,
 - zbiór testowy - od 2017 do 2020.
- To daje 124 obserwacje (pary miasto-rok) w zbiorze uczącym i 30 w zbiorze testowym - standardowy podział 80% – 20%.

1. pylenie
2. maksymalna temperatura dobowa
3. minimalna temperatura dobowa
4. średnia temperatura dobowa
5. suma dobowa opadu
6. średnia dobowa prędkość wiatru
7. średnie dobowe zachmurzenie ogólne
8. średnia dobowa wilgotność względna

1. $\ln(\text{pylenie} + 1)$
2. $\ln(\max \text{ temp dobowa} + 1 + |\min(\text{maksymalna temp dob})|)$
3. $\ln(\min \text{ temp dobowa} + 1 + |\min(\text{minimalna temp dob})|)$
4. $\ln(\text{średnia temp. dobowa} + 1 + |\min(\text{średnia temp. dobowa})|)$
5. $\ln(\text{suma dobowa opadu} + 1)$
6. $\ln(\text{średnia dobowa predkość wiatru} + 1)$
7. $\ln(\text{średnie dobowe zachmurzenie ogólne} + 1)$
8. $\ln(\text{średnia dobowa wilgotność względna} + 1)$

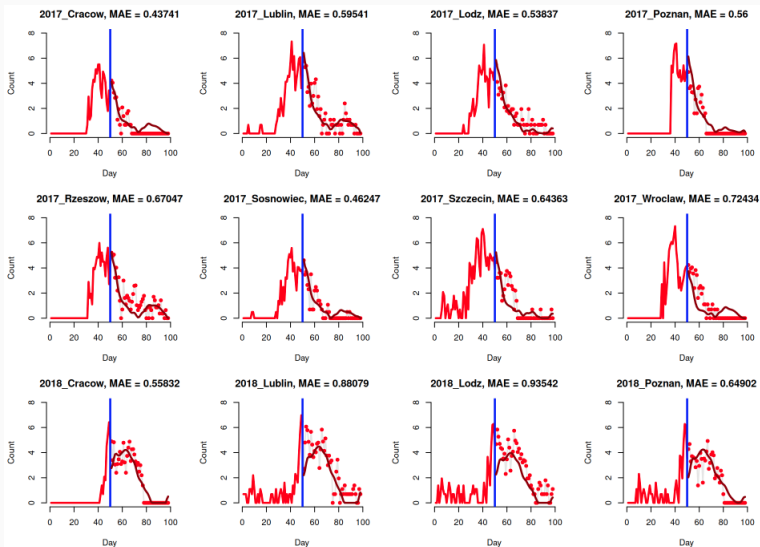
- Y_{ij} oznacza wartość i -tej obserwacji funkcjonalnej w j -tym punkcie czasowym w zbiorze testowym.
- $RMSE = \sqrt{\frac{1}{I \cdot J} \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \hat{Y}_{ij})^2}$
- $MAE = \frac{1}{I \cdot J} \sum_{i=1}^I \sum_{j=1}^J |Y_{ij} - \hat{Y}_{ij}|$
- $MAPE = 100 \cdot \frac{1}{|I| \cdot |J|} \sum_{i \in I} \sum_{j \in J} \left| \frac{Y_{ij} - \hat{Y}_{ij}}{Y_{ij}} \right|$
- $SMAPE = 100 \cdot \frac{1}{|I| \cdot |J|} \sum_{i \in I} \sum_{j \in J} \frac{|Y_{ij} - \hat{Y}_{ij}|}{(|Y_{ij}| + |\hat{Y}_{ij}|)/2}$
- Tutaj zbiory I i J są zbiorami indeksów, dla których Y_{ij} jest różne od zera.

10 najlepszych modeli względem MAE

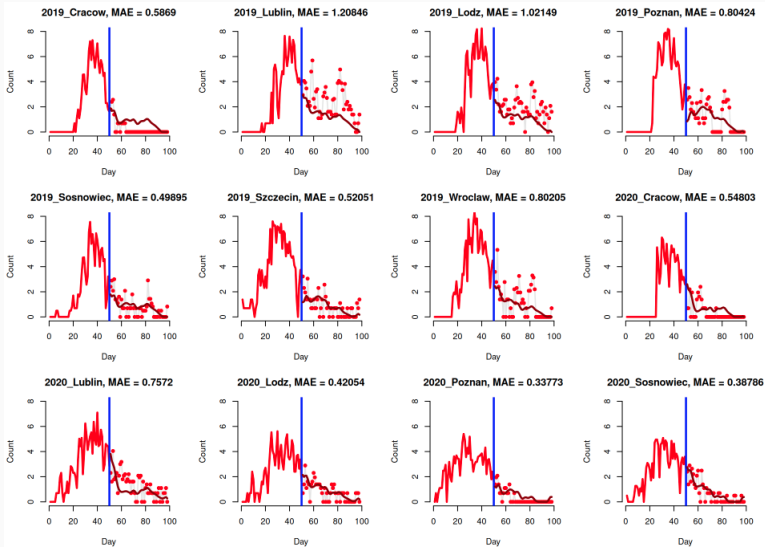
- Oceny błędu zostały policzone na zbiorze testowym.

Model	RMSE	MAE	MAPE	SMAPE
pylenie • srednia temp dobowa • suma dobowa opadu	0.9483723	0.6820621	46.27250	56.59703
pylenie • max temp dobowa • suma dobowa opadu	0.9283679	0.6835761	44.21476	49.80598
min temp dobowa • srednia temp dobowa • suma dobowa opadu	0.9639578	0.6914994	50.22651	56.21168
pylenie • min temp dobowa • srednia temp dobowa • suma dobowa opadu	0.9491779	0.6927701	45.80184	51.67203
max temp dobowa • suma dobowa opadu	0.9879915	0.6944810	50.08804	58.13726
max temp dobowa • min temp dobowa • suma dobowa opadu	0.9998568	0.7058032	53.66269	60.43695
pylenie • max temp dobowa • min temp dobowa • suma dobowa opadu	0.9819094	0.7064519	48.12462	51.96799
pylenie • srednia temp dobowa	0.9500371	0.7079343	48.16724	57.24291
max temp dobowa	0.9942908	0.7146617	53.02336	61.98991
srednia temp dobowa	1.0005901	0.7147146	53.45588	62.52690

Przykładowe prognozy na zbiorze testowym



Przykładowe prognozy na zbiorze testowym



- Uwzględnienie pogody sprzed okresu pylenia
- Analiza dla dekad lub tygodni

1. Bouveyron C., Côme E., Jacques J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics* 9(4), 1726-1760.
2. Crainiceanu C.M., Goldsmith J., Leroux A., Cui E. (2024). *Functional Data Analysis with R*. Chapman and Hall/CRC, New York.
3. Valderrama M.J., Ocaña F.A., Aguilera A.M., Ocaña-Peinado F.M. (2010). Forecasting pollen concentration by a two-step functional model. *Biometrics* 66, 578-585.
4. <https://www.pogodairadar.pl/wiadomosci-pogodowe/jak-mierzy-sie-pylki-w-powietrzu--87cfbc77-ec0a-405a-97a2-ad4243acea74>