Wpływ parametrów modelu GAM na estymacje stopni swobody

Zuzanna Nogala

Czym jest model GAM?
Co go łączy z modelami GLM?

GLM (Generalized Linear Model)

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Założenia:

- Y należy do rodziny wykładniczej (Exponential Family), gdzie E[Y] = μ
- g(.) to funkcja linkująca, która przekształca wartość oczekiwaną E[Y] w predyktor liniowy

GLM (Generalized Linear Model)

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Założenia:

- Y należy do rodziny wykładniczej (Exponential Family), gdzie E[Y] = μ
- g(.) to funkcja linkująca, która przekształca wartość oczekiwaną E[Y] w predyktor liniowy

- β_0 – wyraz wolny
- β_i - Współczynniki regresji określa globalną i stałą zmianę predyktora liniowego wywołaną wzrostem zmiennej X_j o jedną jednostkę, przy założeniu niezmienności pozostałych zmiennych

Estymacja parametrów (Maximum Likelihood)

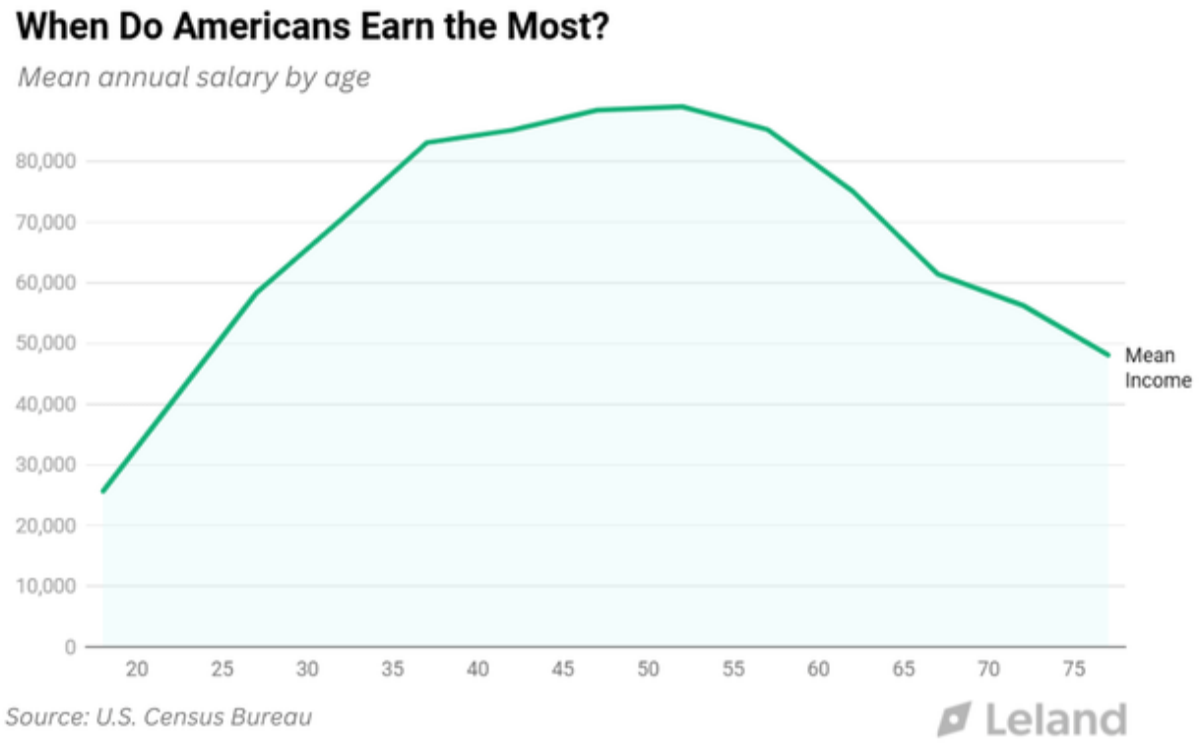
Maksymalizujemy logarytm funkcji wiarygodności (log-likelihood)

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \phi)$$

GLM (Generalized Linear Model)

- Zakładamy, że wpływ predyktorów na wynik jest **liniowy.**
- Wysoka interpretowalność.
- Niskie ryzyko przeuczenia.

Co jeżeli nie możemy dopasować funkcji liniowe?



GLM:

- Model spróbowałby poprowadzić przez te dane jedną, prostą linię.
- Wniosek: zarobki rosną wraz z wiekiem
- Ekspercko

GAM (Generalized Additive Model)

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

Założenia:

- Y należy do rodziny wykładniczej (Exponential Family), gdzie E[Y] = μ
- g(.) to funkcja linkująca, która przekształca wartość oczekiwaną E[Y] w predyktor liniowy

- β_0 – wyraz wolny

GAM (Generalized Additive Model)

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

Założenia:

- Y należy do rodziny wykładniczej (Exponential Family), gdzie E[Y] = μ
- g(.) to funkcja linkująca, która przekształca wartość oczekiwaną E[Y] w predyktor liniowy

- β_0 – wyraz wolny
- f_i(.) - funkcje wygładzające, które postaci:

$$f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x)$$

Estymacja parametrów (Penalized Maximum Likelihood)

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \int \left[f_j''(x)\right]^2 dx$$

gdzie:

• l(β) – Klasyczny logarytm funkcji wiarygodności (log-likelihood) dla danych z rozkładu rodziny wykładniczej.

Estymacja parametrów (Penalized Maximum Likelihood)

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \int \left[f_j''(x)\right]^2 dx$$

gdzie:

• l(β) – Klasyczny logarytm funkcji wiarygodności (log-likelihood) dla danych z rozkładu rodziny wykładniczej.

• $\int \left[f_j''(x)\right]^2$ - **Funkcjonał kary (penalty functional)** mierzący stopień pofalowania (wiggliness).

Estymacja parametrów (Penalized Maximum Likelihood)

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \int \left[f_j''(x)\right]^2 dx$$

gdzie:

• l(β) – Klasyczny logarytm funkcji wiarygodności (log-likelihood) dla danych z rozkładu rodziny wykładniczej.

• $\int \left[f_j''(x)\right]^2$ - **Funkcjonał kary (penalty functional)** mierzący stopień pofalowania (wiggliness).

• λ_j – **Parametry wygładzania (smoothing parameters)**. Skalary sterujące kompromisem (trade-off) pomiędzy wiernością dopasowania do danych empirycznych a gładkością estymowanej krzywej. Są one optymalizowane analitycznie(z reguły przy wykorzystaniu kryteriów)

GAM (Generalized Additive Model)

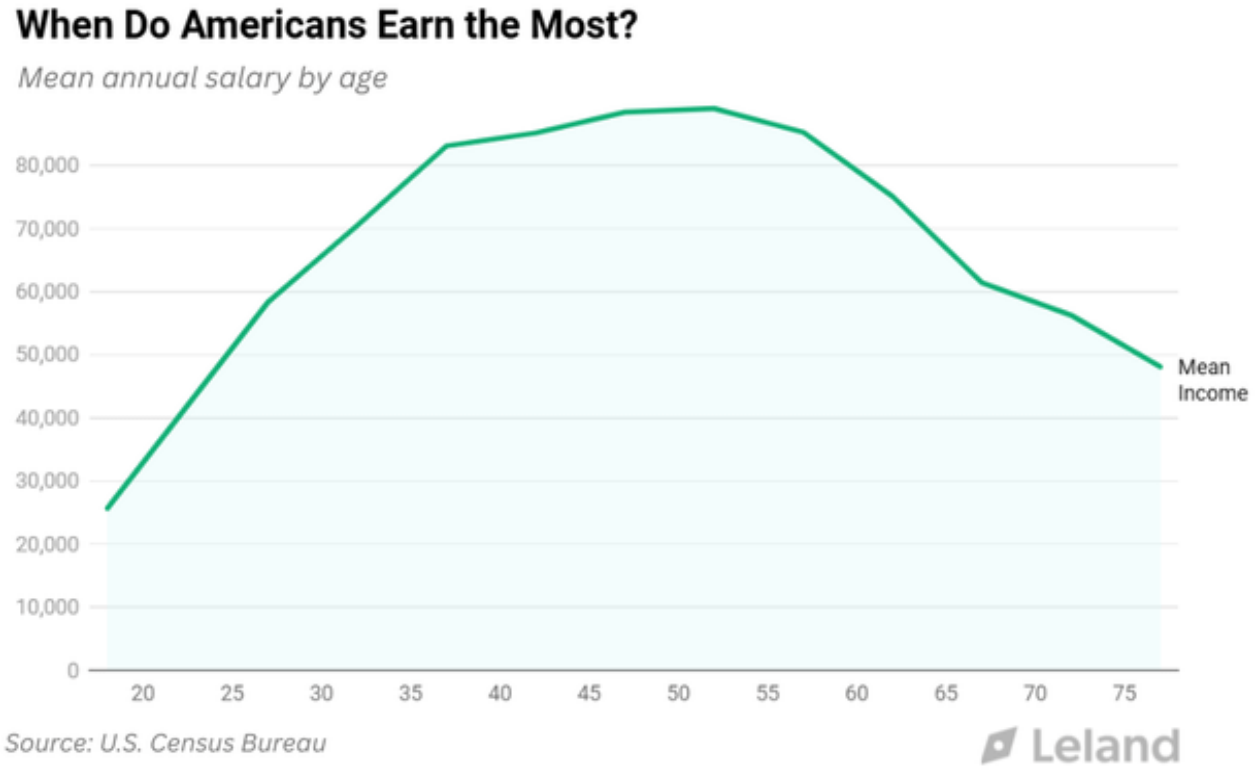
- Nie zakładamy, że wpływ na wynik jest liniowy.
- Również dobra interpretowalność
- Możliwość przeuczenia

Porównanie GLM i GAM

Badamy wpływ wieku (X) na średnie zarobki ($g(\mu)$)

GLM:

- Model spróbowałby poprowadzić przez te dane jedną, prostą linię.
- Wniosek: zarobki rosną wraz z wiekiem
- Ekspercko



Porównanie GLM i GAM

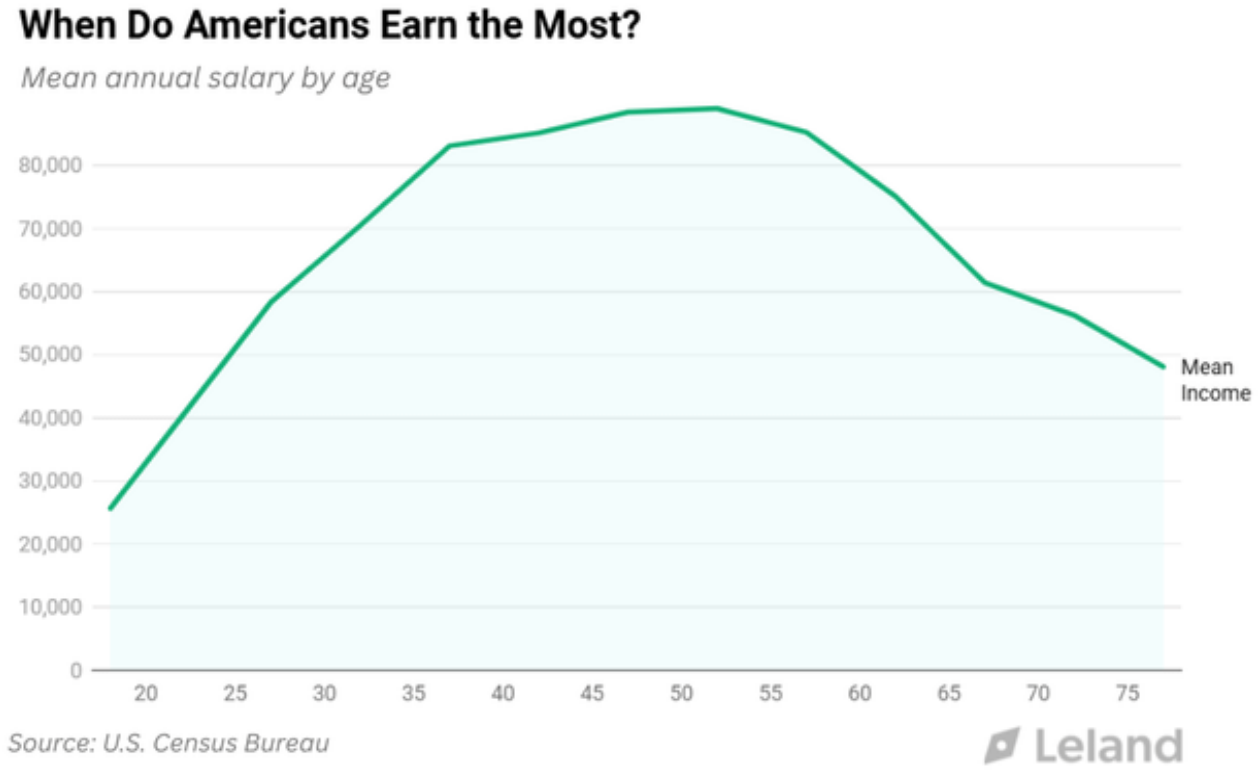
Badamy wpływ wieku (X) na średnie zarobki ($g(\mu)$)

GLM:

- Model spróbowałby poprowadzić przez te dane jedną, prostą linię.
- Wniosek: zarobki rosną wraz z wiekiem
- Ekspercko

GAM:

- Model sam za pomocą funkcji f_i szuka odpowiedniego wygięcia funkcji



Podsumowanie modelu GAM

- GAM jest rozszerzenie GLM
- Model pozwala na modelowanie **nieliniowych zależności** w danych
- Model nie wymusza założenia, że zmienne objaśniające mają stały, liniowy wpływ na wynik

GAM - największa zaleta

Kompromis pomiędzy:

- regresją liniową (interpretowalność, ale mało elastyczna)
- a Machine Learning (skuteczny złożony algorytmem, ale "czarna skrzynka")

Różnice pomiędzy GAM i GLM

Cecha	GLM	GAM
Kształt zależności	Sztywny (stały trend)	Elastyczny (zmienny trend)
Podejście do danych	Z góry relacja liniowa	Dopasowuje się do danych (mogą być relacje nieliniowe)

Różnice pomiędzy GAM i GLM

Cecha	GLM	GAM
Kształt zależności	Sztywny (stały trend)	Elastyczny (zmienny trend)
Podejście do danych	Z góry relacja liniowa	Dopasowuje się do danych (mogą być relacje nieliniowe)
Estymacja parametrów	Jednowarstowa	Dwuwarstowa (β, lambda)
Parametry dla zmiennej	Współczynniki β_i	Cała krzywa f_i

Różnice pomiędzy GAM i GLM

Cecha	GLM	GAM
Kształt zależności	Sztywny (stały trend)	Elastyczny (zmienny trend)
Podejście do danych	Z góry relacja liniowa	Dopasowuje się do danych (mogą być relacje nieliniowe)
Estymacja parametrów	Jednowarstowa	Dwuwarstowa (β, lambda)
Parametry dla zmiennej	Współczynniki β_i	Cała krzywa f_i
Podatność na szum	Niska (ignorowanie małych odchyleń)	Wysoka (ale kontrolowane)

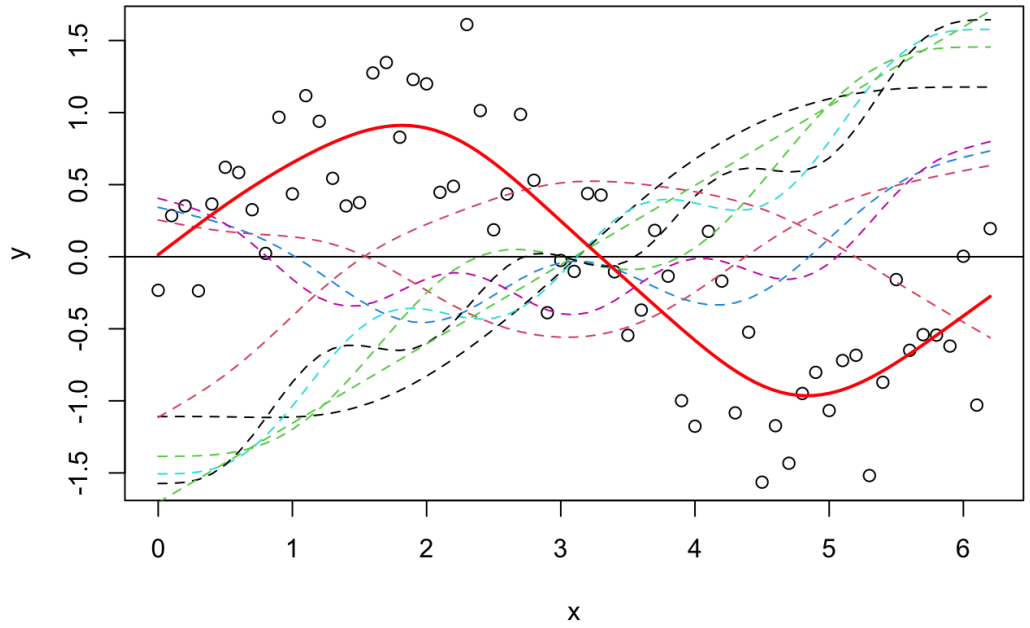
Podstawowe pojęcia w modelu GAM

Funkcja wygładzająca

$$f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x)$$

Gdzie dla każdej zmiennej j:

- b_{jk}(x) to znana k-ta funkcja bazowa zdefiniowana na dziedzinie zmiennej X_j.
- β_{jk} to współczynniki rozwinięcia w bazie, które podlegają estymacji. Nadają wagę i określają wpływ funkcji bazowej

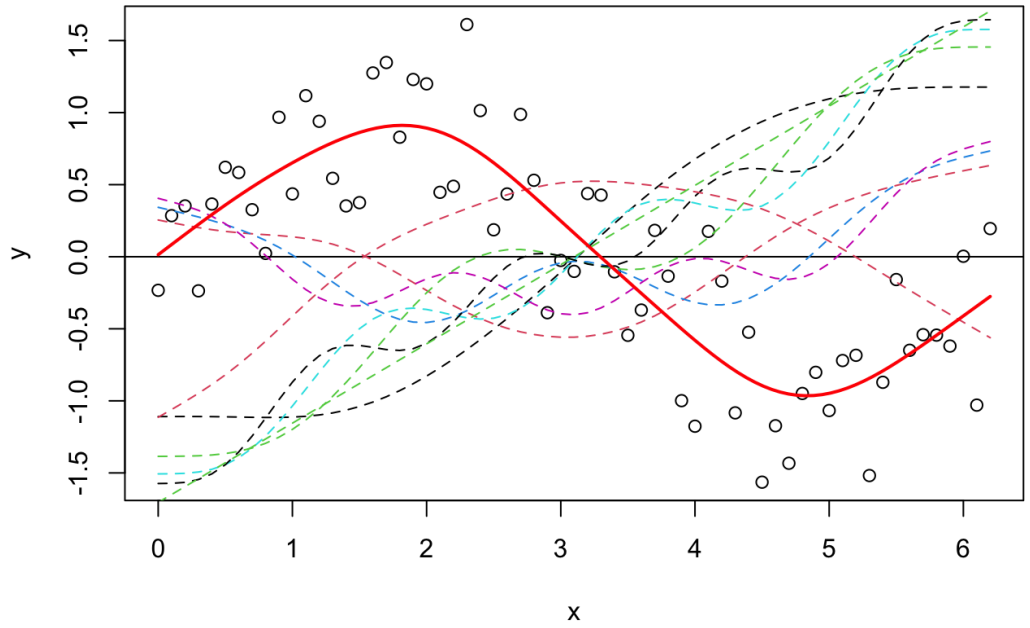


Funkcja wygładzająca

$$f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x)$$

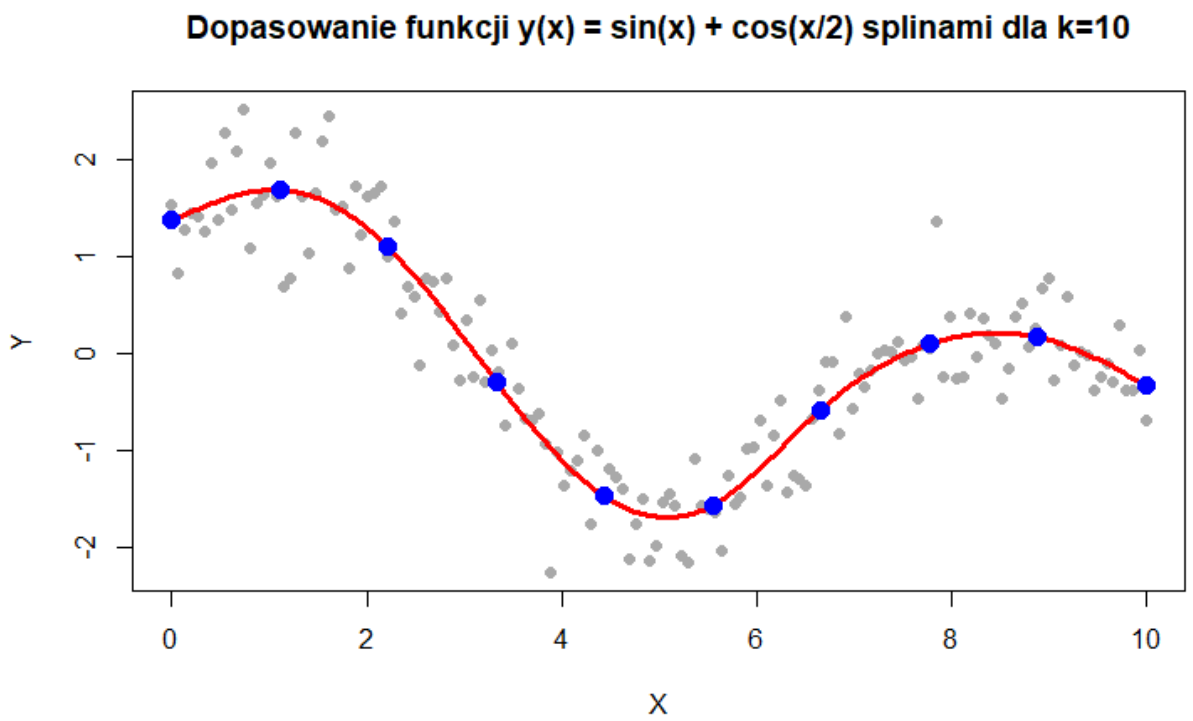
Gdzie dla każdej zmiennej j:

- b_{jk}(x) to znana k-ta funkcja bazowa zdefiniowana na dziedzinie zmiennej X_j.
- β_{jk} to współczynniki rozwinięcia w bazie, które podlegają estymacji. Nadają wagę i określają wpływ funkcji bazowej
- K_j to z góry zadany wymiar bazy (maksymalna złożoność funkcji).



Funkcja sklejana (spline)

- Metoda implementacji funkcji wygładzającej
- służy do modelowania gładkich krzywych.
- buduje krzywą **odcinkowo z wielu mniejszych, prostych fragmentów**



Funkcja sklejana (spline)

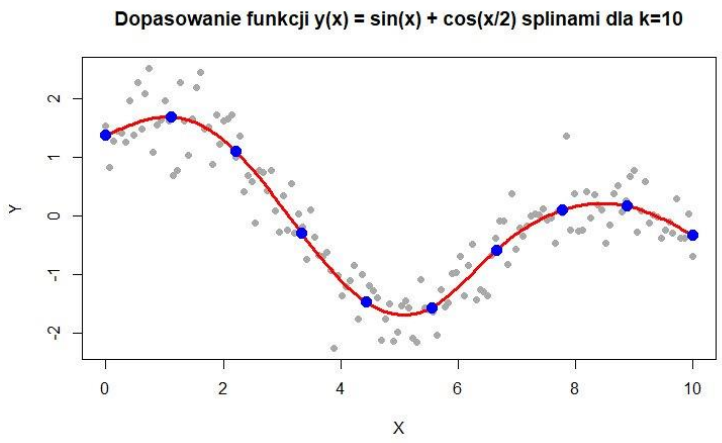
Metoda implementacji funkcji wygładzających

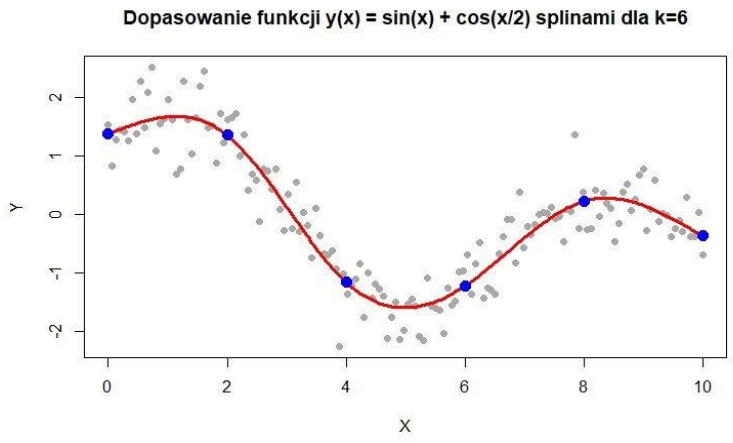
Cubic Splines	P-Spline	Thin Plate Regression Spline

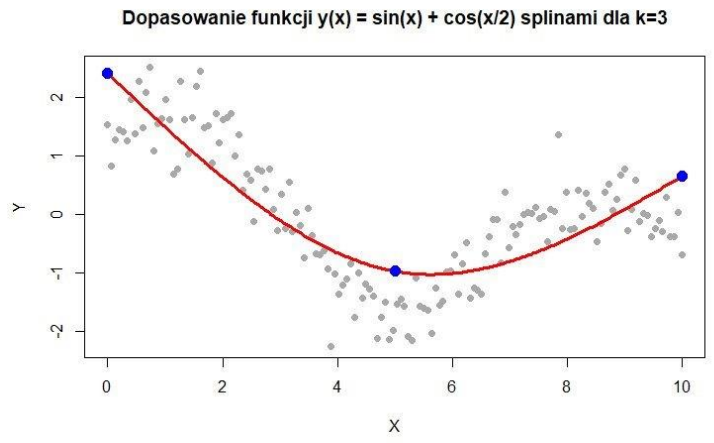
Funkcja sklejana (spline)

Metoda implementacji funkcji wygładzających

Cubic Splines	P-Spline	Thin Plate Regression Spline
• zdefiniowane odcinkami za pomocą wielomianów, które łączą się w punktach zwanych **węzłami (knots)** • Problem: wybór węzłów		
• Do danych jednowymiarowych; szybki		


Dopasowanie funkcji y(x) = sin(x) + cos(x/2) splinami dla k=10
Y
X
-2
-1
0
1
2
0
2
4
6
8
10


Dopasowanie funkcji y(x) = sin(x) + cos(x/2) splinami dla k=6
Y
X
-2
-1
0
1
2
0
2
4
6
8
10


Dopasowanie funkcji y(x) = sin(x) + cos(x/2) splinami dla k=3
Y
X
-2
-1
0
1
2
0
2
4
6
8
10

Funkcja sklejana (spline)

Metoda implementacji funkcji wygładzających

Cubic Splines	P-Spline	Thin Plate Regression Spline
• zdefiniowane odcinkami za pomocą wielomianów, które łączą się w punktach zwanych **węzłami (knots)** • Problem: wybór węzłów	• równomiernej siatce węzłów nakładają dyskretną karę na różnicę pomiędzy współczynnikami sąsiadujących funkcji bazowych	
• Do danych jednowymiarowych; szybki	• Dobre do nieregularnych danych (braki w danych; nierównomierne zagęszczenie)	

Funkcja sklejana (spline)

Metoda implementacji funkcji wygładzających

Cubic Splines	P-Spline	Thin Plate Regression Spline
• zdefiniowane odcinkami za pomocą wielomianów, które łączą się w punktach zwanych **węzłami (knots)** • Problem: wybór węzłów	• równomiernej siatce węzłów (omijająca problem ich ręcznego rozmieszczania) • nakładają dyskretną karę na różnicę pomiędzy współczynnikami sąsiadujących funkcji bazowych	**Thin Plate Spline** • Eliminuja potrzebe stosowania węzłów • Opiera się na fizycznej energii zgięcia
• Do danych jednowymiarowych; szybki	• Dobre do nieregularnych danych (braki w danych; nierównomierne zagęszczenie)	• Domyślna opcja w mgcv, "złoty standard"

Funkcja sklejana (spline)

Metoda implementacji funkcji wygładzających

Cubic Splines	P-Spline	Thin Plate Regression Spline
• zdefiniowane odcinkami za pomocą wielomianów, które łączą się w punktach zwanych **węzłami (knots)** • Problem: wybór węzłów	• równomiernej siatce węzłów (omijająca problem ich ręcznego rozmieszczania) • nakładają dyskretną karę na różnicę pomiędzy współczynnikami sąsiadujących funkcji bazowych	**Thin Plate Spline** • Eliminuja potrzebe stosowania węzłów • Opiera się na fizycznej energii zgięcia **Thin Plate Regression Spline** • Za pomocą dekompozycji SVD wybiera k najważniejszych wygięć
• Do danych jednowymiarowych; szybki	• Dobre do nieregularnych danych (braki w danych; nierównomierne zagęszczenie)	• Domyślna opcja w mgcv, "złoty standard"

EDF (Effective Degrees of Freedom)

Miara złożoności krzywej, określa stopień nieliniowości funkcji

- **EDF = 1:** zależność liniowa
- **EDF > 1:** Zależność jest nieliniowa (np. EDF=2 przypomina parabolę)
- **Im wyższe EDF** (zbliżające się do wartości k-1), tym bardziej złożona i pofalowana jest funkcja.

Analiza EDF pozwala natychmiast ocenić, czy zmienna ma na wynik wpływ liniowy, czy nieliniowy.

EDF (Effective Degrees of Freedom)

EDF nie jest stopniem wielomianu!

- EDF może być liczbą rzeczywistą (np.. EDF = 2.4)

- **EDF = 2:** Krzywa ma jedno wygięcie
- **EDF = 3:** Krzywa ma dwa wygięcia
- **EDF > 5:** Funkcja jest mocno pofalowa

Skąd ta część ułamkowa w EDF?

Klasyczna regresja liniowa (LM)

$$\hat{y} = Hy$$

$$H = X(X^T X)^{-1} X^T$$

$$DF = \mathrm{tr}(H) = p$$

Uogólnione modele addytywne (GAM)

Skąd ta część ułamkowa w EDF?

Klasyczna regresja liniowa (LM)

$$\hat{y} = Hy$$

$$H = X(X^T X)^{-1} X^T$$

$$DF = \mathrm{tr}(H) = p$$

Uogólnione modele addytywne (GAM)

$$\hat{y} = H_\lambda y$$

$$H_\lambda = X(X^T X + \lambda S)^{-1} X^T$$

$$EDF = \mathrm{tr}(H_\lambda)$$

$$EDF_i = \frac{1}{1 + \lambda \cdot s_i}$$

Wymiar bazy (k) a EDF

- k to górny limit złożoność a EDF to "zużyte" stopnie swobody po nałożeniu kary

Gamma - **mnożnik kary za złożoność (penalty multiplier)**

Domyślne działanie (gamma = 1):

- Model traktuje każdy wykorzystany stopień swobody standardowo.
- Ocenia kompromis między dokładnością a złożonością w sposób 1:1.

Gamma - **mnożnik kary za złożoność (penalty multiplier)**

Domyślne działanie (gamma = 1):

- Model traktuje każdy wykorzystany stopień swobody standardowo.
- Ocenia kompromis między dokładnością a złożonością w sposób 1:1.

Zwiększenie kary (np. gamma = 1.4):

- *traktowanie każdego stopienia swobody tak, jakby kosztował o 40% więcej*.
- złożone, pofalowane krzywe (wysokie EDF) są dla niego bardzo "drogie".
- algorytm preferuje wybór większego lambda, co prowadzi do spłaszczenia i wygładzenia krzywej.

Metody wygładzania

REML (Restricted Maximum Likelihood)	**GCV.CP** (Generalized Cross Validation)	**ML** (Maximum Likelihood)

Metody wygładzania

REML (Restricted Maximum Likelihood)	**GCV.CP** (Generalized Cross Validation)	**ML** (Maximum Likelihood)
Metoda najpierw odfiltrowuje efekty stałe/ liniowe i skupia się na estymacji wariancji i gładkości funkcji. "Złoty standard" w mgcv (opcja domyślna)		

Metody wygładzania

REML (Restricted Maximum Likelihood)	**GCV.CP** (Generalized Cross Validation)	**ML** (Maximum Likelihood)
Metoda najpierw odfiltrowuje efekty stałe/ liniowe i skupia się na estymacji wariancji i gładkości funkcji. "Złoty standard" w mgcv (opcja domyślna)	Metoda oparta na predykcji. Szacuje, jak dobrze model poradzi sobie na nowych danych (Leave-One-Out Cross Validation) i dodaje karę za złożoność (Cp).	

Metody wygładzania

REML (Restricted Maximum Likelihood)	**GCV.CP** (Generalized Cross Validation)	**ML** (Maximum Likelihood)
Metoda najpierw odfiltrowuje efekty stałe/ liniowe i skupia się na estymacji wariancji i gładkości funkcji. "Złoty standard" w mgcv (opcja domyślna)	Metoda oparta na predykcji. Szacuje, jak dobrze model poradzi sobie na nowych danych (Leave-One-Out Cross Validation) i dodaje karę za złożoność (Cp).	Metoda jednocześnie chce estymować efekty stałe/liniowe i wariancję. Przez co bywa zbyt optymistycznie podchodzi do wariancji i ją zaniża.

Plan pracy

1. Wstęp teoretyczny

2. Część symulacyjna – badanie wpływu parametrów modelu GAM na estymowanie EDF

3. Część praktyczna – model GAM w praktyce

2. Część symulacyjna

Przeprowadzenie 5 rodzajów analiz wrażliwość EDF na wybrane parametry

ANALIZA A: Wybór metody doboru wygładzania

ANALIZA B: Wpływ gammy i k

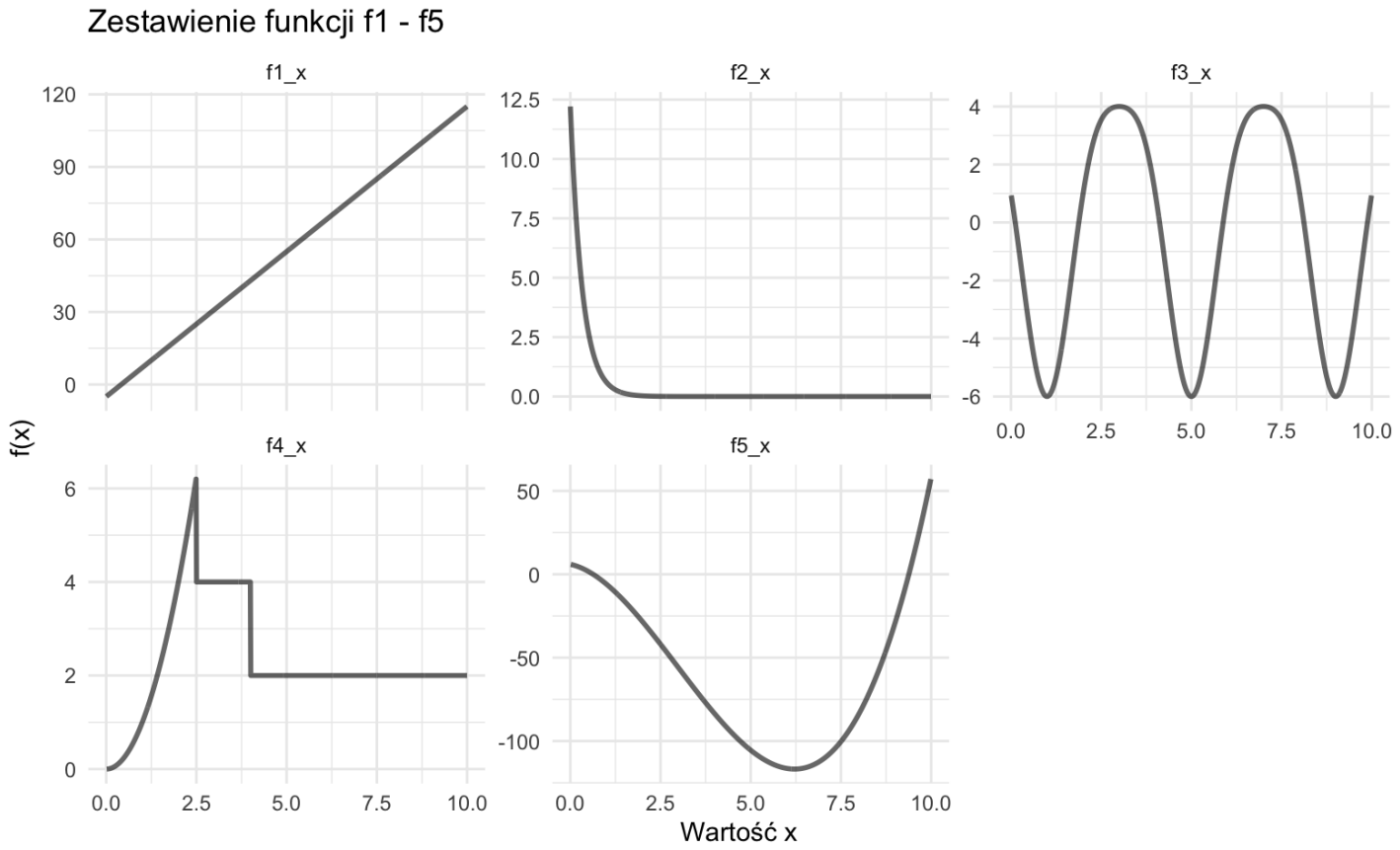
ANALIZA C: Typ bazy splina

ANALIZA D: Selekcja zmiennych

ANALIZA E: Wpływ liczby zmiennych

Liczba powtórzeń dla danego zestawu parametrów: K = 500

Analiza A - wpływ metody wygładzania



$$X = (X_1, X_2, X_3, X_4, X_5)$$

$$X_i \sim U[0, 10]$$

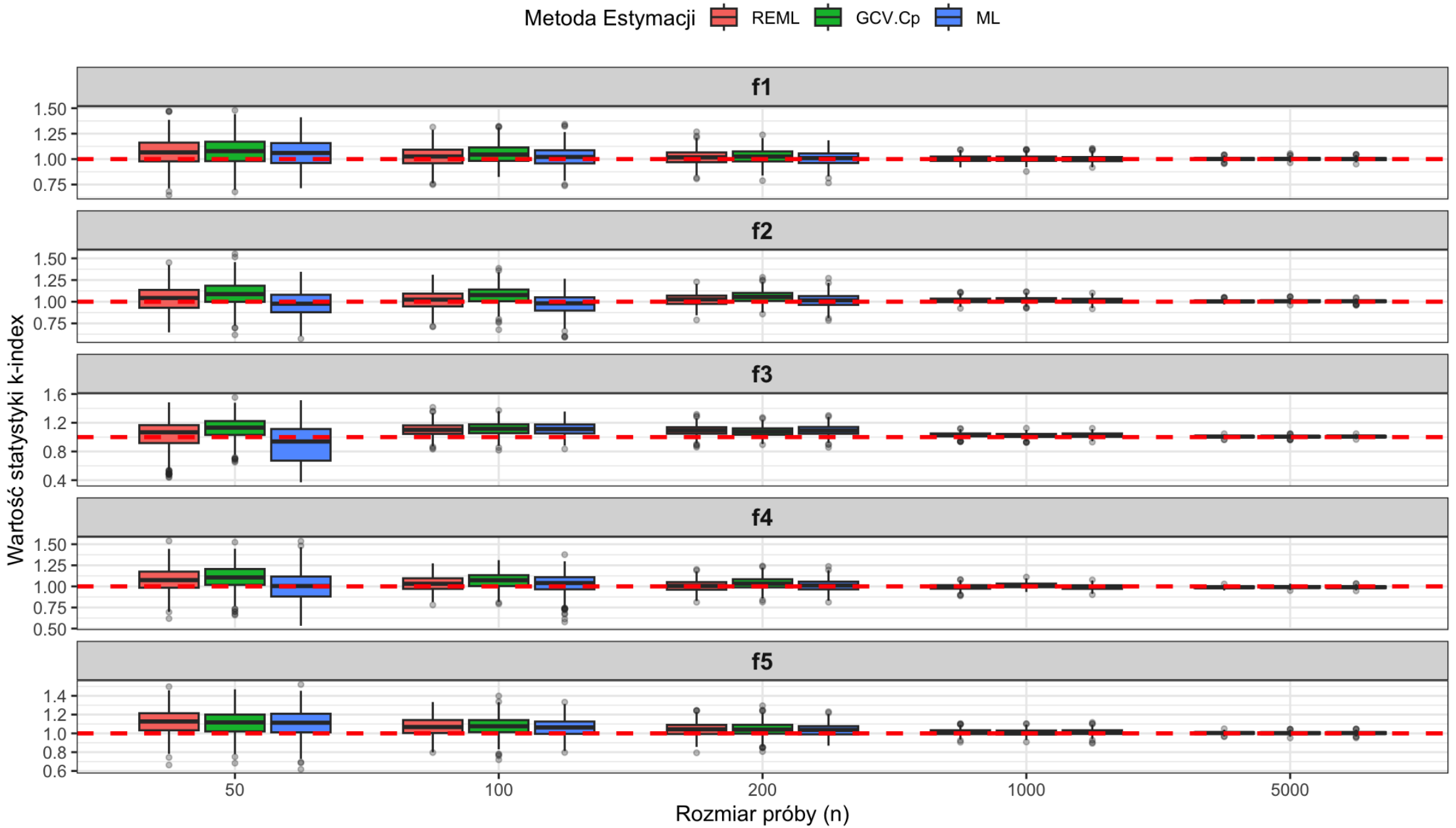
$$y(X) = \sum_{i=1}^{5} f_i(X_i)$$

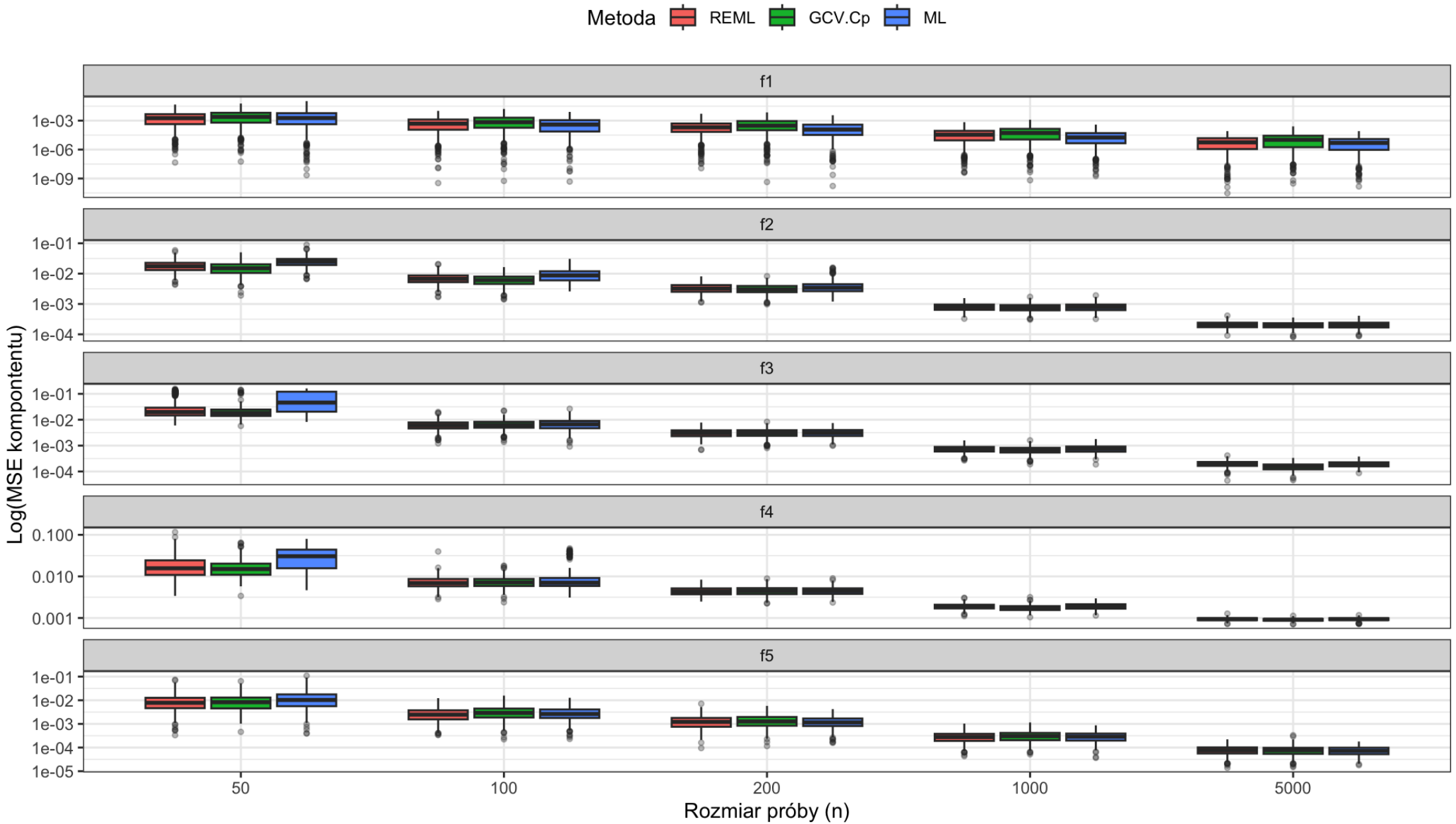
Parametry:

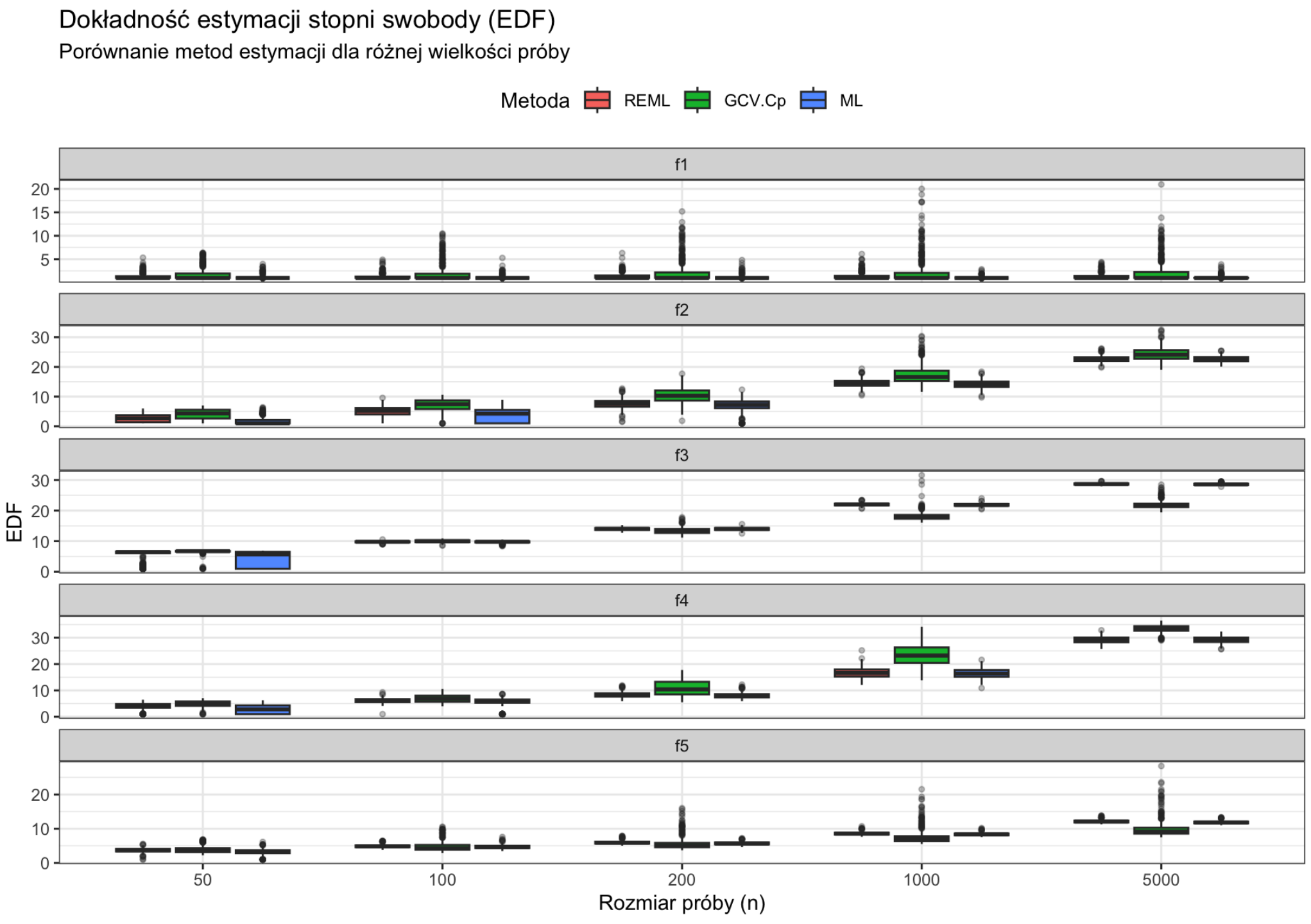
- **Metoda:**
  REML, GCV.Cp, ML
- **Rozmiar próby (n):**
  50, 100, 200, 1000, 5000

Diagnoza autokorelacji reszt (k-index)

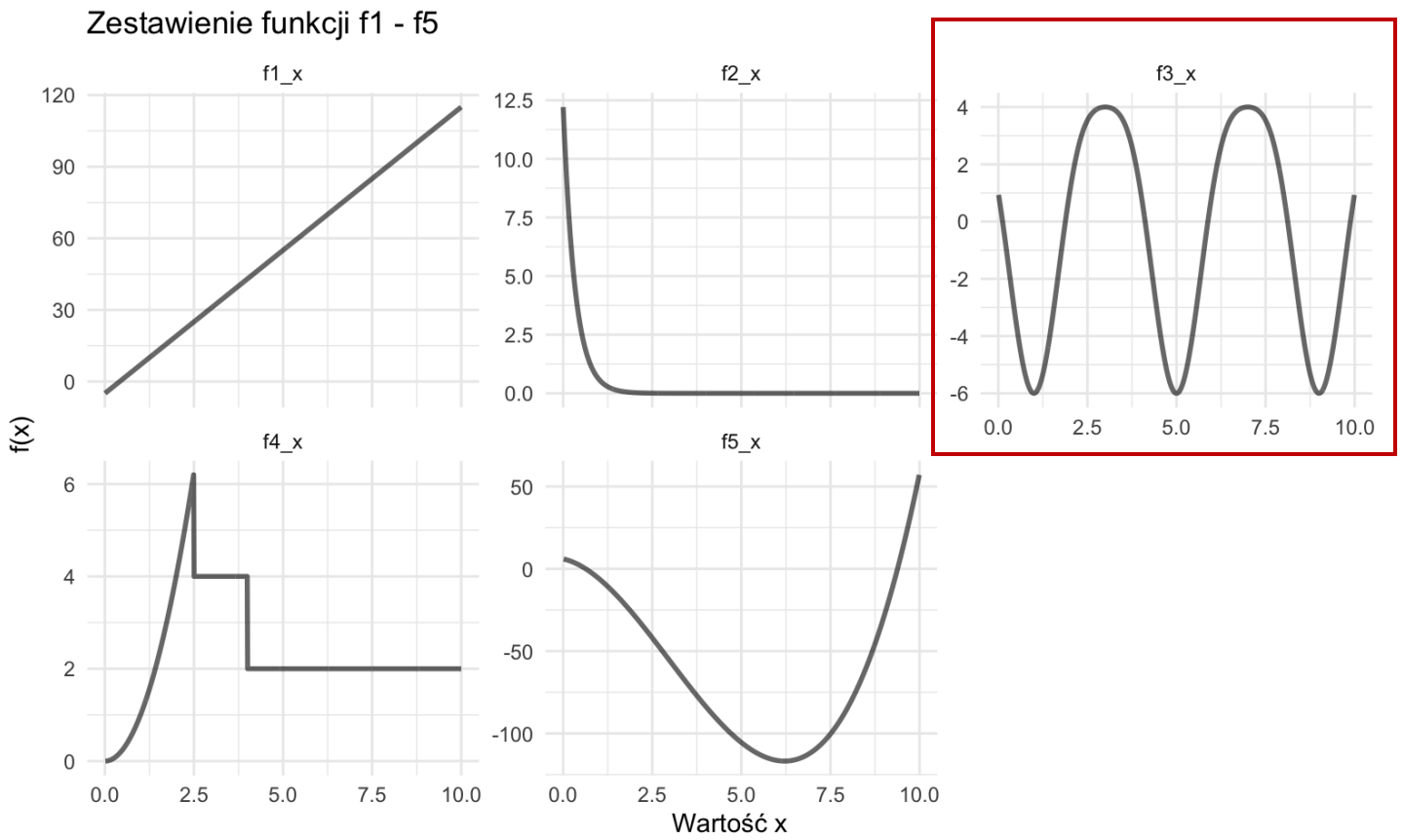
Zależność k-index od rozmiaru próby, metody estymacji i złożoności funkcji.
Wartości poniżej czerwonej przerywanej linii sugerują niewystarczający wymiar bazy k.

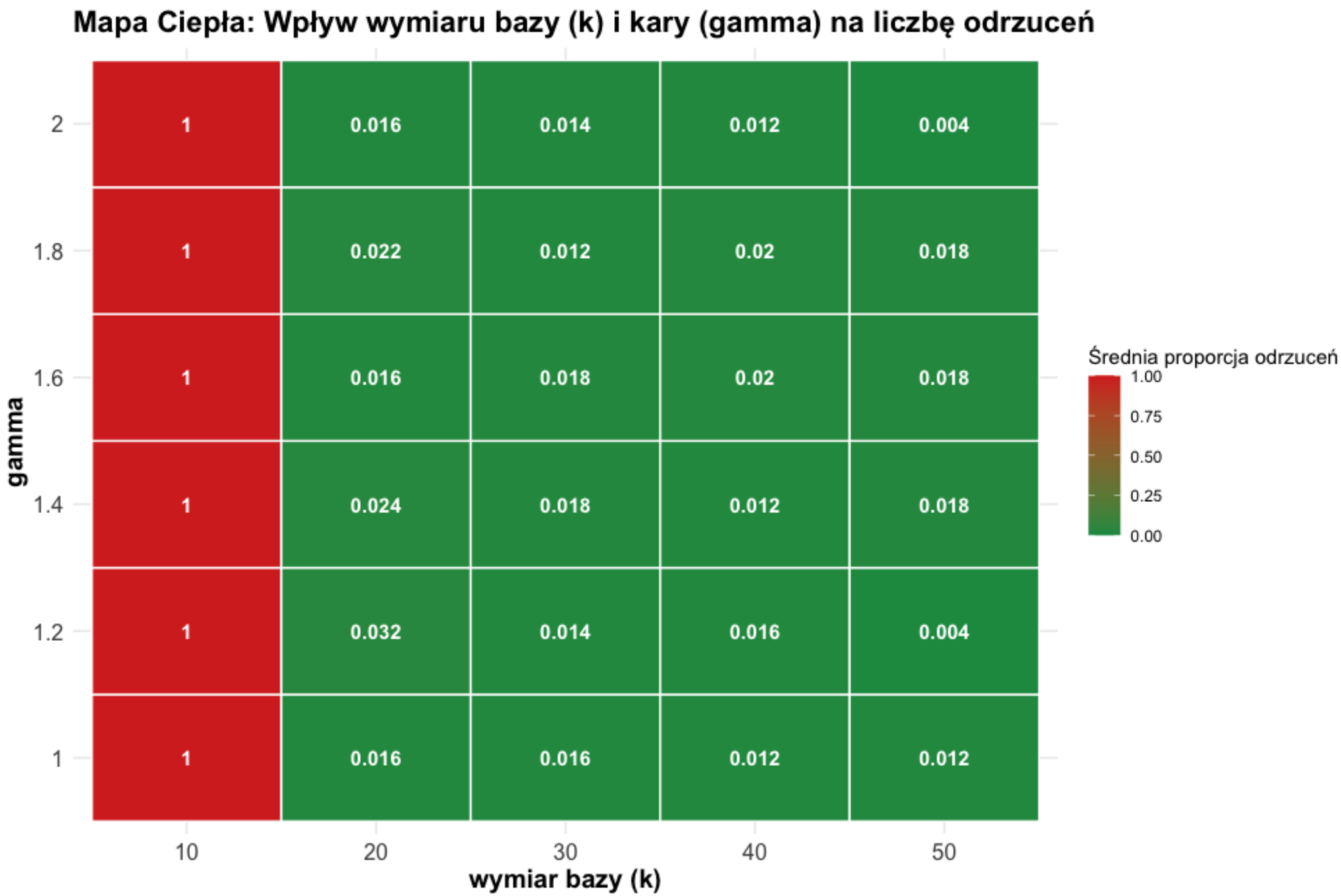


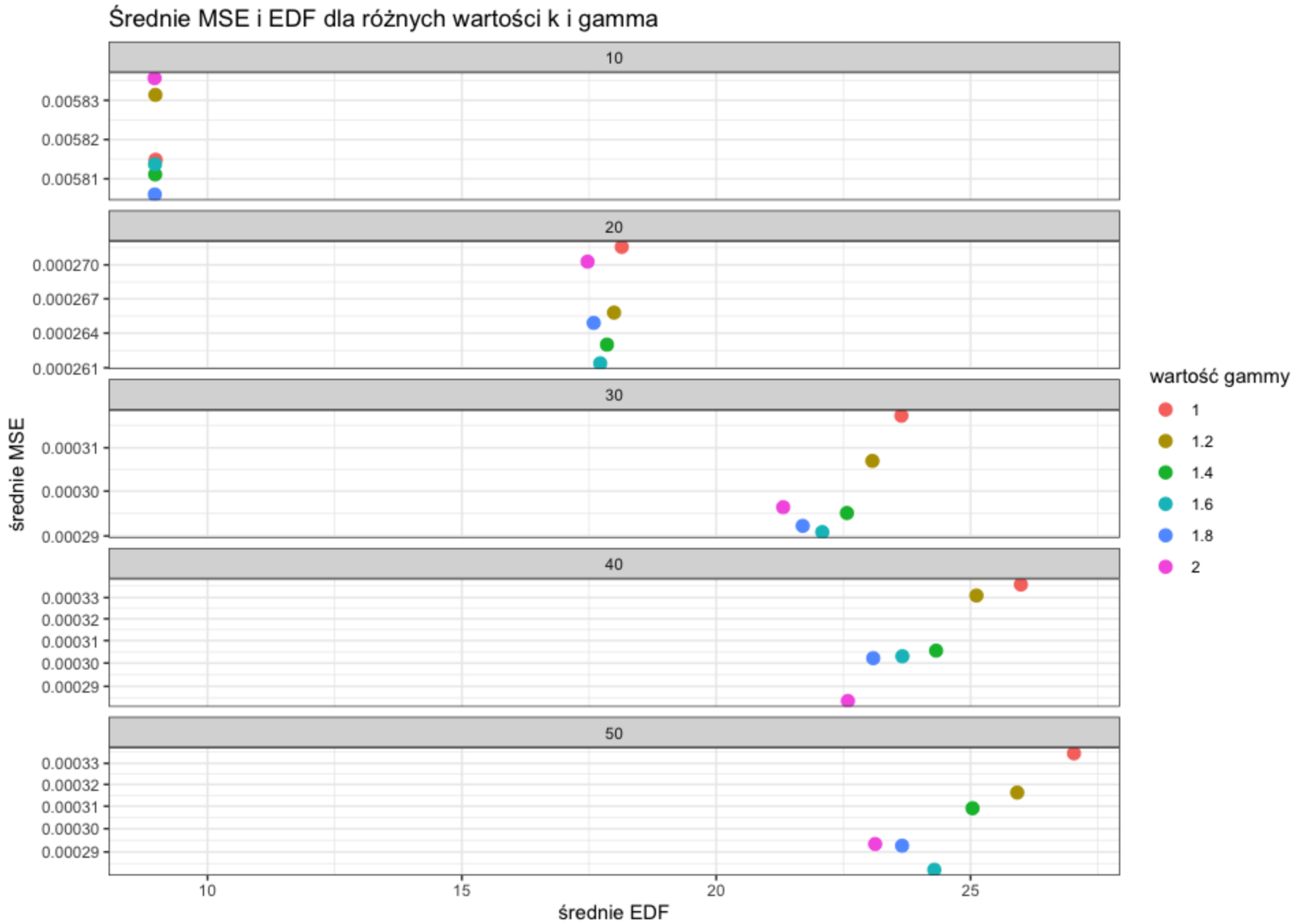

Dokładność estymacji kształtu funkcji (MSE)
Porównanie metod estymacji dla różnej wielkości próby
Metoda
REML
GCV.Cp
ML
f1
f2
f3
f4
f5
1e-03
1e-06
1e-09
1e-01
1e-02
1e-03
1e-04
1e-01
1e-02
1e-03
1e-04
0.100
0.010
0.001
1e-01
1e-02
1e-03
1e-04
1e-05
Log(MSE komponentu)
50
100
200
1000
5000
Rozmiar próby (n)


Dokładność estymacji stopni swobody (EDF)
Porównanie metod estymacji dla różnej wielkości próby
Metoda
REML
GCV.Cp
ML
f1
f2
f3
f4
f5
EDF
Rozmiar próby (n)
50
100
200
1000
5000

Analiza B - wpływ gammy i k na EDF




Mapa Ciepła: Wpływ wymiaru bazy (k) i kary (gamma) na liczbę odrzuceń
gamma
2
1.8
1.6
1.4
1.2
1
1 0.016 0.014 0.012 0.004
1 0.022 0.012 0.02 0.018
1 0.016 0.018 0.02 0.018
1 0.024 0.018 0.012 0.018
1 0.032 0.014 0.016 0.004
1 0.016 0.016 0.012 0.012
10
20
30
40
50
wymiar bazy (k)
Średnia proporcja odrzuceń
1.00
0.75
0.50
0.25
0.00


Średnie MSE i EDF dla różnych wartości k i gamma
10
20
30
40
50
0.00583
0.00582
0.00581
0.000270
0.000267
0.000264
0.000261
0.00031
0.00030
0.00029
0.00033
0.00032
0.00031
0.00030
0.00029
średnie MSE
10
15
20
25
średnie EDF
wartość gammy
1
1.2
1.4
1.6
1.8
2