

POLITECHNIKA KRAKOWSKA
WYDZIAŁ INŻYNIERII ELEKTRYCZNEJ I KOMPUTEROWEJ

Zenon Cyganek, Szymon Łukasik

**IDENTYFIKACJA ROZKŁADU
W SYSTEMACH RZECZYWISTYCH
ZA POMOCĄ ESTYMATORÓW
JĄDROWYCH**

Praca magisterska
Kierunek: elektrotechnika
Specjalność: automatyka

Opiekun naukowy:
dr hab. inż. Piotr Kulczycki, prof. PK

Kraków, 22 marca 2006

*"Engineers like to solve problems. If there are no problems handily available,
they will create their own problems."*

Scott Adams

Autorzy składają serdeczne podziękowania

Prof. Piotrowi Kulczyckiemu

za opiekę i pomoc podczas tworzenia niniejszej pracy,

Paniom: mgr Barbarze Łukasik, Joannie Kopeć, a także Panom:

prof. Januszowi Chwastowskiemu oraz dr. Konradowi Weinrebowi

za pomoc w gromadzeniu potrzebnych danych i cenne uwagi

SPIS TREŚCI

<i>Wstęp</i>	4
Rozdział 1. Preliminaria matematyczne	6
1.1. Podstawowe pojęcia statystyki	6
1.1.1. Charakterystyki liczbowe rozkładu zmiennej losowej	7
1.1.2. Charakterystyki funkcyjne rozkładu zmiennej losowej	8
1.2. Estymacja	9
1.2.1. Własności estymatorów	10
1.2.2. Estymacja parametryczna i nieparametryczna	11
1.2.3. Ocena jakości estymatora	13
1.3. Estymatory jądrowe	14
1.4. Dobór postaci jądra	17
1.5. Wyznaczanie parametru wygładzania	20
1.5.1. Metoda przybliżona	21
1.5.2. Metoda podstawień	22
1.5.3. Metoda krzyżowego uwiarygodniania	24
1.6. Metody poprawy jakości estymacji	26
1.6.1. Transformacja liniowa	27
1.6.2. Modyfikacja parametru wygładzania	29
1.6.3. Ograniczenie nośnika	32
Rozdział 2. Opis programu KDEstim	35
Rozdział 3. Estymacja rozkładów w wybranych systemach rzeczywistych	42
3.1. Charakterystyka cech gości górskiego pensjonatu	42
3.1.1. Analiza jednowymiarowa	43
3.1.2. Analiza dwuwymiarowa	44
3.1.3. Zastosowanie ograniczenia nośnika	46
3.1.4. Wpływ randomizacji danych na działanie algorytmu krzyżowego uwiarygodniania	49
3.2. Analiza wyników sondażu PGSS	52
3.3. Wykrywanie uszkodzeń silnika asynchronicznego	58
3.4. Analiza wybranych rozkładów z zakresu fizyki wysokich energii	69
3.4.1. Analiza procesu fotoprodukcji	69
3.4.2. Analiza zjawiska bremsstrahlungu	73
Rozdział 4. Podsumowanie	77
Dodatek A. Opis zawartości płyty CD	79
Dodatek B. Estymatory jądrowe w dostępnych narzędziach statystycznych	80
Spis rysunków i tabel	85
Bibliografia	88

WSTĘP

Estymacja funkcji gęstości rozkładu zmiennej losowej jest jednym z głównych zagadnień z zakresu stosowania metod statystycznych. Tradycyjne metody estymacji polegają na założeniu z góry ustalonego typu rozkładu zmiennej losowej, a następnie określeniu parametrów definiujących ten rozkład. Z tego też powodu ogólnie nazywa się je *metodami parametrycznymi*. Istnieje bogata literatura przedmiotowa opisująca algorytmy pomocne przy ich stosowaniu, jednak w obecnych czasach możliwości ich użycia do modelowania systemów rzeczywistych stają się coraz bardziej niewystarczające. Głównym powodem takiego stanu jest właśnie konieczność ograniczenia się w praktyce do kilkunastu dostępnych typów rozkładów. Sytuacja ta staje się jeszcze bardziej złożona w przypadku rozważania wielowymiarowych zmiennych losowych.

Wymienionej powyżej wady nie posiadają *metody nieparametryczne*, takie jak np. najprostsze *histogramy* czy, będące tematem niniejszej pracy, *estymatory jądrowe*. W metodach tych nie określa się z góry typu rozkładu charakteryzującego badane zmienne. Doboru funkcji opisujących rozkład i ich współczynników dokonuje się stosując odpowiednie kryteria optymalizacyjne. Kolejnym przyczyną coraz częstszego sięgania po metody nieparametryczne jest gwałtowny wzrost mocy obliczeniowej współczesnych komputerów, umożliwiającą rozwiązywanie w coraz krótszym czasie coraz bardziej złożonych problemów.

Celem niniejszej pracy jest przedstawienie zarysu wiedzy na temat estymatorów jądrowych, wykonanie programu pozwalającego, przy ich pomocy, na identyfikację rozkładu jedno- lub wielowymiarowej zmiennej losowej, a także użycie tego programu do określenia rozkładu kilku konkretnych systemów rzeczywistych.

W rozdziale pierwszym opisane zostały estymatory jądrowe, wraz z procedurami doboru typu jądra i parametru wygładzania. Ponadto rozdział ten zawiera metody poprawiające jakość estymacji jądrowej: transformację liniową, modyfikację parametru wygładzania oraz ograniczenie nośnika.

Rozdział drugi stanowi opis programu *KDEstim*, napisanego w ramach niniejszej pracy. Program ten posłużył autorom do obliczeń estymatorów jądrowych i wizualizacji otrzymanych wyników.

Rozdział trzeci charakteryzuje systemy rzeczywiste, wybrane do prezentacji możliwości estymatorów jądrowych. Rozdział ten zawiera również graficzną reprezentację wyników estymacji oraz ich analizę.

Podsumowanie pracy znajduje się w rozdziale czwartym.

Rysunki zawarte w pracy zostały wykonane w programach statystycznych *R 2.2.1*, *SigmaPlot 9.0*, *Statistica* oraz, w napisanym przez autorów, programie *KDEstim*.

Rozdział 1. PRELIMINARIA MATEMATYCZNE

Poniższy rozdział w zwięzły sposób przedstawia podstawowe wiadomości z zakresu statystyki oraz teorii estymacji. W kolejnych podrozdziałach zdefiniowano estymatory jądrowe, metody doboru postaci jądra oraz algorytmy wyznaczania parametru wygładzania. Wiadomości te stanowią podstawę pomocną w implementacji programu *KDEstim*. Niniejszy rozdział został opracowany na bazie dostępnej literatury przedmiotowej [7], [8], [9], [11], [12].

1.1. Podstawowe pojęcia statystyki

Podstawowym narzędziem badawczym nauk przyrodniczych, a częściowo nauk technicznych i społecznych jest *eksperyment*. Celem eksperymentu jest ustalenie prawa opisującego przebieg badanego zjawiska. Ze względu na odpowiedź badanego obiektu można wyróżnić dwa rodzaje eksperymentu:

- eksperyment o odpowiedzi zdeterminowanej – w którym dla tej samej wartości wejściowej X odpowiada zawsze ta sama (z zadaną dokładnością) reakcja Y ;
- eksperyment o odpowiedzi niezdeterminowanej – przy powtórzeniach eksperymentu, tej samej wartości X opowiadają różne wartości Y .

W drugiej grupie eksperymentów wyróżnia się eksperymenty, charakteryzujące się tym, że wraz ze wzrostem liczby powtórzeń eksperymentu, częstość pojawiania się danej odpowiedzi stabilizuje się. Są to *eksperymenty losowe*.

W modelu matematycznym opisującym eksperyment losowy, zbiorowi wszystkich możliwych wyników doświadczenia odpowiada przestrzeń zdarzeń elementarnych, oznaczana symbolem Ω . Elementy $\omega \in \Omega$ nazywa się *zdarzeniami elementarnymi*.

Zdarzenia to podzbiory przestrzeni zdarzeń elementarnych Ω , z góry wyróżnione przez obserwatora. Zbiór zdarzeń Σ musi spełniać następujące postulaty:

$$1. \phi \in \Sigma; \quad (1.1)$$

$$2. \text{jeśli } A \in \Sigma \text{ to } A' \in \Sigma; \quad (1.2)$$

$$3. \text{jeśli } A_1, A_2, \dots \in \Sigma \text{ to } \bigcup_{n=1}^{\infty} A_n \in \Sigma. \quad (1.3)$$

Rodzinę Σ spełniającą powyższe własności nazywa się σ -ciałem lub σ -algebrą zbiorów.

Zmienną losową nazywamy funkcję X , określoną na przestrzeni zdarzeń elementarnych Ω , o wartościach rzeczywistych, taką, że dla każdego $t \in \mathbb{R}^n$ zbiór

$$\{\omega \in \Omega : X(\omega) < t\} \quad (1.4)$$

jest zdarzeniem (czyli należy do Σ). Dla $n = 1$ mamy do czynienia z rzeczywistą zmienną losową, w przeciwnym wypadku z n -wymiarową zmienną losową.

Dla dowolnego $A \in \Sigma$ można wyznaczyć *prawdopodobieństwo* zdarzenia A oznaczane przez $P(A)$. Prawdopodobieństwo, stanowiące *wartość miary probabilistycznej*, musi spełniać następujące aksjomaty:

$$1. P(\emptyset) = 0; \quad (1.5)$$

$$2. 0 \leq P(A) \leq 1 \text{ dla każdego } A \in \Sigma; \quad (1.6)$$

$$3. P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n), \text{ dla dowolnego ciągu parami rozłącznych zdarzeń } A_1, A_2, \dots, A_n. \quad (1.7)$$

Trójka (Ω, Σ, P) nazywana jest *przestrzenią probabilistyczną*.

1.1.1. Charakterystyki liczbowe rozkładu zmiennej losowej

Wartość oczekiwana zmiennej losowej X to liczba $E(X)$ określona wzorem:

$$E(X) = \int_{\Omega} X(\omega) dP(\omega) \quad (1.8)$$

Wariancją zmiennej losowej X o wartości oczekiwanej $E(X)$ nazywa się liczbę $V(X)$, równą:

$$V(X) = \int_{\Omega} (X(\omega) - E(X))^2 dP(\omega) = E(X - E(X))^2. \quad (1.9)$$

Wariancja jest miarą rozrzutu. Jej wadą jest to, że jej wymiar jest kwadratem wymiaru zmiennej losowej X . Aby pozbyć się tej niedogodności definiuje się *odchylenie standardowe* $\sigma(X)$ jako

$$\sigma(X) = \sqrt{V}. \quad (1.10)$$

Kowariancja zmiennej losowej X to macierz $Cov(X)$ o wymiarach $n \times n$, o wyrazach:

$$\begin{aligned} c_{i,j}(X) &= \int_{\Omega} (X_i(\omega) - E(X_i))(X_j(\omega) - E(X_j)) dP(\omega) = \\ &= E[(X_i - E(X_i))(X_j - E(X_j))] \end{aligned} \quad (1.11)$$

Na głównej przekątnej tej macierzy znajdują się wyrazy

$$c_{ii}(X) = E[(X_i - E(X_i))(X_i - E(X_i))] = E(X_i - E(X_i))^2 = V(X). \quad (1.12)$$

1.1.2. Charakterystyki funkcyjne rozkładu zmiennej losowej

Funkcję $F: \mathbb{R}^n \rightarrow [0,1]$, której wartość określa wzór

$$F(u) = P(\{\omega \in \Omega : X_1(\omega) < u_1, \dots, X_n(\omega) < u_n\}), \quad (1.13)$$

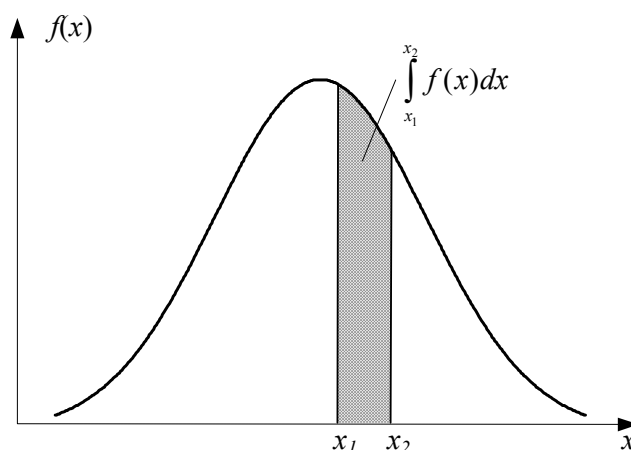
nazywa się *dystrybuantą* zmiennej losowej X . Dystrybuanta określa prawdopodobieństwo przyjęcia przez zmienną losową wartości mniejszej niż $u = [u_1, u_2, \dots, u_n]$.

Zmienna losowa X jest *absolutnie ciągła*, jeżeli istnieje nieujemna funkcja f , określona i spełniająca warunek $\int_{-\infty}^{\infty} f(x) dx = 1$, taka, że dla każdego przedziału $[x_1, x_2]$ zachodzi:

$$P(\{\omega : x_1 \leq X(\omega) \leq x_2\}) = \int_{x_1}^{x_2} f(x) dx. \quad (1.14)$$

Funkcja f jest nazywana *gęstością rozkładu prawdopodobieństwa zmiennej losowej* X . Jej interpretację graficzną dla zmiennej jednowymiarowej przedstawia rysunek 1.1

(zgodnie ze wzorem (1.14) całka $\int_{x_1}^{x_2} f(x)dx$ określa prawdopodobieństwo, że zmienna losowa X przyjmuje wartości z przedziału $[x_1, x_2]$). Wyznaczanie gęstości rozkładu prawdopodobieństwa jest jednym z głównych zagadnień niniejszej pracy.



Rysunek 1.1. Gęstość prawdopodobieństwa rozkładu normalnego

Modyfikując daną gęstość f w skończonej liczbie punktów, otrzymamy nową funkcję f_1 , która również spełnia równość (1.14). Zatem gęstość nie jest określona jednoznacznie. Zależność pomiędzy dystrybuantą a gęstością prawdopodobieństwa jest następująca:

$$f(x) = \frac{\partial F(x)}{\partial x}. \quad (1.15)$$

1.2. Estymacja

Z punktu widzenia statystyki matematycznej otrzymane w wyniku eksperymentu wartości określonych parametrów rozkładu są jedynie *estymatorami parametrów* zmiennej losowej w populacji generalnej. *Estymacją* nazywa się dział statystyki zajmujący się wyznaczaniem parametrów populacji generalnej za pośrednictwem wartości tych parametrów, które zostały wyznaczone w *populacji próbnej*. Wartości parametrów rozkładu w populacji generalnej nazywa się *wartościami prawdziwymi*, natomiast ich przybliżenia otrzymywane z próbki reprezentatywnej są *estymatorami*.

1.2.1. Własności estymatorów

Dla danego parametru φ można utworzyć wiele estymatorów $\hat{\varphi}_n(x_1, x_2, \dots, x_n)$, lecz dla uzyskania estymatora o wymaganych własnościach, pożądane jest, aby miał on pewne, z góry narzucone, cechy. Zrozumiałe jest to, że należy wymagać, aby ze wzrostem liczebności próbki wzrastała dokładność oszacowania parametru φ . Wymaganie to prowadzi do spełnienia dla każdej liczby $\varepsilon > 0$ warunku:

$$\lim_{n \rightarrow \infty} P(|\hat{\varphi}_n - \varphi| \geq \varepsilon) = 0. \quad (1.16)$$

Warunek ten oznacza, że dla dostatecznie dużych liczebności próby estymator przyjmuje z dowolnie dużym prawdopodobieństwem wartości bliskie estymowanemu parametrowi φ . Estymator $\hat{\varphi}_n(x_1, x_2, \dots, x_n)$ spełniający powyższe równanie nosi nazwę *estymatora zgodnego*.

Obciążenie $B_n(\varphi)$ estymatora to różnica

$$B_n(\varphi) = E(\hat{\varphi}_n) - \varphi. \quad (1.17)$$

Gdy wartość oczekiwana estymatora jest równa wartości rzeczywistej parametru φ , wówczas obciążenie jest równe zero a estymator nazywa się *nieobciążonym*. W przeciwnym razie estymator jest *obciążony*. Obciążenie estymatora określa wokół jakiej wielkości „oscylować” będą uzyskiwane wartości estymatora. W praktyce pożądane jest również, by wielkość owych oscylacji (określana przez wariancję) dążyła do zera, gdy liczebność próby dąży do nieskończoności, czyli by prawdziwa była zależność:

$$\lim_{n \rightarrow \infty} V(\hat{\varphi}_n) = 0. \quad (1.18)$$

Estymator asymptotycznie nieobciążony to estymator, spełniający warunek:

$$\lim_{n \rightarrow \infty} E(\hat{\varphi}_n) = \varphi. \quad (1.19)$$

Dla danego parametru φ może istnieć więcej niż jeden estymator nieobciążony. Jeżeli zatem $\hat{\varphi}_n^*$ i $\hat{\varphi}_n^{**}$ są dwoma estymatorami nieobciążonymi parametru φ , o wariancjach $V(\hat{\varphi}_n^*)$ oraz $V(\hat{\varphi}_n^{**})$ spełniających nierówność:

$$V(\hat{\varphi}_n^*) < V(\hat{\varphi}_n^{**}), \quad (1.20)$$

czyli, że skupienie wartości estymatora $\hat{\varphi}_n^*$ wokół φ jest większe niż skupienie wartości $\hat{\varphi}_n^{**}$, to mówimy, że $\hat{\varphi}_n^*$ jest *estymatorem efektywniejszym* parametru φ niż estymator $\hat{\varphi}_n^{**}$.

Wobec powyższego, estymator nieobciążony $\hat{\varphi}_n(x_1, x_2, \dots, x_n)$ parametru φ , który ma najmniejszą wariancję spośród wszystkich nieobciążonych estymatorów danego parametru φ , wyznaczonych z prób n -elementowych, nosi nazwę *estymatora najefektywniejszego*.

1.2.2. Estymacja parametryczna i nieparametryczna

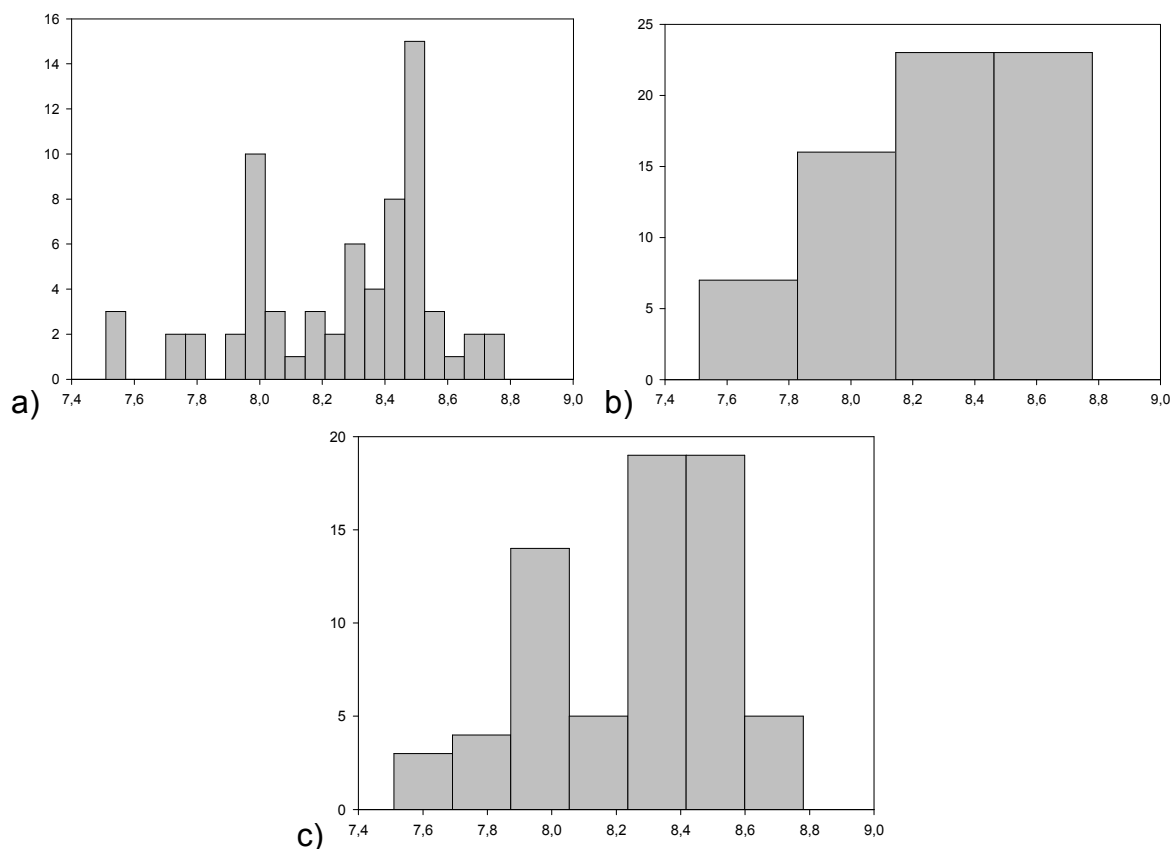
Klasyczne metody estymacji polegają na wyborze jednego z kilkunastu typowych rozkładów prawdopodobieństwa, a następnie dobraniu właściwych parametrów dopasowujących ten rozkład do badanych danych. Najpopularniejsze z tych rozkładów to normalny, jednostajny, trójkątny, beta, gamma, wykładniczy, Weibulla, t -Studenta, χ^2 (chi kwadrat), Poissona. Obecnie jednak, wraz z rozwojem techniki, wybór spośród tych rozkładów staje się coraz bardziej niewystarczający. W związku z tym rozpoczęły się poszukiwania nowych, doskonalszych metod, które nie byłyby obciążone przymusem arbitralnego wyboru konkretnego typu rozkładu – tzw. *metod nieparametrycznych*. Rozwój tych metod staje się coraz bardziej dynamiczny w związku z ciągłym wzrostem mocy obliczeniowej współczesnych komputerów.

Najprostszą nieparametryczną metodą estymacji jest *histogram*. Polega ona na podziale zbioru, w którym zawierają się wszystkie próby losowe x_1, x_2, \dots, x_m , na przedziały H_k o jednakowej szerokości h . Dla każdego przedziału definiuje się histogram jako funkcję równą liczbie tych wartości próby losowej, które należą do tego przedziału (oznaczanej jako $\#\{x_i \in H_k\}$) podzielonej przez mh (normalizacja):

$$\hat{f}(x) = \frac{\#\{x_i \in H_k\}}{mh} \quad \text{dla każdego } k \text{ całkowitego.} \quad (1.21)$$

Zasadnicze znaczenie dla właściwości histogramu ma wybór szerokości przedziałów h . Zbyt duża jej wartość powoduje ukrycie charakterystycznych cech rozkładu, z kolei zbyt mała jej wartość daje wiele nieistniejących maksimów i minimów lokalnych. Rysunek 1.2 przedstawia trzy histogramy (nieznormalizowane) dla tej samej zmiennej losowej, charakteryzującej się dwoma maksimami (modami). Histogramy te różnią się jedynie

szerokością przedziału h . Jak widać na podstawie tego przykładu, histogram jest bardzo wrażliwy na wartość tego parametru. Podobnie silną wrażliwość wykazują histogramy względem „punktu zaczepienia” (położenia początku pierwszego z przedziałów).



Rysunek 1.2. Wpływ doboru szerokości przedziałów h na własności histogramu: a) zbyt mała, b) zbyt duża, c) optymalna wartość

Kolejną wadą histogramu jest jego nieciągłość oraz fakt, że jego pochodna jest równa zero we wszystkich punktach (z wyjątkiem punktu styku przedziałów, gdzie w ogóle nie istnieje), co bardzo utrudnia jego analizę.

Na korzyść histogramu przemawia jedynie jego prostota oraz łatwość interpretacji – z tego powodu jest dość skutecznym narzędziem wizualizującym, szczególnie polecanym w początkowej fazie analizy.

Spośród innych metod nieparametrycznych wyróżnić można *estymatory najbliższego sąsiedztwa*, *estymator Fouriera*, metodę funkcji sklepanych (tzw. *splinów*). Jednak obecnie podstawową metodą nieparametryczną są *estymatory jądrowe*. Cechuje je naturalność konstrukcji, łatwość interpretacji i analizy matematycznej. Ponieważ estymatory te są tematem niniejszej pracy, zostały one szczegółowo opisane w osobnym rozdziale (1.3).

1.2.3. Ocena jakości estymatora

Jedną z najczęściej stosowanych metod oceny jakości estymatorów jest *kryterium błędu średniokwadratowego* - *MSE* (ang. Mean Square Error). Jeżeli dla rzeczywistego parametru b wyznaczony został jego estymator \hat{b} , wówczas wartością tego kryterium jest wartość oczekiwana kwadratu błędu estymacji:

$$MSE = E\left((\hat{b} - b)^2\right). \quad (1.22)$$

Po przekształceniach można otrzymać wzór, określający błąd średniokwadratowy jako sumę kwadratu obciążenia estymatora \hat{b} i jego wariancji:

$$MSE = \left[E(\hat{b}) - b\right]^2 + V(\hat{b}). \quad (1.23)$$

Z punktu widzenia niniejszej pracy interesujący jest przypadek jakości estymacji funkcji gęstości prawdopodobieństwa f n -wymiarowej zmiennej losowej. Wówczas wyżej wymienione kryterium może być stosowane przy ustalonej wartości x :

$$MSE_x = E\left((\hat{f}(x) - f(x))^2\right) \quad \text{dla dowolnego } x \in \mathbb{R}^n, \quad (1.24)$$

bądź równoważnie:

$$MSE_x = \left[E(\hat{f}(x)) - f(x)\right]^2 + V(\hat{f}(x)) \quad \text{dla dowolnego } x \in \mathbb{R}^n. \quad (1.25)$$

Kryterium błędu średniokwadratowego określa, jaki błąd popełniany jest podczas estymacji tylko dla jednego, danego x . W celu określenia całkowitej jakości estymacji definiuje się globalny wskaźnik *MISE* (*Mean Integrated Square Error*) – tzw. *kryterium scałkowanego błędu średniokwadratowego*, będący całką wartości *MSE* na całej przestrzeni \mathbb{R}^n :

$$MISE = \int_{\mathbb{R}^n} E\left((\hat{f}(x) - f)^2\right) dx, \quad (1.26)$$

bądź równoważnie:

$$MISE = \int_{\mathbb{R}^n} \left[\left[E(\hat{f}(x)) - f(x)\right]^2 + V(\hat{f}(x)) \right] dx. \quad (1.27)$$

Do określenia jakości estymacji gęstości prawdopodobieństwa f dogodniejsza jest druga postać tego kryterium.

1.3. Estymatory jądrowe

W celu zdefiniowania estymatora jądrowego, zakłada się daną n -wymiarową zmienną losową X o funkcji gęstości rozkładu f . Wówczas wynikiem m niezależnych eksperymentów, jest m -elementowa próba losowa x_1, x_2, \dots, x_m , na podstawie której wyznaczyć można estymator $\hat{f}: \mathbb{R}^n \rightarrow [0, \infty)$ funkcji gęstości rozkładu f zmiennej losowej X .

Estymator jądrowy definiowany jest wzorem:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^m K\left(\frac{x - x_i}{h}\right), \quad (1.28)$$

gdzie:

$m \in \mathbb{N} \setminus \{0\}$ to liczność próby losowej,

$h > 0$ nazywany jest *parametrem wygładzania*.

$K: \mathbb{R}^n \rightarrow [0, \infty)$ to funkcja symetryczna względem zera i posiadająca w tym punkcie słabe maksimum, a więc spełniająca poniższe warunki:

$$1. \int_{\mathbb{R}^n} K(x) dx = 1 \quad (1.29)$$

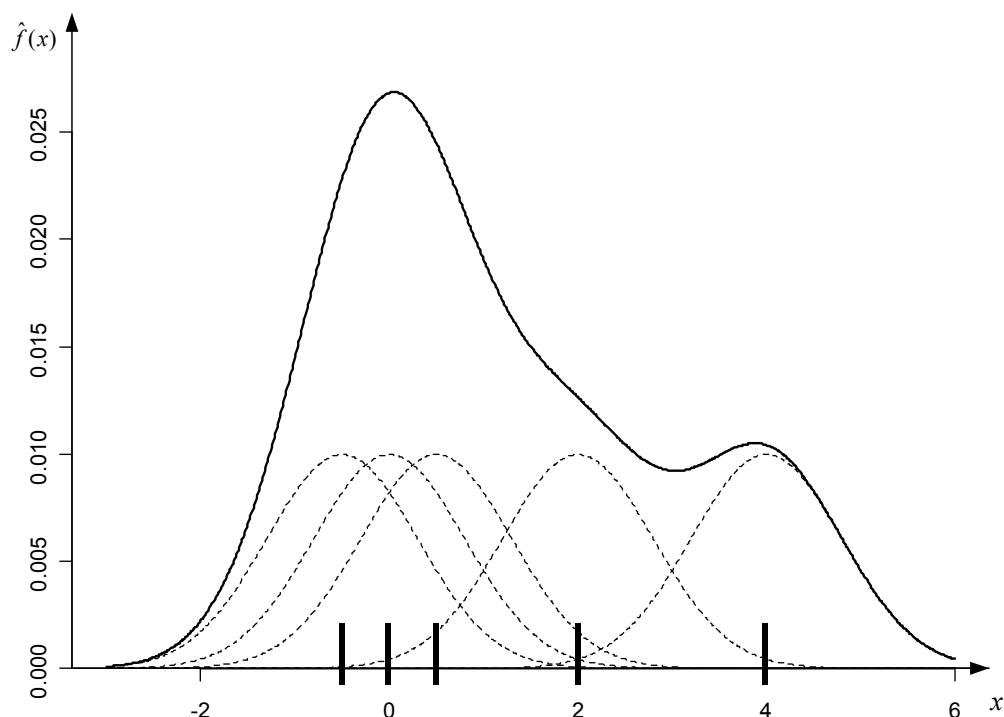
$$2. K(x) = K(-x) \text{ dla każdego } x \in \mathbb{R}^n \quad (1.30)$$

$$3. K(0) \geq K(x). \quad (1.31)$$

Odwzorowanie K jest nazywane *jądrem* (ang. *kernel*) jako analogia do jądra operatora całkowego. Najczęściej stosowane typy jąder przedstawia tabela 1.1.

Rysunek 1.3 ilustruje definicję estymatora jądrowego dla jednowymiarowej zmiennej losowej. Dla pojedynczego eksperymentu x_i , estymatorem rozkładu zmiennej losowej X jest funkcja K przesunięta o wektor x_i oraz przeskalowana za pomocą współczynnika h . W przypadku m wyników niezależnych eksperymentów x_1, x_2, \dots, x_m staje się on sumą pojedynczych oszacowań. Jednak, aby otrzymana funkcja spełniała warunek $\int_{\mathbb{R}^n} f(x) dx = 1$, stawiany mierze probabilistycznej, przed tą sumą znajduje się

współczynnik $\frac{1}{mh^n}$.



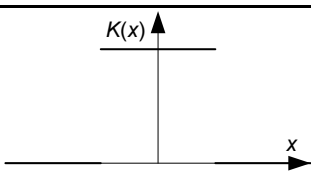
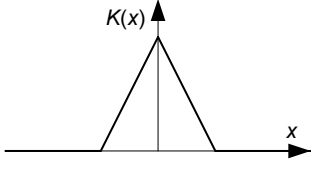
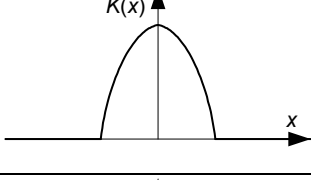
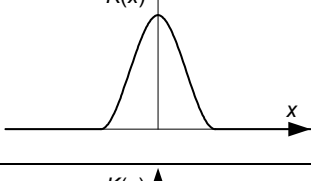
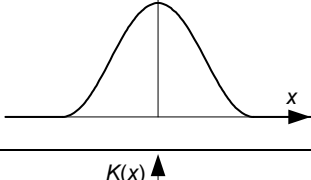
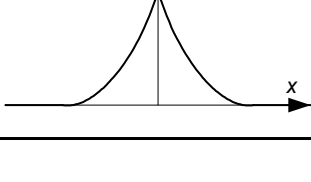
Rysunek 1.3. Interpretacja graficzna jednowymiarowego jądrowego estymatora gęstości prawdopodobieństwa / $m = 5$ /

Estymator jądrowy pozwala wyznaczyć gęstość bez konieczności ograniczenia się do arbitralnie założonego rozkładu. Z tego powodu estymatory te warto stosować w przypadku niestandardowych rozkładów gęstości, dla których metody parametryczne są nieskuteczne, np. w przypadkach rozkładów wielomodalnych. Dzięki niezależności od konkretnego typu rozkładu możliwe jest określenie, wielu właściwości funkcji f , takich jak położenie modów, symetria, czy zachowanie rozkładu dla skrajnych wartości zmiennej losowej.

Wyniki otrzymywane przy pomocy estymatorów jądrowych dają możliwość łatwej interpretacji, szczególnie jeśli chodzi o wartość modalne. Zazwyczaj, w przypadku analizy rzeczywistych systemów technicznych, ekonomicznych, czy socjologicznych, mody stanowią informację o różnych aspektach rozważanego problemu. Wyodrębnienie ich umożliwia lepsze poznanie badanego obiektu, znalezienie nowych czynników mających wpływ na jego zachowanie, diagnozę czy detekcję uszkodzeń.

Kluczową kwestią w stosowaniu estymatorów jądrowych jest właściwa odpowiedź na pytanie: jak dobrać postać jądra oraz wartość parametru h , tak by estymator jądrowy najlepiej oddawał charakter badanej populacji? Jest to problem optymalizacji, który można rozwiązać na podstawie kryterium scalkowanego błędu średniokwadratowego (opisanego w poprzednim rozdziale). Zagadnieniem wyboru typu jądra oraz wartości parametru wygładzania poświęcono kolejne dwa podrozdziały niniejszej pracy.

Tabela 1.1. Najczęściej stosowane typy jąder

Typ jądra	Radialna postać jądra $K(x)$	Wykres (dla $n = 1$)
Jednostajne	$\frac{1}{k_n}$ dla $\ x\ \leq 1$ 0 dla $\ x\ > 1$	
Trójkątne	$\frac{n(n+1)}{k_n}(1-\ x\)$ dla $\ x\ \leq 1$ 0 dla $\ x\ > 1$	
Epanechnikowa (kwadratowe)	$\frac{n+2}{2k_n}(1-\ x\ ^2)$ dla $\ x\ \leq 1$ 0 dla $\ x\ > 1$	
Dwuwagowe	$\frac{(n+2)(n+4)}{8k_n}(1-\ x\ ^2)^2$ dla $\ x\ \leq 1$ 0 dla $\ x\ > 1$	
Normalne	$(2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\ x\ ^2}$	
Gamma	$\frac{1}{k_n \Gamma(n)} e^{-\ x\ }$	
<p>Oznaczenia:</p> <ul style="list-style-type: none"> k_n - objętość n-wymiarowej kuli jednostkowej zadanej równaniami: $k_n = \begin{cases} \frac{(2\pi)^{\frac{1}{2}}}{2 \cdot 4 \cdot \dots \cdot n} & \text{dla } n \text{ parzystych} \\ \frac{2(2\pi)^{\frac{n-1}{2}}}{1 \cdot 3 \cdot \dots \cdot n} & \text{dla } n \text{ nieparzystych} \end{cases}$ Γ - funkcja Eulera $\Gamma: C \rightarrow C$ opisana wzorem $\Gamma(z) = \int_0^{\infty} e^{-x} x^{z-1} dx$; 		

1.4. Dobór postaci jądra

Podczas doboru postaci jądra, jak również doboru parametru wygładzania h , korzysta się z kryterium minimalizacji scałkowanego błędu średniokwadratowego MISE, opisanego w rozdziale 1.2.3. Do określenia wartości MISE czyni się wówczas poniższe założenia:

- gęstość f ma ciągłą drugą pochodną i jest ona całkowalna z kwadratem,
- funkcja K spełnia warunek:

$$\int_{-\infty}^{\infty} x^2 K(x) dx < \infty .$$

Dodatkowo dogodnie jest zdefiniować następujące zależności:

$$U(K) = \int_{\mathbb{R}^n} x^2 K(x) dx \quad (1.32)$$

$$W(K) = \int_{\mathbb{R}^n} K(x)^2 dx \quad (1.33)$$

$$Z(f) = \int_{\mathbb{R}^n} [\nabla^2 f(x)]^2 dx \quad (1.34)$$

gdzie:

$$\nabla^2 f(x) = \sum_{i=1}^n \frac{\partial^2 f(x)}{\partial x_i^2} . \quad (1.35)$$

Wówczas konieczna do obliczenia MISE wartość obciążenia wynosi:

$$E(\hat{f}(x)) - f(x) = \frac{1}{2} h^2 U(K) \nabla^2 f(x) + o(h^2) \quad (1.36)$$

zaś wartość wariancji jest równa:

$$V(\hat{f}(x)) = \frac{1}{mh^n} W(K) f(x) + o\left(\frac{1}{mh^n}\right) . \quad (1.37)$$

W powyższych wzorach funkcja $o(h^2)$ spełnia warunek:

$$\lim_{h \rightarrow 0} \frac{o(h^2)}{h^2} = 0 , \quad (1.38)$$

natomiast $o\left(\frac{1}{mh^n}\right)$:

$$\lim_{h \rightarrow 0} \frac{o\left(\frac{1}{mh^n}\right)}{\frac{1}{mh^n}} = 0 . \quad (1.39)$$

Analizując powyższe równania, można stwierdzić, że aby obciążenie estymatora dążyło do zera wraz ze wzrostem liczności próby losowej, parametr wygładzania h należy dobierać tak by spełniał warunek

$$\lim_{m \rightarrow \infty} h = 0, \quad (1.40)$$

natomiast, aby wariancja estymatora zmierzała do zera – parametr wygładzania powinien być dobierany tak, by zachodziła zależność:

$$\lim_{m \rightarrow \infty} \frac{1}{mh^n} = 0. \quad (1.41)$$

Spełnienie powyższych warunków w dowolnym punkcie x zapewnia, że estymator jądrowy jest asymptotycznie nieobciążony i jego wariancja maleje do zera wraz ze wzrostem liczności próby losowej. Wobec tego wartość MSE_x maleje wówczas do zera.

Scałkowany błąd średniokwadratowy MISE po podstawieniu wzorów (1.36) oraz (1.37) wynosi

$$MISE = \frac{1}{4} h^4 U(K)^2 Z(f) + \frac{1}{mh^n} W(K) + o(h^4) + o\left(\frac{1}{mh^n}\right). \quad (1.42)$$

Po uwzględnieniu warunków (1.38), (1.39), (1.40) oraz (1.41) otrzymuje się:

$$MISE = \frac{1}{4} h^4 U(K)^2 Z(f) + \frac{1}{mh} W(K). \quad (1.43)$$

Z powyższego równania wynika, że zmniejszanie parametru wygładzania h w celu zmniejszenia pierwszego składnika sumy (obciążenia) powoduje zwiększenie drugiego (wariancji), natomiast odwrotnie jest w przypadku zwiększania parametru h . Konieczne jest więc znalezienie kompromisu pomiędzy wariancją a obciążeniem, tak aby zminimalizować wartość scałkowanego błędu średniokwadratowego MISE. Minimum to występuje dla

$$h_0 = {}^{n+4}\sqrt{\frac{nW(K)}{U(K)^2 Z(f)m}}, \quad (1.44)$$

a jego wartość jest równa

$$MISE = d(n) {}^{n+4}\sqrt{\frac{U(K)^{2n} W(K)^4 Z(f)^n}{m^4}}. \quad (1.45)$$

gdzie nieistotna z punktu widzenia niniejszych rozważań stała $d(n)$ wynosi:

$$d(n) = \frac{4+n}{4n^{n/(n+4)}}. \quad (1.46)$$

Wzór (1.45) jest pomocny przy wyborze jądra. Funkcjonał średniokwadratowy jest proporcjonalny do wyrażenia $\sqrt[n+4]{U(K)^{2n}W(K)^4}$, które jest najmniejsze w przypadku jądra Epanecznikowa. Dla innych jąder wyrażenie to przyjmuje wartość niewiele większą, ze względu na wpływ pierwiastka stopnia $(n+4)$. W przypadku jąder, które przedstawia tabela 1.1, zmniejszenie efektywności innych jąder w stosunku do jądra Epanecznikowa, dla $n = 1$, wynosi od 1 do 7 %.

Dla przypadków wielowymiarowych problem wyboru postaci jądra rozszerza się, poprzez możliwość stosowania dwóch postaci każdego z jąder: jądra radialnego bądź produktowego.

Jądro radialne definiuje się w praktyce następująco:

$$K^n(x) = c^n K^1(\sqrt{x^T x}), \quad (1.47)$$

gdzie K^1 oznacza jądro jednowymiarowe, natomiast c^n stałą tak dobraną, aby jądro K^n spełniało warunek (1.29). $\sqrt{x^T x}$ to *norma euklidesowa* czyli długość odcinka w n -wymiarowej przestrzeni:

$$\sqrt{x^T x} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \|x\|. \quad (1.48)$$

Wzory najczęściej stosowanych typów jąder radialnych przedstawia tabela 1.1.

Wartość jądra radialnego jest stała na każdym okręgu o środku w punkcie zero – jest to tzw. *symetria radialna*.

Jądro produktowe opisane jest wzorem:

$$K^n(x) = K^n \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right) = K^1(x_1) \cdot K^1(x_2) \cdot \dots \cdot K^1(x_n). \quad (1.49)$$

Wartość n -wymiarowego jądra K^n jest w tym przypadku iloczynem (produktem) jednowymiarowych jąder K^1 dla każdej ze współrzędnych.

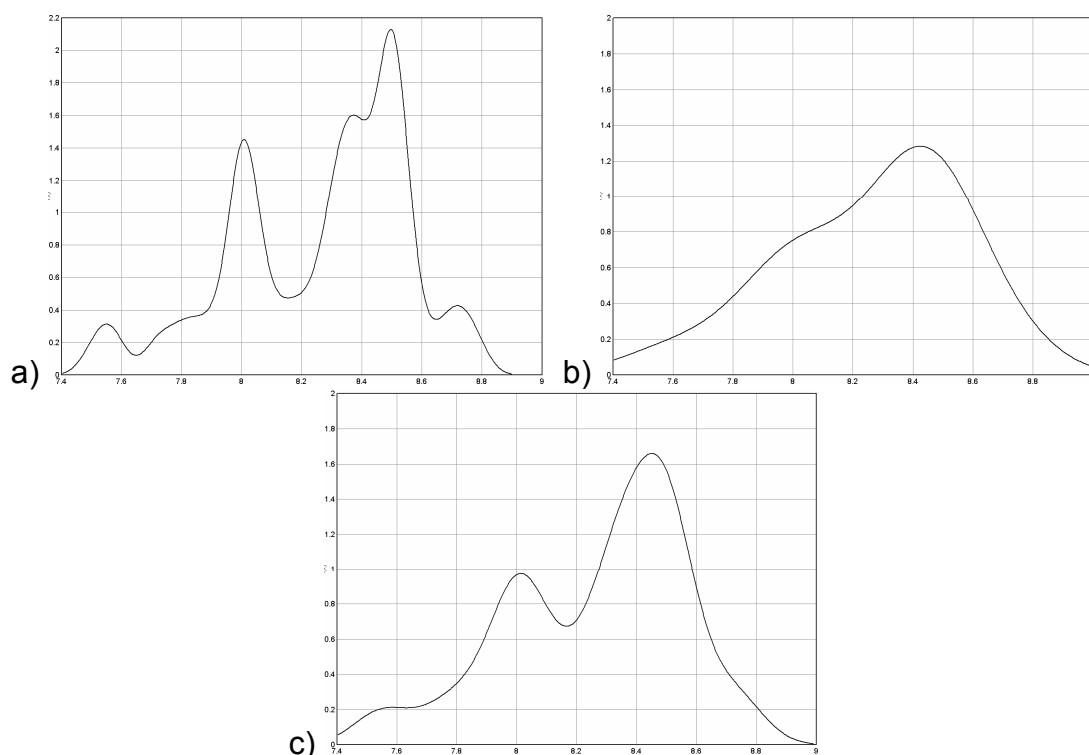
Z kryterium scałkowanego błędu kwadratowego wynika, że zarówno wśród jąder radialnych i produktowych najefektywniejsze jest jądro Epanecznikowa. Dodatkowo porównując parami jądra typu radialnego i produktowego, można zaobserwować w każdym przypadku większą efektywność jąder radialnych. Ponadto, różnica ta zwiększa się w miarę wzrostu wymiaru zmiennej losowej. Za stosowaniem jąder produktowych

przemawia z kolei łatwość implementacji – całkowanie i różniczkowanie tych jąder jest zbliżone do przypadku jednowymiarowego.

Podsumowując, z powodu niewielkiej różnicy efektywności poszczególnych jąder, bardzo często przy wyborze typu jądra można kierować się innymi własnościami estymatora takimi jak: klasa regularności, ograniczoność nośnika czy prostota obliczeniowa. W związku z tym w niniejszej pracy zdecydowano się na stosowanie radialnego jądra normalnego. Wybór tego jądra podyktowany był jego dużą efektywnością i łatwością implementacji.

1.5. Wyznaczanie parametru wygładzania

Jakość estymacji jądrowej w największym stopniu zależy od wartości parametru wygładzania h . Wpływ ten ilustruje rysunek 1.4.



Rysunek 1.4. Wpływ parametru wygładzania h na estymator jądrowy: a) zbyt mała wartość, b) zbyt duża wartość, c) optymalna wartość parametru wygładzania

Na rysunku tym przedstawiono estymatory jądrowe zmiennej losowej (rozkład tej samej zmiennej został przedstawiony na histogramach - rysunek 1.2) różniące się parametrem h . Gdy parametr wygładzania jest za mały (rysunek 1.4a), pojawia się wiele ekstremów lokalnych (w skrajnym przypadku może ich być m – dla każdego elementu próby), co jest sprzeczne z rzeczywistymi właściwościami realnych populacji. W przypadku, gdy parametr h przyjmuje zbyt duże wartości (rysunek 1.4b), estymator jądrowy cechuje się

nadmiernym wygładzeniem. Obie sytuacje uniemożliwiają wyciąganie prawidłowych wniosków na temat badanego przy pomocy estymatora jądrowego rozkładu zmiennej losowej.

Powyższe rozważania stanowią pewną analogię do zagadnień doboru szerokości h przedziałów dla histogramu. O ile jednak w jego przypadku brak jest skutecznych metod doboru wartości h , o tyle dla parametru wygładzania estymatora jądrowego istnieje wiele takich metod. Niniejszy rozdział przedstawia trzy najbardziej popularne: metodę przybliżoną, podstawień oraz krzyżowego uwiarygodniania.

1.5.1. Metoda przybliżona

W metodzie tej wykorzystuje się równanie (1.44), które określa na podstawie kryterium MISE optymalną wartość parametru wygładzania. Jak łatwo zauważyć, jego wartość zależna jest od estymowanej, czyli nieznannej funkcji gęstości f , poprzez wyrażenie $Z(f)$. Aby wyznaczyć przybliżoną wartość h , można podstawić wartość funkcjonau $Z(f)$ odpowiadającą standardowemu rozkładowi normalnemu. Dla n -wymiarowej zmiennej losowej przyjmuje on postać:

$$Z(f) = \frac{n(n+2)}{4 \cdot 2^n \pi^{\frac{n}{2}}}. \quad (1.50)$$

Obliczając odpowiednie całki otrzymuje się dla radialnego jądra normalnego:

$$\frac{W(K)}{U(K)^2} = \frac{1}{2^n \pi^{\frac{n}{2}} n^2}. \quad (1.51)$$

Po wstawieniu powyższych zależności do równania (1.44), otrzymuje się gotowy wzór na przybliżony (suboptymalny) parametr wygładzania h dla jądra normalnego:

$$h_0^* = \sqrt[n+4]{\frac{4}{n^2(n+2)m}} \cdot \prod_{i=1}^n \hat{\sigma}_i, \quad (1.52)$$

gdzie $\hat{\sigma}_i$ oznaczają estymatory odchylenia standardowego poszczególnych współrzędnych.

Metoda ta jest najszybszą z wszystkich przedstawionych, daje jednak tylko przybliżoną wartość parametru wygładzania. Otrzymany wynik jest tym lepszy, im bardziej badany rozkład jest zbliżony do rozkładu normalnego. Ze względu na powyższe fakty, metoda ta ma znaczenie jedynie do wstępnego oszacowania parametru wygładzania, np. w celu wykorzystania w innych, dokładniejszych, lecz bardziej złożonych procedurach.

1.5.2. Metoda podstawień

Metoda podstawień (ang. *plug-in*), polega na minimalizacji wartości kryterium scałkowanego błędu średniokwadratowego MISE. O ile jednak daje ona najlepsze wyniki, to jej wykorzystanie jest ograniczone jedynie do rozkładów zmiennych losowych jednowymiarowych.

Metoda plug-in wykorzystuje zależność (1.44). Tak jak w poprzedniej metodzie, pojawia się tu więc problem z określeniem wartości funkcjonału $Z(f)$. Dla zmiennej jednowymiarowej przyjmuje on postać:

$$Z(f) = \int_{\mathbb{R}^n} f''(x)^2 dx. \quad (1.53)$$

Ponieważ w wyrażeniu tym występuje druga pochodna funkcji gęstości f , w pierwszej fazie tej metody tworzony jest estymator jądrowy f'' . W celu wyznaczenia dla niego parametru wygładzania, należy w następnym etapie stworzyć estymator jądrowy czwartej pochodnej $f^{(4)}$. Po wykonaniu k takich kroków, aby wyznaczyć parametr wygładzania dla estymatora jądrowego funkcji $f^{(2k)}$, należy skorzystać z opisanej powyżej metody przybliżonej.

Aby zmniejszyć niedokładność powodowaną przez zastosowanie metody przybliżonej, ilość kroków k nie powinna być zbyt mała. Z drugiej jednak strony duża ilość kroków wymaga estymacji pochodnych wysokich rzędów, co obarczone jest znacznym błędem. Z praktycznego punktu widzenia ilość kroków k powinna być nie mniejsza niż 2, a najlepiej gdy $k = 2$. Dla ustalonego kroku k , metodę tą nazywa się *metodą podstawień k -tego rzędu*.

W celu bezpośredniego wykorzystania metody *plug-in* k -tego rzędu, wprowadza się pomocniczy parametr $\xi \in \{4, 6, \dots, 2k + 4\}$, dla którego stała $c_\xi \in \mathbb{R}$ wynosi:

$$c_\xi = \frac{\xi!}{(0,5\xi)! \sqrt{\pi} (2\hat{\sigma})^{\xi+1}}, \quad (1.54)$$

natomiast stała $C_{\xi,h} \in \mathbb{R}$ dla $h > 0$ przyjmuje wartość:

$$C_{\xi,h} = \frac{1}{m^2 h^{\xi+1}} \sum_{i=1}^m \sum_{j=1}^m \tilde{K}^{(\xi)} \left(\frac{x_i - x_j}{h} \right). \quad (1.55)$$

W powyższych wzorach $\hat{\sigma}$ oznacza estymator odchylenia standardowego (1.10), \tilde{K} - jądro $(2k + 2)$ -krotnie różniczkowalne, natomiast x_i oraz $x_j - i$ -tą i j -tą wartość próby losowej.

W niniejszej pracy rozważona zostanie metoda podstawień drugiego rzędu. Postępowanie w przypadku wyższych rzędów jest bardzo podobne.

W sytuacji, gdy $k = 2$ stała ξ wynosi 8. Wówczas wzór (1.54) przyjmuje postać:

$$c_8 = \frac{105}{32\sqrt{\pi}\hat{\sigma}^9}. \quad (1.56)$$

Po wyznaczeniu wartości c_8 wykonuje się kolejne dwa kroki (przyjmując numerację odwrotną od naturalnej). Najpierw wyznacza się wartość:

$$h_{II} = \left(\frac{-2\tilde{K}^{(6)}(0)}{U(\tilde{K})C_8 m} \right)^{\frac{1}{9}}, \quad (1.57)$$

przy czym dla normalnego, najczęściej stosowanego, jądra \tilde{K} , zachodzi:

$$U(\tilde{K}) = 1, \quad (1.58)$$

$$\tilde{K}^{(6)}(0) = -\frac{15}{\sqrt{2\pi}}, \quad (1.59)$$

Następnie wyznacza się wartość:

$$h_I = \left(\frac{-2\tilde{K}^{(4)}(0)}{U(\tilde{K})C_{6,h_{II}} m} \right)^{\frac{1}{7}}, \quad (1.60)$$

gdzie:

$$\tilde{K}^{(4)}(0) = \frac{3}{\sqrt{2\pi}}, \quad (1.61)$$

a wartość $C_{6,h_{II}}$ oblicza się ze wzoru (1.55), w którym:

$$\tilde{K}^{(6)}(x) = \frac{1}{\sqrt{2\pi}} (x^6 - 15x^4 + 45x^2 - 15) e^{-\frac{1}{2}x^2}. \quad (1.62)$$

Poszukiwaną wartość $Z(f)$ oblicza się na podstawie:

$$Z(f) = C_{4,h_I}, \quad (1.63)$$

korzystając ze wzoru (1.55) i przyjmując, że:

$$\tilde{K}^{(4)}(x) = \frac{1}{\sqrt{2\pi}} (x^4 + 6x^2 - 3) e^{-\frac{1}{2}x^2} \quad (1.64)$$

Po wyznaczeniu wartości $Z(f)$ można już bez przeszkód skorzystać ze wzoru (1.44) i wyznaczyć optymalną wartość parametru wygładzania h .

Metoda podstawień umożliwia szybkie wyznaczenie parametru wygładzania, niestety jedynie dla jednowymiarowych zmiennych losowych. W przypadku zmiennej losowej wielowymiarowej może być stosowana tylko w połączeniu z jądrem produktowym, gdzie n -krotnie oblicza się parametr wygładzania, osobno dla każdej współrzędnej.

1.5.3. Metoda krzyżowego uwiarygodniania

W metodzie krzyżowego uwiarygodniania (ang. *cross-validation*) ponownie korzysta się z kryterium scałkowanego błędu średniokwadratowego MISE. W tym przypadku, wyznacza się wartość h_0 realizującą minimum funkcji $g: \mathbb{R} \rightarrow [0, \infty)$, zdefiniowanej równością

$$g(h) = \frac{1}{m^2 h^n} \sum_{i=1}^m \sum_{j=1}^m \tilde{K}\left(\frac{x_j - x_i}{h}\right) + \frac{2}{mh^n} K(0), \quad (1.65)$$

gdzie x_i oraz x_j oznaczają i -tą i j -tą wartość próby losowej, natomiast

$$\tilde{K}(x) = K^{*2}(x) - 2K(x), \quad (1.66)$$

przy czym K^{*2} oznacza kwadrat splotowy funkcji K , tzn.

$$K^{*2}(x) = \int_{\mathbb{R}^n} K(y)K(x-y)dy. \quad (1.67)$$

Dla jądra normalnego kwadrat ten wyraża się wzorem

$$K^{*2}(x) = (4\pi)^{-\frac{n}{2}} e^{-\frac{\|x\|^2}{4}} \quad (1.68)$$

Minimalizacji funkcji (2.65) można dokonać przy użyciu numerycznych metod optymalizacji, bądź po prostu znaleźć najmniejszą wartość tej funkcji w zadanym przedziale z ustalonym krokiem. Po znalezieniu minimum czynność tą można powtórzyć zmniejszając zakres oraz krok, w celu dokładniejszego ustalenia minimum funkcji g .

W pierwszej fazie odpowiedni wydaje się zakres $[\frac{1}{4}h_0^*, 4h_0^*]$ i krok $\frac{h_0^*}{100}$, gdzie h_0^* oznacza przybliżony parametr wygładzania.

Aby ograniczyć ilość iteracji, w których poszukiwane jest minimum funkcji g można skorzystać z tzw. *metody złotego podziału*.

W przeważającej części przypadków zdefiniowana w metodzie krzyżowego uwiarygodniania funkcja g przypomina kształtem parabolę o niesymetrycznych ramionach. Metoda złotego podziału polega na stworzeniu zstępującego ciągu przedziałów $[a_0, b_0] \supset [a_1, b_1] \supset \dots \supset [a_k, b_k]$, dla których zachodzi zależność:

$$\frac{b_i - a_i}{b_{i-1} - a_{i-1}} = \frac{\sqrt{5} - 1}{2} \cong 0,618 \text{ dla } i = 1, 2, \dots, k. \quad (1.69)$$

Podział odcinka w takim stosunku jest powszechnie znany pod nazwą *złotego podziału*, stąd też nazwę wzięła opisywana metoda.

Uwzględniając powyższe informacje algorytm postępowania w metodzie złotego podziału jest następujący:

1. Przed rozpoczęciem obliczeń należy określić przedział $[a_0, b_0]$ określający zakres poszukiwań oraz liczby rzeczywiste

$$p_0 = b_0 - \frac{\sqrt{5} - 1}{2}(b_0 - a_0), \quad (1.70)$$

$$q_0 = a_0 + \frac{\sqrt{5} - 1}{2}(b_0 - a_0) \quad (1.71)$$

oraz wartości $g(p_0)$ i $g(q_0)$.

2. W każdym następnym i -tym kroku ($i = 1, 2, \dots, k$, natomiast $k \in \mathbb{IN} \setminus \{0\}$ to założona liczba kroków) wykonuje się porównanie wartości $g(p_{i-1})$ i $g(q_{i-1})$. Wtedy:

- jeżeli $g(p_{i-1}) \leq g(q_{i-1})$, wówczas

$$a_i = a_{i-1}, \quad (1.72)$$

$$b_i = q_{i-1}, \quad (1.73)$$

$$p_i = b_i - \frac{\sqrt{5} - 1}{2}(b_i - a_i), \quad (1.74)$$

$$q_i = p_{i-1}, \quad (1.75)$$

następnie wyznacza się wartość $g(p_i)$

oraz dokonuje podstawienia

$$g(q_i) = g(p_{i-1}); \quad (1.76)$$

- natomiast jeżeli $g(p_{i-1}) > g(q_{i-1})$, wówczas

$$a_i = p_{i-1}, \quad (1.77)$$

$$b_i = b_{i-1}, \quad (1.78)$$

$$p_i = q_{i-1}, \quad (1.79)$$

$$q_i = a_i + \frac{\sqrt{5}-1}{2}(b_i - a_i), \quad (1.80)$$

następnie dokonuje się podstawienia

$$g(p_i) = g(q_{i-1}) \quad (1.81)$$

oraz wyznacza wartość $g(q_i)$.

3. Na zakończenie w ostatnim, k -tym, kroku wyznacza się wartość parametru h :

- jeżeli $g(p_{k-1}) \leq g(q_{k-1})$, wówczas

$$h = \frac{a_k + q_k}{2}; \quad (1.82)$$

- natomiast jeżeli $g(p_{k-1}) > g(q_{k-1})$, wówczas

$$h = \frac{p_k + b_k}{2}. \quad (1.83)$$

Jeżeli funkcja g ma w przedziale $[a_0, b_0]$ tylko jedno minimum, to jest ono zawarte w każdym z kolejnych przedziałów, a w szczególności w ostatnim $[a_k, b_k]$. Wyznaczanie minimum funkcji g tą metodą pozwala znacznie zmniejszyć czas obliczeń (wartość funkcji jest obliczana dwukrotnie we wstępnym kroku oraz jednokrotnie w każdym z następnych kroków). Aby uzyskać dokładność wyznaczenia współczynnika wygładzania $h_0^*/100$ metodą złotego podziału wystarczy wykonać co najmniej 10 kroków.

1.6. Metody poprawy jakości estymacji

Podczas praktycznego stosowania estymatorów jądrowych okazuje się, że w wielu przypadkach samo wykorzystywanie wyżej opisanych metod doboru typu jądra oraz wielkości parametru wygładzania nie pozwala na oddanie specyficznych cech rozkładu. Wymagane jest wówczas dobranie własności estymatora do konkretnego rozważanego problemu.

Kolejne podrozdziały zawierają opis procedur pozwalających lepiej dopasować estymator do rzeczywistych cech rozważanego rozkładu. Metody te obejmują:

transformację liniową, modyfikację parametru wygładzania i ograniczenie nośnika badanej zmiennej losowej.

1.6.1. Transformacja liniowa

Parametr wygładzania h jednakowo wpływa na poszczególne składowe rozważanej zmiennej losowej. Ponieważ w ogólnym przypadku skale współrzędnych tych składowych mogą być różne, zatem optymalna wartość parametru wygładzania dla niektórych składowych może okazać się za mała, a dla niektórych za duża. Z tego względu bardzo często pomocne może być zastosowanie transformacji liniowej

$$Y \equiv \sqrt{R^{-1}} X, \quad (1.84)$$

w której macierz R jest macierzą kowariancji:

$$R = \text{Cov}(X). \quad (1.85)$$

Pierwiastek macierzy R^{-1} to taka macierz dodatnio półokreślona, która pomnożona przez samą siebie jest równa R^{-1} . Analitycznie pierwiastek taki uzyskuje się przekształcając macierz kowariancji do postaci kanonicznej Jordana, będącej w tym przypadku macierzą diagonalną, następnie obliczając jej pierwiastek i stosując przekształcenie odwrotne.

Uproszczona (diagonalna) wersja transformacji (1.84) polega na ograniczeniu macierzy R do jej diagonalii:

$$R = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_n \end{bmatrix}, \quad (1.86)$$

gdzie V_1, V_2, \dots, V_n to wariancje kolejnych składowych zmiennej losowej X .

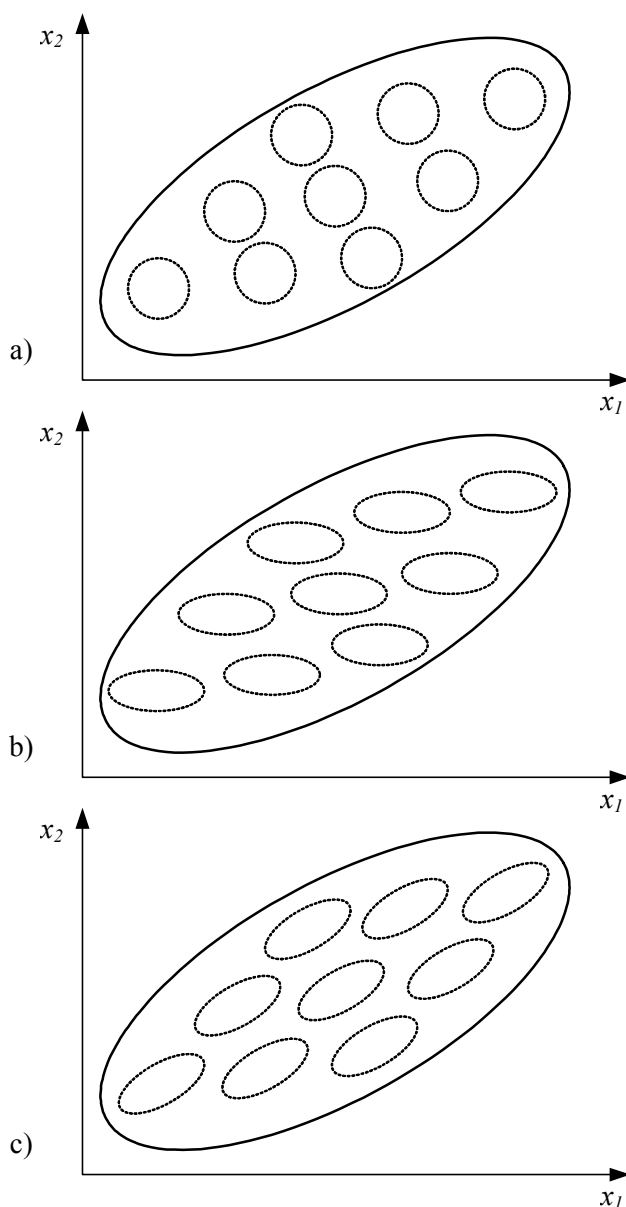
W praktyce można założyć, że macierz R , zarówno dla postaci pełnej jak i postaci diagonalnej, jest nieosobliwa.

Stosowanie transformacji (1.84) uogólnia definicję jądra radialnego (1.47) do postaci:

$$K^n(x) = \frac{c^n}{\sqrt{\det(R)}} K^1\left(\sqrt{x^T R^{-1} x}\right) \quad (1.87)$$

równoważnej wzorowi (1.47), gdyby R było macierzą jednostkową.

Wzór (1.84) definiuje nową, n -wymiarową zmienną losową. W przypadku postaci diagonalnej, wariancje wszystkich jej współrzędnych są równe jedności. Jeżeli stosowana jest postać pełna, współrzędne te są również liniowo niezależne. Dzięki powyższym własnościom kształt radialnych jąder jest lepiej dopasowany do kształtu badanego rozkładu.



Rysunek 1.5. Interpretacja transformacji liniowej: a) brak transformacji, b) postać diagonalna, c) postać pełna

Sytuację tą ilustruje rysunek 1.5. Przedstawia on poziomice funkcji gęstości prawdopodobieństwa przykładowej, dwuwymiarowej zmiennej losowej wraz z poziomiami poszczególnych jąder (linia przerywana). W przypadku braku transformacji liniowej (rysunek 1.5a) poziomice jąder są okręgami. Gdy zastosowano transformację diagonalną (rysunek 1.5b), poziomice te stają się elipsami o osiach równoległych do osi

współrzędnych. Pełna transformacja liniowa (rysunek 1.5c) pozwala w pełni ukierunkować owe elipsy względem „rozciągnięcia” rozkładu gęstości.

W poniższej pracy, w celu uproszczenia obliczeń, ograniczono się do postaci diagonalnej omawianej transformacji. W tym przypadku, mając daną próbę losową, do wyznaczenia estymatora jądrowego stosuje się następujące wzory:

$$R^{-1} = \begin{bmatrix} \frac{1}{\hat{V}_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\hat{V}_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\hat{V}_n} \end{bmatrix}, \quad (1.88)$$

$$\det(R) = \hat{V}_1 \cdot \hat{V}_2 \cdot \dots \cdot \hat{V}_n, \quad (1.89)$$

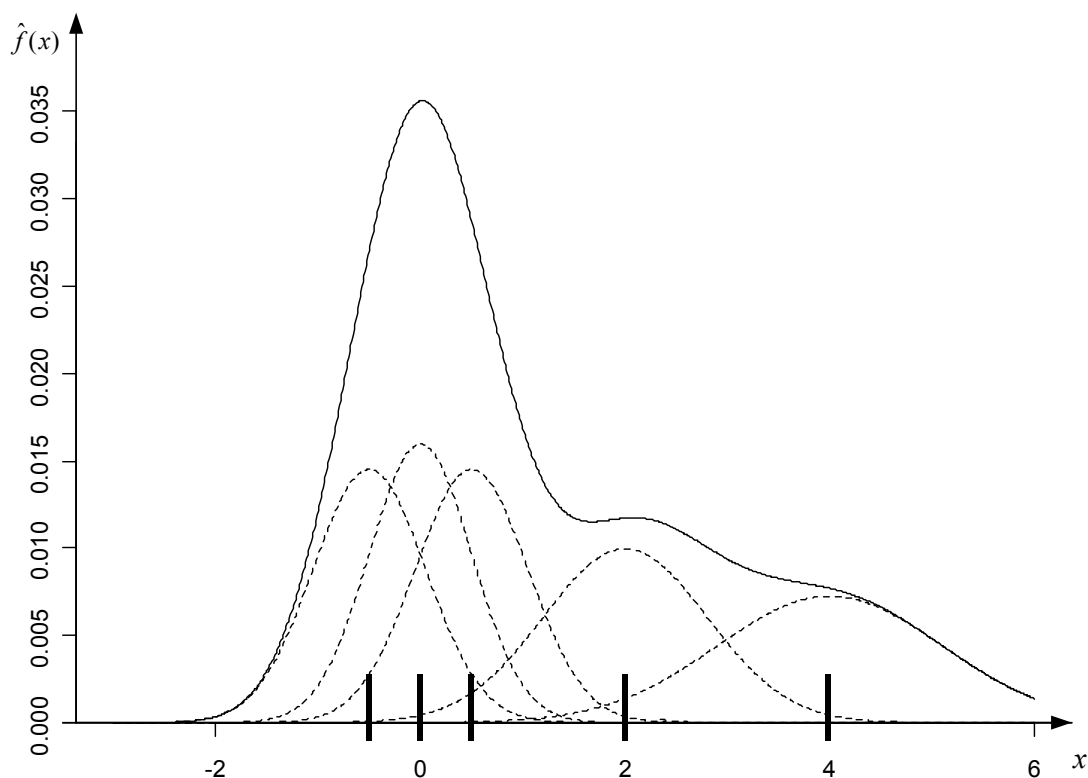
natomiast wzór estymatora jądrowego przyjmuje postać:

$$\hat{f}(x) = \frac{1}{mh^n \sqrt{\det(R)}} \sum_{i=1}^m K \left(\sqrt{\left[\frac{x-x_i}{h} \right]^T R^{-1} \left[\frac{x-x_i}{h} \right]} \right). \quad (1.90)$$

Podsumowując, transformacja liniowa mimo swej prostoty jest bardzo efektywną metodą. Jej celem jest dopasowanie kształtu jądra do badanego rozkładu, a wyniki dzięki niej otrzymywane zgadzają się z intuicyjnymi rozważaniami. Ze względu na wyżej wymienione cechy, dla jąder radialnych, zalecane jest stosowanie tej metody przynajmniej w postaci diagonalnej.

1.6.2. Modyfikacja parametru wygładzania

W przypadku znacznej części systemów rzeczywistych ekstremalne wartości zmiennej losowej są mało prawdopodobne. W związku z małą ilością próbek, w obszarze tym powstają niepożądane zniekształcenia estymatora jądrowego - co nie odpowiada własnościom rzeczywistych systemów. Celem zmodyfikowanego parametru wygładzania jest dodatkowe wygładzenie estymatora tam, gdzie przyjmuje on małą wartość, natomiast w obszarach o dużych wartościach – jego wyostrenie. Koncepcję tą ilustruje rysunek 1.6. Rysunek ten warto porównać z rysunkiem 1.3 uzyskanym dla tych samych danych.



Rysunek 1.6. Konceptcja zmodyfikowanego parametru wygładzania

Kolejność działań w przypadku obliczeń estymatora jądrowego ze zmodyfikowanym parametrem wygładzania jest następująca:

1. Zdefiniowanie estymatora jądrowego \hat{f} według wcześniejszych wzorów;
2. Określenie parametrów modyfikujących $s_i > 0$ dla każdego jądra (każdego elementu próby losowej) postaci

$$s_i = \left(\frac{\hat{f}(x_i)}{\bar{s}} \right)^{-c}, \quad (1.91)$$

gdzie:

- parametr $c \in [0,1]$; na podstawie kryterium minimum błędu średniokwadratowego c dobierane jest zwykle jako równe $1/2$;
- \bar{s} to średnia geometryczna liczb $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_m)$ obliczana ze wzoru

$$\ln(\bar{s}) = \frac{1}{m} \sum_{i=1}^m \ln(\hat{f}(x_i)); \quad (1.92)$$

3. Wyznaczenie estymatora jądrowego ze zmodyfikowanym parametrem wygładzania za pomocą wzoru

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^m \frac{1}{s_i^n} K\left(\frac{x-x_i}{hs_i}\right). \quad (1.93)$$

Wzór (1.93) jest uogólnieniem wzoru (1.28), które otrzymuje się z (1.93), przyjmując $c = 0$ (a więc $s_i = 1$).

Analizując rysunek 1.6 można zaobserwować, że w obszarze dużej wartości gęstości (w przykładzie przedział $[-1,1]$) wartości estymatorów poszczególnych próbek są większe od średniej geometrycznej wszystkich estymatorów. Zatem, zgodnie ze wzorem (1.91) ich parametry modyfikujące są większe od 1, co powoduje „wyszczuplenie” jąder w tamtym obszarze. Odwrotnie, w obszarach małej wartości gęstości (przedział $[3, \infty)$) wartości estymatorów są mniejsze od średniej, co skutkuje parametrami modyfikującymi mniejszymi od 1 oraz „wygładzeniem” estymatorów.

Istnieje też oczywiście możliwość zastosowania modyfikacji parametru wygładzania w połączeniu z transformacją liniową. Definicja takiego estymatora jądrowego, otrzymana z połączenia wzorów (1.87) i (1.93), jest następująca:

$$\hat{f}(x) = \frac{1}{mh^n \sqrt{\det(R)}} \sum_{i=1}^m \frac{1}{s_i^n} K\left(\sqrt{\begin{bmatrix} x-x_i \\ hs_i \end{bmatrix}^T R^{-1} \begin{bmatrix} x-x_i \\ hs_i \end{bmatrix}}\right). \quad (1.94)$$

Zaletami estymatora ze zmodyfikowanym parametrem wygładzania są:

- zmniejszona wrażliwość na nieprawidłowy dobór stałej h ;
- zwiększenie efektywności poszczególnych typów jąder względem jądra Epanecznikowa;
- w przypadku wielowymiarowym zwiększenie efektywności jądra produktowego względem radialnego;
- prostota obliczeniowa.

Ze względu na wyżej wymienione zalety koncepcji modyfikacji parametru wygładzania, metoda ta powinna być stosowana w większości przypadków, szczególnie przy korzystaniu z mniej efektywnych jąder oraz z jąder produktowych.

1.6.3. Ograniczenie nośnika

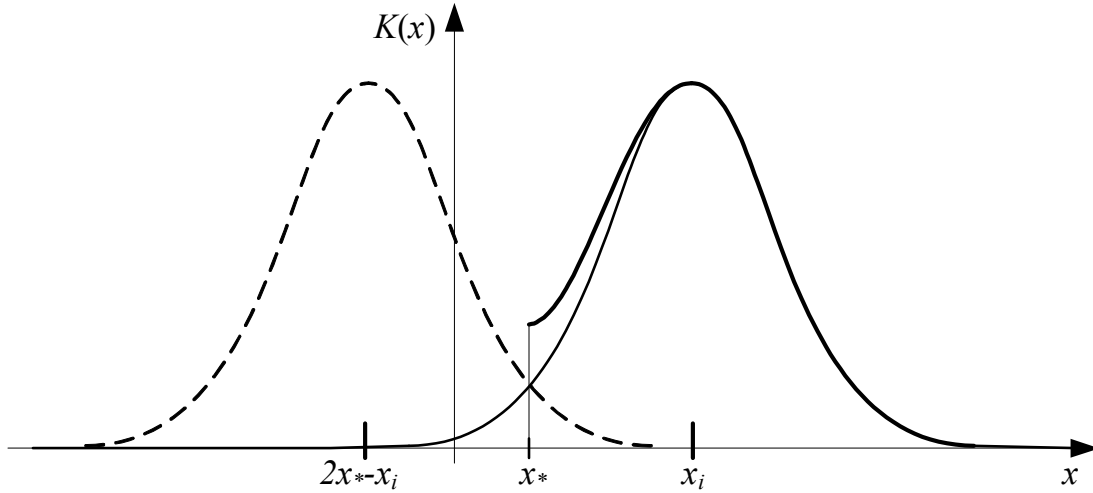
Kolejnym przypadkiem, gdy korzystne wyniki daje modyfikacja postaci estymatora jądrowego, jest sytuacja, gdy zmienna losowa może przyjmować wartości jedynie z pewnego przedziału. Jest to zjawisko dość powszechne, dla przykładu wiek człowieka czy grubość blachy mogą być jedynie dodatnie. Natomiast nawet przy prawidłowo stworzonym estymatorze jądrowym zdarza się, że prawdopodobieństwo poza tym ograniczonym przedziałem jest niezerowe. Dzieje się tak, ponieważ niezależnie od stosowania ograniczonego (np. Epanecznikowa), bądź nieograniczonego (np. normalnego) jądra, części jąder znajdujących się na obrzeżu przedziału mogą częściowo wykraczać poza te dozwolone przedziały. W większości przypadków sytuacja taka nie ma większego znaczenia, tym bardziej, że wartości estymatora poza tym przedziałem są zwykle pomijalnie małe.

Są jednak sytuacje, gdy korzystne jest zastosowanie ograniczenia nośnika. Z tego powodu poniżej przedstawiona zostanie metodologia owego ograniczenia. Wpierw opisane zostanie postępowanie dla *lewostronnego ograniczenia* jednowymiarowej zmiennej losowej. W takim przypadku spełniony ma być warunek:

$$\hat{f}(x) = 0 \text{ dla każdego } x < x_*, \quad (1.95)$$

gdzie x_* jest ograniczeniem lewostronnym.

Sposób postępowania w tym przypadku polega na symetrycznym odbiciu tej części każdego jądra, która leży poza przedziałem $[x_*, \infty)$. Estymator realizowany jest wtedy jako suma fragmentu owego jądra oraz fragmentu jądra symetrycznie odbitego względem ograniczenia x_* , co graficznie przedstawia rysunek 1.7. Linia przerywaną zaznaczono symetryczne odbicie jądra względem ograniczenia x_* , zaś pogrubioną linią sumę obu jąder w dozwolonym przedziale. Łatwo można stwierdzić, że odbiciem punktu x_i względem punktu x_* jest punkt $2x_* - x_i$.



Rysunek 1.7. Lewostronne ograniczenie nośnika estymatora jądrowego

Sumę jąder odpowiadających punktom x_i oraz $2x_* - x_i$ można zapisać w następujący sposób:

$$K\left(\frac{x - x_i}{h}\right) + K\left(\frac{x - (2x_* - x_i)}{h}\right) \text{ dla każdego } i = 1, 2, \dots, m. \quad (1.96)$$

W takim przypadku definicja estymatora jądrowego przyjmuje postać:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^m \chi_{[x_*, \infty)}(x) \left[K\left(\frac{x - x_i}{h}\right) + K\left(\frac{x + x_i - 2x_*}{h}\right) \right], \quad (1.97)$$

gdzie $\chi_{[x_*, \infty)}(x)$ to funkcja charakterystyczna przedziału $[x_*, \infty)$:

$$\chi_{[x_*, \infty)}(x) = \begin{cases} 1, & \text{gdy } x \geq x_* \\ 0, & \text{gdy } x < x_* \end{cases}. \quad (1.98)$$

Analogicznie można zdefiniować *prawostronne ograniczenie nośnika*. Wówczas formułowany jest warunek:

$$\hat{f}(x) = 0 \text{ dla każdego } x > x^*, \quad (1.99)$$

gdzie x^* jest ograniczeniem prawostronnym. Definicja estymatora jądrowego przyjmuje wtedy postać:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^m \chi_{(-\infty, x^*]}(x) \left[K\left(\frac{x - x_i}{h}\right) + K\left(\frac{x + x_i - 2x^*}{h}\right) \right], \quad (1.100)$$

gdzie $\chi_{(-\infty, x^*]}(x)$ oznacza funkcję charakterystyczną przedziału $(-\infty, x^*]$:

$$\chi_{(-\infty, x^*]}(x) = \begin{cases} 1, & \text{gdy } x \leq x^* \\ 0, & \text{gdy } x > x^* \end{cases} \quad (1.101)$$

Metodę tę można łączyć z transformacją liniową oraz z modyfikacją współczynnika wygładzania. W przypadku wielowymiarowego jądra radialnego można ją stosować względem poszczególnych współrzędnych, wymagających ograniczenia.

Rozdział 2. OPIS PROGRAMU *KDEstim*

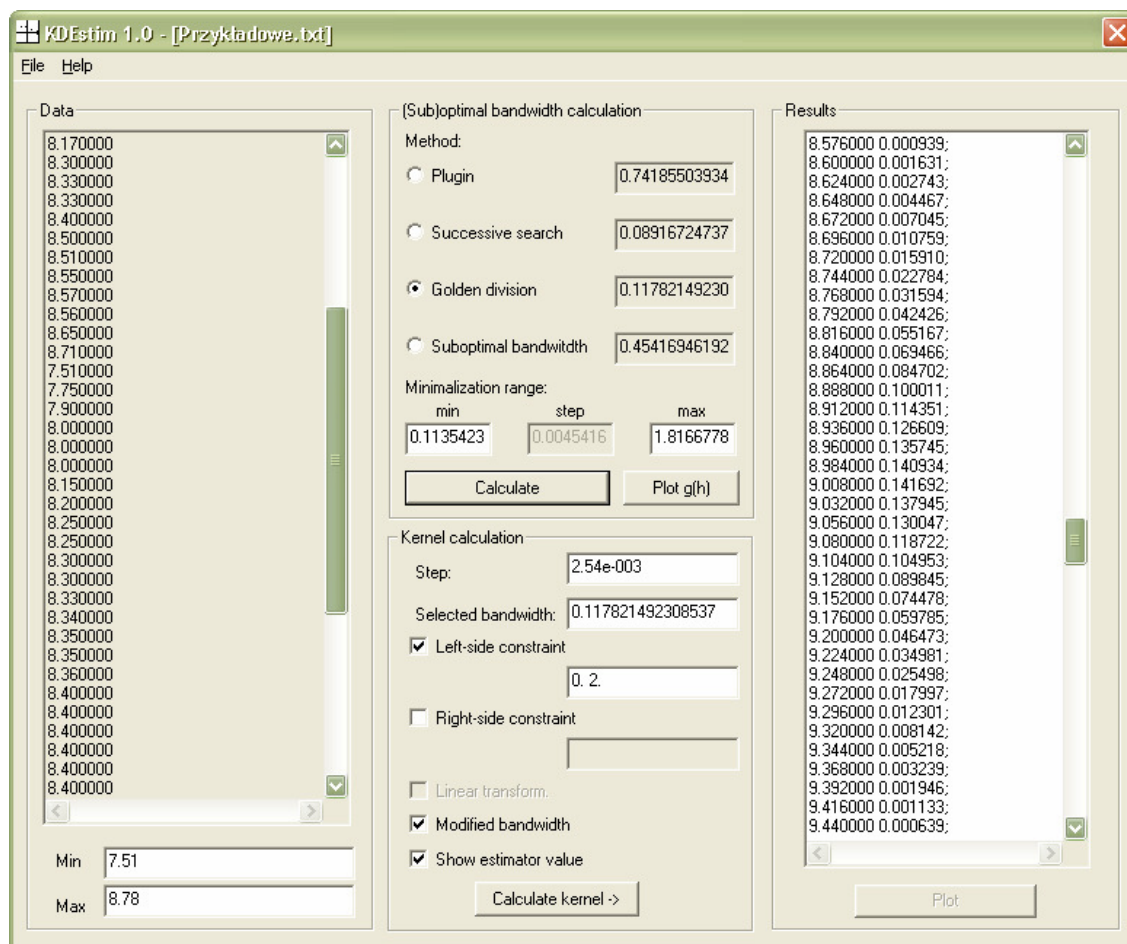
W ramach niniejszej pracy magisterskiej napisany został program *KDEstim* (Kernel Density Estimation). Jest to narzędzie umożliwiające wyznaczanie funkcji gęstości prawdopodobieństwa przy użyciu estymatorów jądrowych dla danych zarówno jedno- jak i wielowymiarowych. W obliczeniach wykorzystywane są normalne jądra radialne. Program oblicza współczynnik wygładzania kilkoma metodami. Kolejną cechą programu jest wizualizacja otrzymanych wyników w przypadku danych jedno- i dwuwymiarowych.

Program *KDEestim* został wykonany w języku C++, w oparciu o środowisko Microsoft Visual Studio, w związku z czym pracuje on pod kontrolą systemu operacyjnego Microsoft Windows. Wybór platformy został głównie podyktowany ogólną dostępnością rozwiązań firmy z Redmond.

Środowisko Visual C++ wyróżnia się łatwością programowania interfejsu przy użyciu biblioteki MFC (Microsoft Foundation Classes) oraz bardzo dobrą przenośnością skompilowanego kodu – dlatego też zostało ono wybrane jako narzędzie implementacyjne (przy wyborze brano również pod uwagę alternatywne rozwiązanie w postaci środowiska C++ Builder firmy Borland).

Okno główne programu (rysunek 2.1) zostało podzielone na cztery części:

- Data,
- (Sub)optimal bandwidth calculation,
- Kernel calculation,
- Results.



Rysunek 2.1. Okno główne programu KDEstim

Pole *Data* służy do wizualizacji wczytanych danych. Aby wczytać dane należy z menu *File* wybrać opcję *Open data...*. Program akceptuje następujący format danych: w każdym wierszu pliku tekstowego znajduje się jedna próbka, a kolejne wymiary próbek oddzielone są spacją.

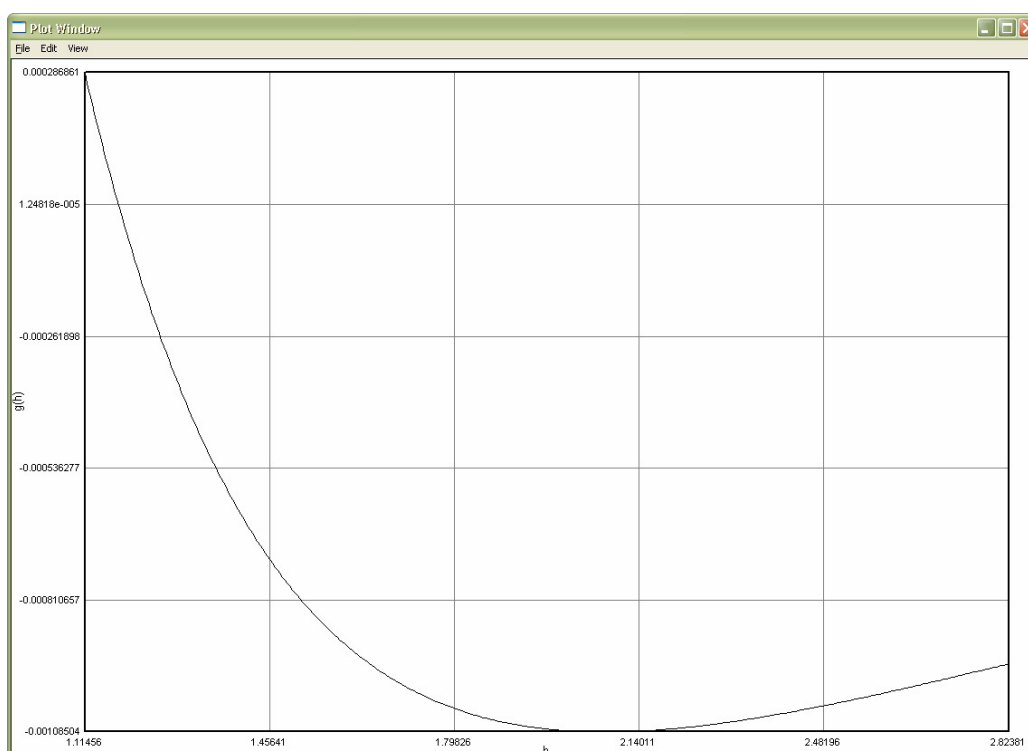
Pola *min* i *max* zawierają wartości maksymalne i minimalne poszczególnych wymiarów wczytanych danych.

Pole *(Sub)optimal bandwidth calculation* umożliwia obliczenie współczynnika wygładzania dla wcześniej otwartych danych. Wybrać można pomiędzy następującymi metodami wyznaczania współczynnika wygładzania:

- *Plug-in* - metoda podstawień (wyłącznie dla danych jednowymiarowych);
- *Successive search* – sukcesywne przeszukiwanie (metoda *cross-validation*);
- *Golden division* – metoda złotego podziału (metoda *cross-validation*);
- *Suboptimal bandwidth* – użycie suboptymalnego współczynnika wygładzania zgodnie ze wzorem (1.52).

W przypadku sukcesywnego przeszukiwania i metody złotego podziału określa się przedział wartości i dokładność z jaką wyszukiwany jest parametr h . Domyślnie program przyjmuje zakres $[\frac{1}{4}h_0^*, 4h_0^*]$ oraz krok $\frac{h_0^*}{100}$, gdzie h_0^* - suboptymalny parametr wygładzania.

Dodatkowo, po obliczeniu parametru h metodą sukcesywnego przeszukiwania, użytkownik ma możliwość wizualizacji funkcji $g(h)$ – po naciśnięciu przycisku *Plot g(h)* pojawia się okno wykresu funkcji (rysunek 2.2). Właściwości tego okna są podobne do właściwości okna wizualizacji wyników, opisanego poniżej.



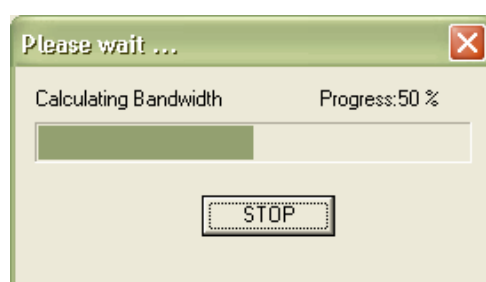
Rysunek 2.2. Okno wizualizacji funkcji $g(h)$

Pole *Kernel calculation* służy do obliczania estymatora jądrowego funkcji gęstości rozkładu. Estymator ten jest liczony w zakresie określonym przez pola *min* i *max*. Krok obliczeń dla każdego wymiaru określany jest przy pomocy pola *Step*. Zaznaczenie opcji *Left-side constraint* lub/i *Right-side constraint* umożliwia ustalenie lewostronnego lub/i prawostronnego ograniczenia nośnika.

Estymator może być liczony z zastosowaniem diagonalnej transformacji liniowej (zaznaczenie opcji *Linear transformation* – opcja ta jest dostępna tylko w przypadku danych wielowymiarowych) oraz/lub zmodyfikowanego współczynnika wygładzania (zaznaczenie opcji *Modified bandwidth*). Ponadto zaznaczenie opcji *Show estimator value* powoduje wyświetlenie obliczonych wartości estymatora w polu *Results*.

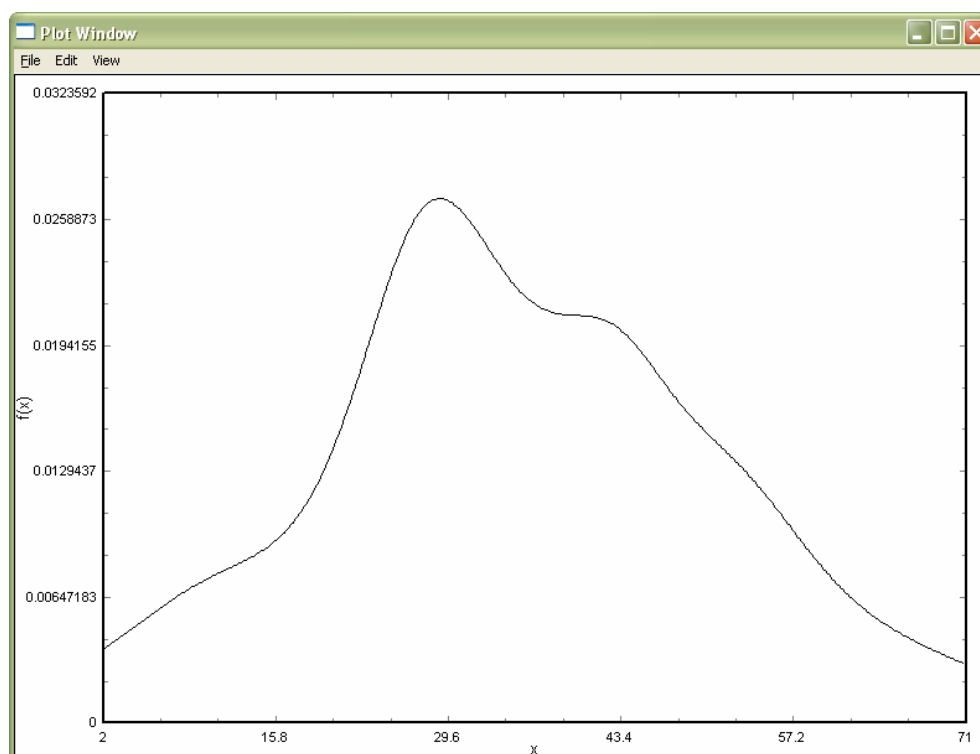
Po wykonaniu obliczeń użytkownik ma możliwość zapisania wyników w pliku tekstowym. W każdym wierszu pliku wyników znajdują się współrzędne n -wymiarowe, po których następuje wartość estymatora jądrowego dla tych współrzędnych. Kolejne wartości liczbowe oddzielone są spacją. Domyślnie wyniki estymacji zapisywane są w plikach z rozszerzeniem *rlt*.

W przypadku naciśnięcia przycisku *Calculate bandwidth* lub *Calculate kernel* otwiera się okno postępu (rysunek 2.3). Umożliwia ono śledzenie zaawansowania obliczeń wykonywanych przez program oraz pozwala na ich zatrzymanie w dowolnym momencie przez użytkownika.

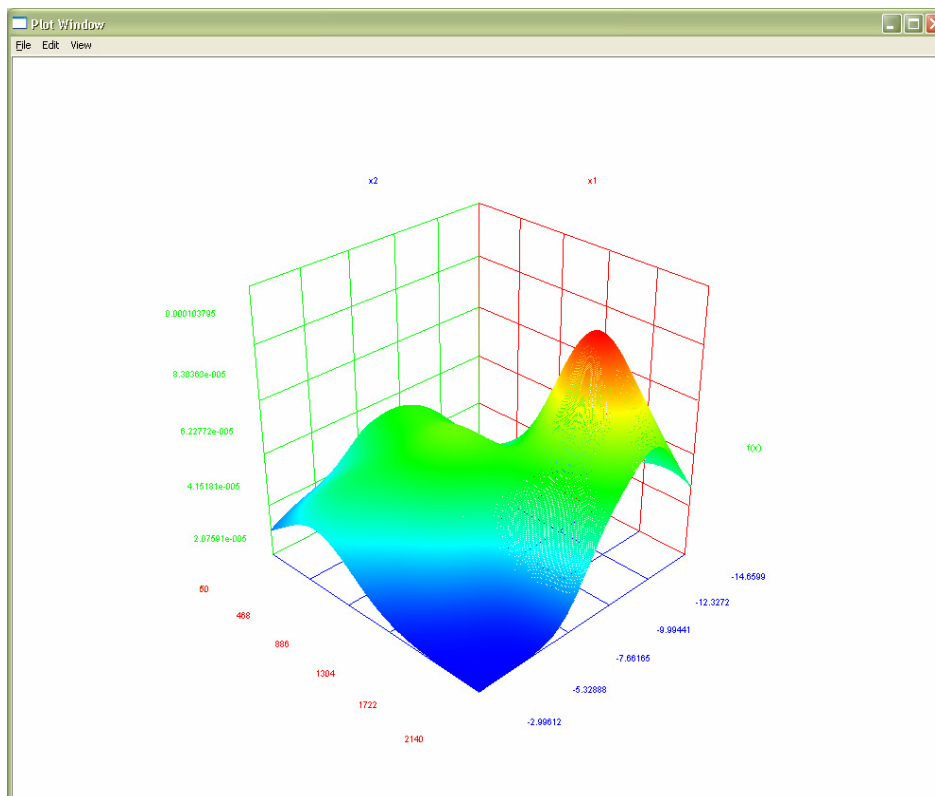


Rysunek 2.3. Okno postępu

Pole *Results* służy do wyświetlania wyników obliczonych przez program. W przypadku danych jedno- jak i dwuwymiarowych naciśnięcie przycisku *Plot* powoduje otwarcie nowego okna – okna wizualizacji wyników (odpowiednio rysunek 2.4 oraz 2.5).



Rysunek 2.4. Okno wizualizacji wyników (dla danych jednowymiarowych)

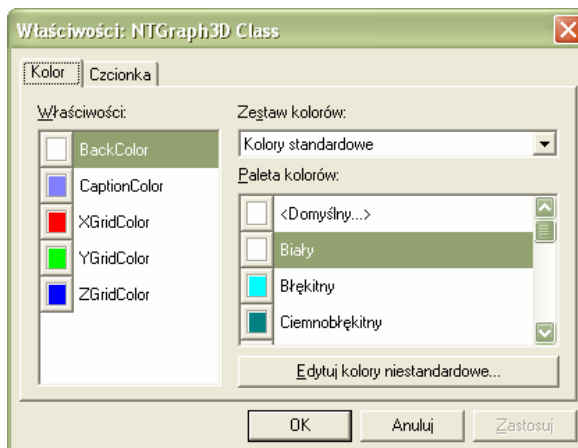


Rysunek 2.5. Okno wizualizacji wyników (dla danych dwuwymiarowych)

W przypadku danych dwuwymiarowych użytkownik ma możliwość dowolnego obracania wykresu – poprzez naciśnięcie i przytrzymanie lewego przycisku myszy a następnie przesuwanie kursora w wybraną stronę. Podczas obrotu wykresu znika przebieg funkcji oraz siatka, a pozostają jedynie osie.

Polecenie *Copy* (menu *Edit* lub menu podręczne) umożliwia skopiowanie zawartości okna do schowka.

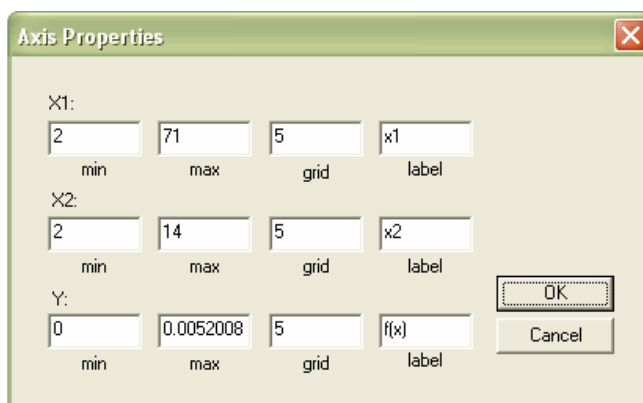
Menu *View* zawiera funkcje umożliwiające zmianę właściwości wyświetlania wyników:



Rysunek 2.6. Okno właściwości wykresu

Polecenie *Preferences* otwiera okno właściwości wykresu (rysunek 2.6). Umożliwia ono zmianę kolorów tła, kolorów siatki oraz rozmiaru i stylu czcionki.

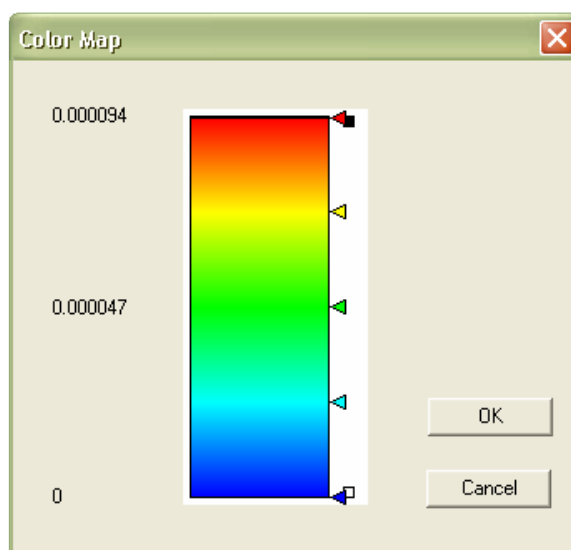
Polecenie *Axis properties* otwiera okno właściwości osi (rysunek 2.7). Umożliwia ono modyfikację zakresu wyświetlanych danych, rozmiaru siatki oraz zmianę opisu dla poszczególnych osi wykresu.



Rysunek 2.7. Okno właściwości osi

Polecenie *Zoom* (dostępne jedynie w przypadku danych dwuwymiarowych) powoduje przełączenie myszki w tryb zbliżania/oddalania. W tym trybie przyciśnięcie i przytrzymanie lewego przycisku myszy w obszarze okna wykresu, a następnie przesuwanie kursora w lewo lub w prawo odpowiednio pomniejsza lub powiększa wykres. Powrót do normalnego trybu następuje po odznaczeniu polecenia *Zoom*.

Polecenie *Color Map* (dostępne jedynie w przypadku danych dwuwymiarowych) otwiera okno kolorów wykresu (rysunek 2.8). Umożliwia ono zmianę gradientu kolorów wykresu, poprzez przesuwanie trójkątów znajdujących się obok palety kolorów.



Rysunek 2.8. Okno kolorów wykresu

Refresh, ostatnie polecenie z menu *View*, służy do odświeżenia okna wykresu.

W programie *KDEstim* wykorzystano następujące gotowe klasy:

- *NTGraph*,
- *NTGraph3D*,
- *CGradient*.

Pierwsze dwie z nich to kontrolki *ActiveX*, umożliwiające wstawianie do okien programu wykresów odpowiednio: dwu- oraz trójwymiarowych. Trzecia klasa służy do ustalania kolorystyki wykresów trójwymiarowych 3D.

Wszystkie wyżej wymienione klasy zostały pobrane ze strony internetowej Code Project [6].

Rozdział 3. ESTYMACJA ROZKŁADÓW W WYBRANYCH SYSTEMACH RZECZYWISTYCH

W niniejszym rozdziale wykorzystano estymatory jądrowe do analizy wybranych rozkładów rzeczywistych. W pierwszej kolejności rozpatrzono dane charakteryzujące osoby wynajmujące pokoje w górskim pensjonacie. Jako kolejny przykład wybrane zostały dane socjologiczne zaczerpnięte z Polskich Generalnych Sondaży Społecznych (PGSS). Trzecim rozważanym przypadkiem – z dziedziny szeroko rozumianej elektrotechniki - jest detekcja uszkodzeń silnika asynchronicznego. Na końcu rozpatrywane są rozkłady z dziedziny fizyki wysokich energii.

Podczas analizy tychże danych wykorzystany został program *KDEstim*, szczegółowo opisany w poprzednim rozdziale.

3.1. Charakterystyka cech gości górskiego pensjonatu

Jako pierwsze estymacji jądrowej poddane zostały dane zaczerpnięte z księgi meldunkowej prywatnego pensjonatu znajdujący się w górskiej miejscowości Szczawnicy.

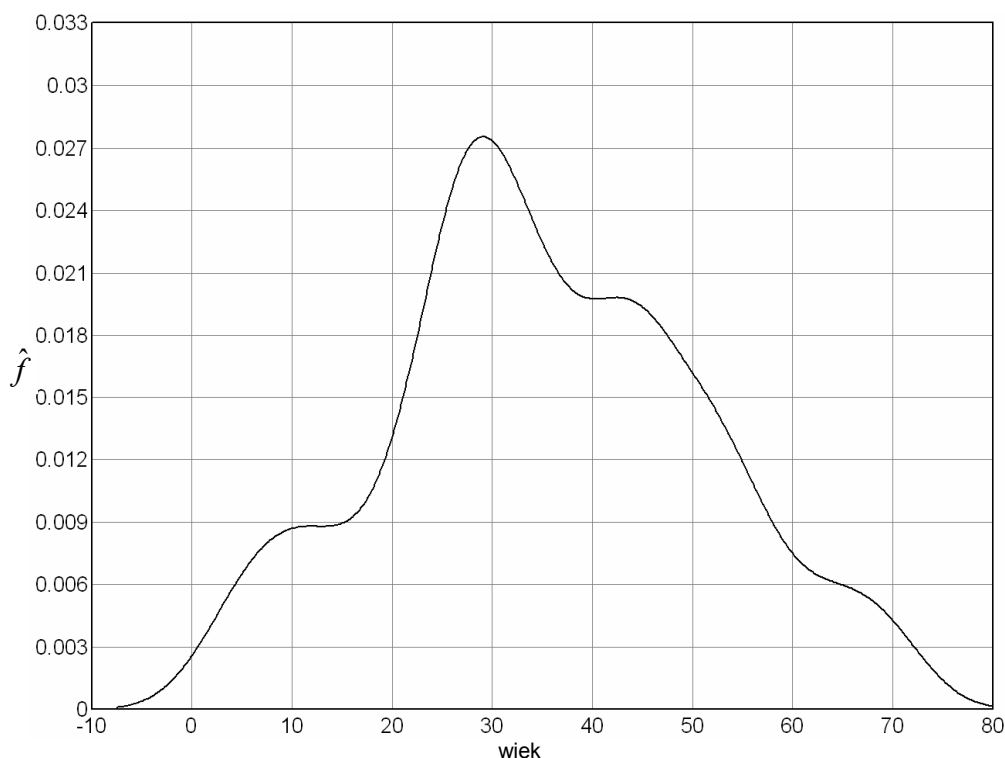
Dla gospodarza takiego pensjonatu ważnymi informacjami są wiek gości oraz czas, na jaki wynajmują oni pokoje. Posiadając takie informacje może on lepiej przygotować swoją ofertę.

W związku z powyższym, zanalizowano rzeczywiste dane dotyczące wybranego pensjonatu dotyczące lat 1999-2005. Dla każdego gościa ustalono *datę urodzenia* oraz jego

czas pobytu. Następnie data urodzenia została zamieniona na wiek, jaki dana osoba miała, w czasie jej pobytu w pensjonacie.

3.1.1. Analiza jednowymiarowa

W pierwszej kolejności analizie poddano przekrój wiekowy osób odwiedzających rozważany pensjonat. Do wyznaczenia współczynnika wygładzania zastosowano metodę *plug-in*. Wynik przeprowadzonej estymacji przedstawia rysunek 3.1.



Rysunek 3.1. Wynik estymacji dla wieku gości pensjonatu / $h = 4,19$ /

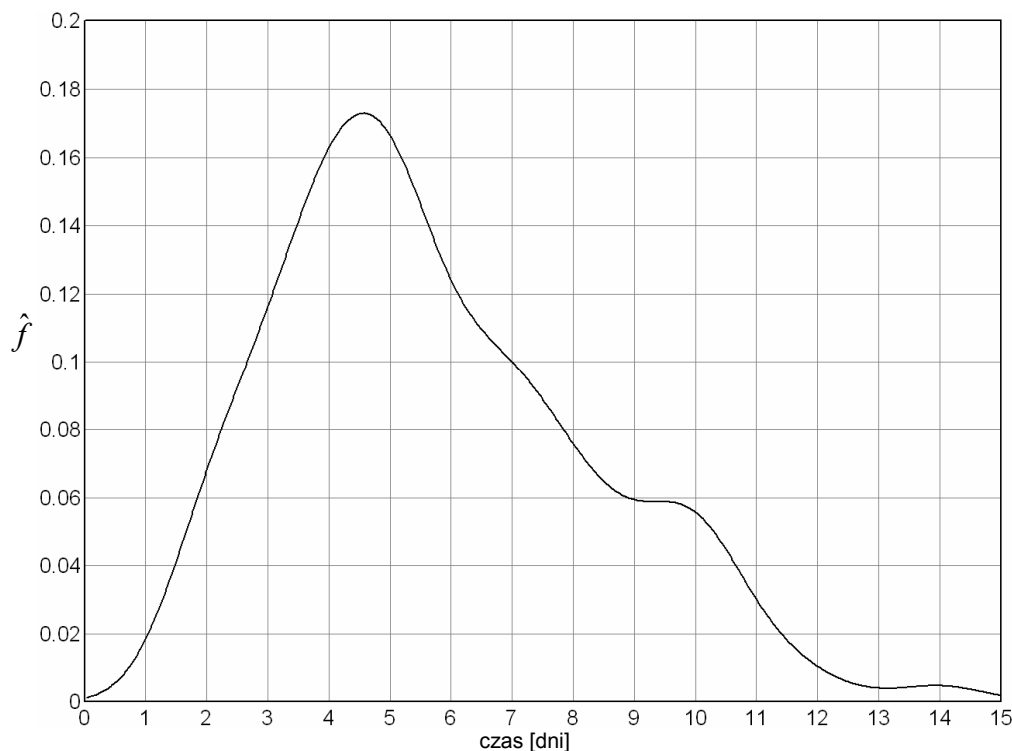
Już wstępna analiza wyników estymacji pozwala stwierdzić, że wśród osób wynajmujących pokoje w pensjonacie zdecydowanie najliczniejszą grupą są ludzie w wieku 25-35 lat. Są to osoby młode, które prawdopodobnie jeszcze nie założyły rodziny. Przyjeżdżają oni do pensjonatu razem z przyjaciółmi (oczywiście w tym samym wieku) w celu aktywnego spędzenia czasu (turystyka górską, spływ Dunajcem).

Kolejną liczną grupą osób wynajmujących pokoje są ludzie dojrzały – w wieku ok. 45 lat. Są to osoby ustawkowane, najczęściej posiadające rodzinę, razem z nimi przyjeżdżają małe dzieci – kolejny punkt przegięcia w okolicach wieku 10 lat.

Najmniej licznymi gośćmi pensjonatu są emeryci – przyjeżdżający do Szczawnicy zapewne w celach rehabilitacyjno-wypoczynkowych.

Drugim czynnikiem poddanym analizie jest *czas pobytu* gości w pensjonacie wyrażony w dniach. W tym przypadku do wyznaczenia współczynnika wygładzania wykorzystano również metodę *plug-in*.

Wynik przeprowadzonej estymacji przedstawia rysunek 3.2.



Rysunek 3.2. Wynik estymacji dla czasu pobytu gości w pensjonacie / $h = 0,74$ /

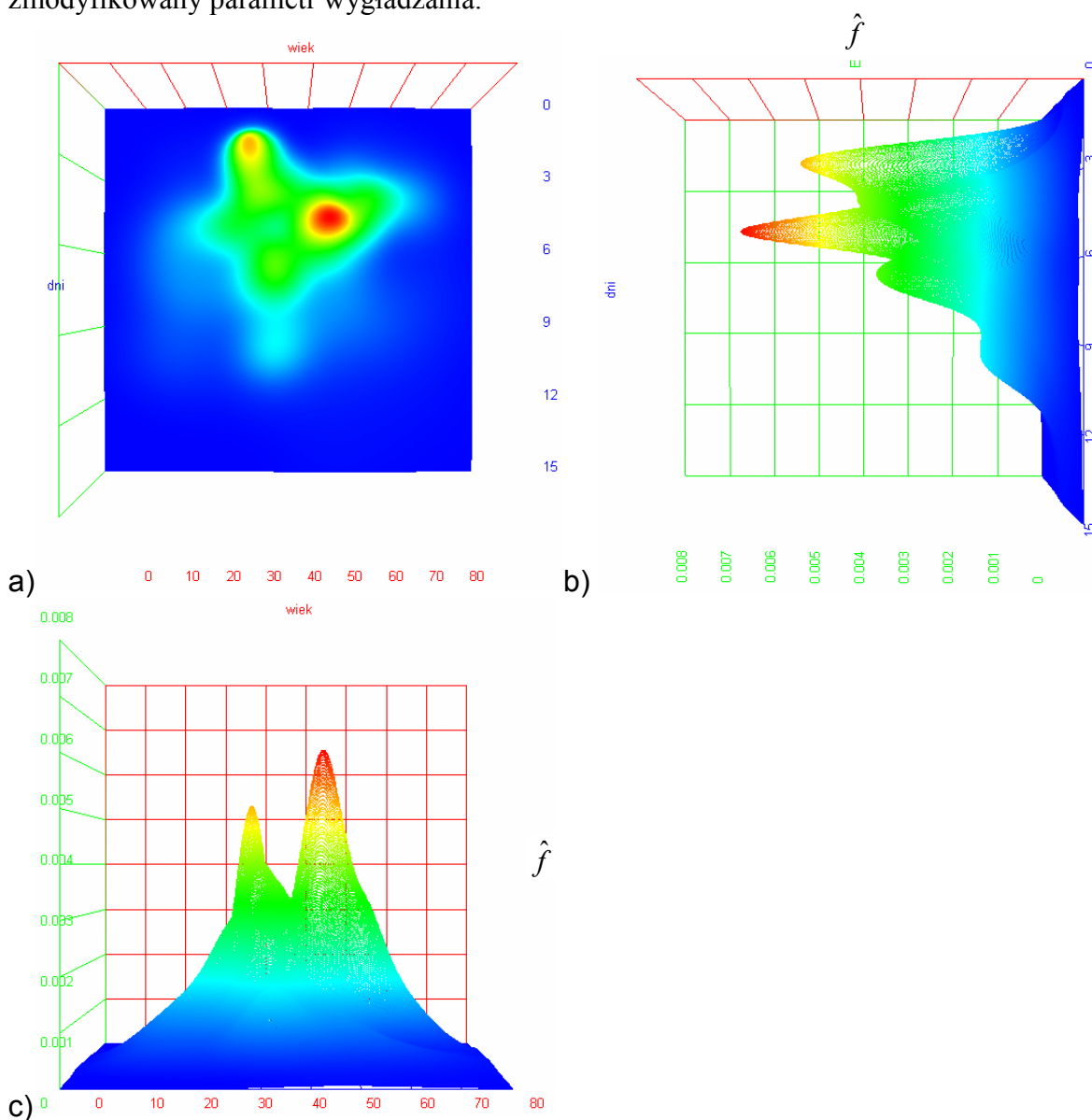
Na podstawie obserwacji wyników estymacji można zauważyć, że goście pensjonatu najczęściej wynajmują pokoje na 4-5 dni. Oznacza to, że na wyjazd do Szczawnicy ludzie wybierają zazwyczaj tzw. „długie” weekendy (np. majowy, czy czerwcowy – w okolicach święta Bożego Ciała), a nie „zwykłe weekendy” (pobyt dwudniowy jest ponad dwa razy rzadziej spotykany).

Powyżej pięciu dni funkcja gęstości stopniowo opada do poziomu zerowego (dla 15 dni). Jednak dla ok. 10 dni pojawia się mała anomalia – punkt przegięcia. Dotyczy on zapewne tych osób, które w Szczawnicy wynajmują pokoje podczas swojego urlopu. „Wczasowiczów” takich jest trzy razy mniej niż „długoweekendowców”.

3.1.2. Analiza dwuwymiarowa

Zastosowanie jednowymiarowych estymatorów jądrowych prowadzi do interesujących wniosków dotyczących cech osób wynajmujących pokoje w pensjonacie. Jednak jeszcze ciekawsze wyniki daje analiza dwuwymiarowa.

Pierwszą składową rozważanej zmiennej losowej jest *wiek* gości, natomiast drugą – ich *czas pobytu* w pensjonacie. W obliczeniach zastosowano transformację liniową oraz zmodyfikowany parametr wygładzania.



Rysunek 3.3. Wynik estymacji (a, b, c) dla zmiennej [wiek gości, czas pobytu] / $h = 0,32$ /

Obliczony rozkład gęstości, który przedstawia rysunek 3.3, cechuje się dwoma, bardzo wyraźnymi modami: pierwszym dla wieku 25-30 lat i czasu pobytu 2-3 dni oraz drugim dla wieku 40-50 lat i czasu pobytu 4-5 dni. Dodatkowo zauważyć można, że dla przedziału wiekowego 30-40 lat stosunkowo duże wartości gęstości występują dla najszerszego przedziału czasu pobytu – od 3 do 9, a nawet 10 dni.

Otrzymane wyniki potwierdzają wcześniej sformułowaną tezę, że osoby młode wybierają krótkie, „weekendowe wypady” do Szczawnicy. Maksymalnie wykorzystują oni

czas, najczęściej z przyjaciółmi, np. wędrując po górach, a pensjonat traktują jako bazę noclegową.

Inaczej przedstawia się sytuacja dla osób 40-50 letnich. Dla nich 4-5 dniowe maksimum jest jeszcze wyraźniejsze – goście w tym wieku preferują prawie wyłącznie „długie weekendy”, zdecydowanie nie śpieszą się tak jak osoby młode. Zauważyć też można, że to właśnie osoby dojrzałe przyjeżdżają do pensjonatu ze swoimi dziećmi (zwiększona gęstość prawdopodobieństwa dla dzieci pokrywa się czasowo z maksimum pięćdziesięciolatków).

Interesujący jest również czas pobytu dla osób 30-40 letnich. Jest to jedyna grupa ludzi, dla których czas pobytu zawiera się w całym zakresie (od 3 do 12 dni), z niewielkim maksimum dla ok. 7 dni. W związku z tym wiek taki wydaje się wiekiem „przejściowym”.

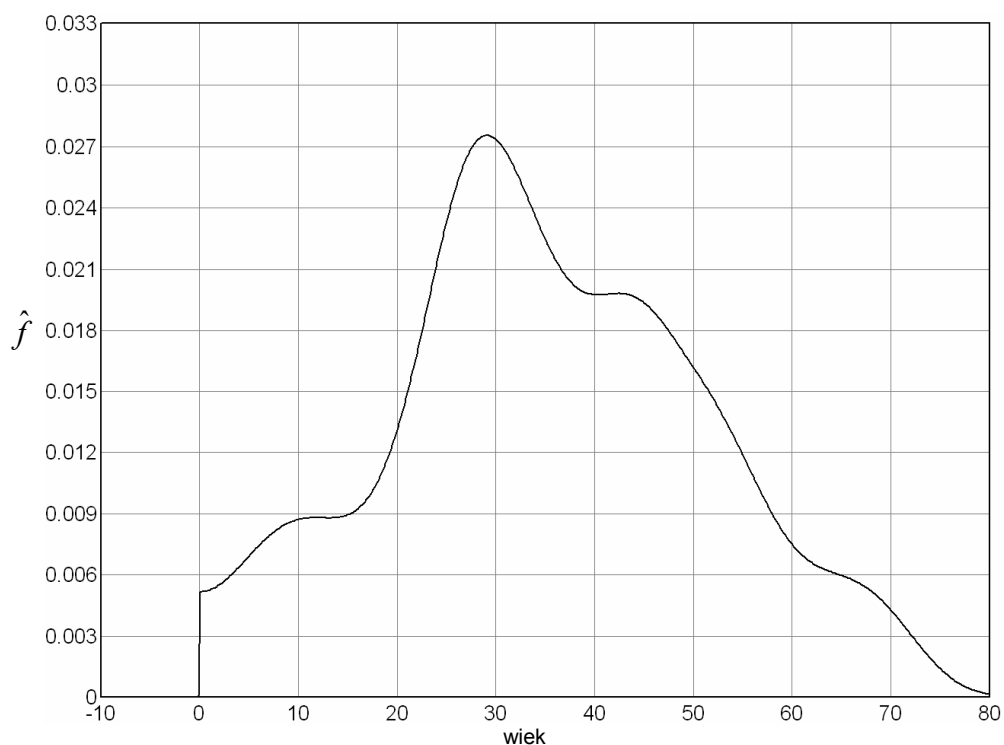
3.1.3. Zastosowanie ograniczenia nośnika

Podczas analizowania estymatorów jądrowych obliczonych dla pensjonatu można zauważyć pewną nieścisłość: prawdopodobieństwo dla ujemnych wartości wieku gości jest niezerowe. Sytuacja taka jest związana z wykorzystaniem nieograniczonego nośnika do analizy problemu ograniczonego. Z praktycznego punktu widzenia błąd ten jest w akceptowalnym zakresie.

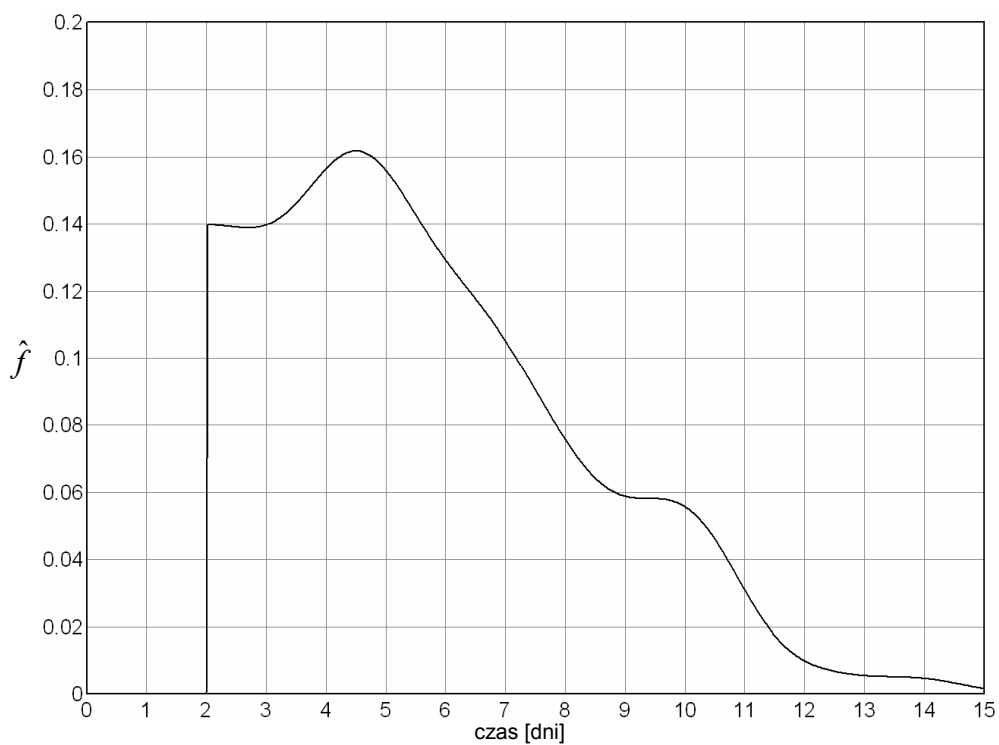
W celu dopasowania właściwości nośnika do uwarunkowań analizowanego systemu, zaproponować można ograniczenie nośnika. Metodyka postępowania w tym przypadku została opisana w części teoretycznej niniejszej pracy. Dla systemu rozważanego w bieżącym podrozdziale, oczywiste wydaje się wprowadzenie lewostronnego ograniczenia dla *wieku* 0 lat.

Kolejne ograniczenie może być wprowadzone na współrzędną *czas pobytu* – bardzo często zdarza się, że właściciele pensjonatu określają minimalną ilość noclegów, np. na dwa, jak to ma miejsce w przypadku rozpatrywanego pensjonatu. Właściciele tłumaczą takie wymagania czynnikami ekonomicznymi.

W związku z powyższymi rozważaniami wykonano ponowne obliczenia estymatora jądrowego z ograniczeniami lewostronnymi: 0 – dla *wieku* gości, 2 – dla *czasu pobytu* w pensjonacie. Otrzymane wyniki przedstawia rysunki 3.4 i 3.5.

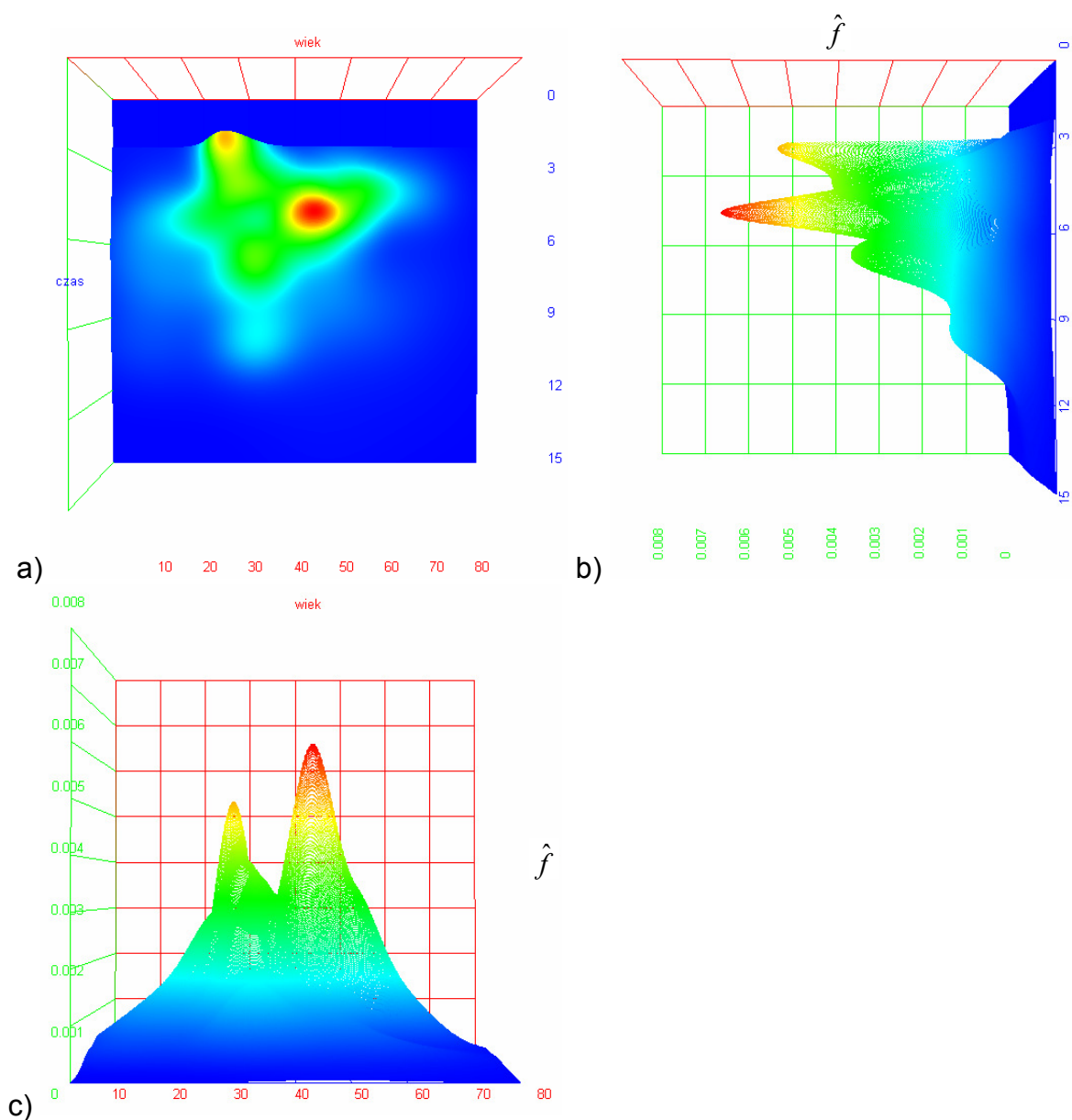


Rysunek 3.4. Wynik estymacji dla wieku gości pensjonatu (z ograniczeniem lewostronnym dla 0 lat)



Rysunek 3.5. Wynik estymacji dla czasu pobytu w pensjonacie (z ograniczeniem lewostronnym dla 2 dni)

Podobne ograniczenie można zastosować w przypadku analizy dwuwymiarowej (rysunek 3.6).



Rysunek 3.6. Wynik estymacji dla zależności czasu pobytu w pensjonacie od wieku gości w przypadku ograniczenia nośnika $/h = 0,32/$

Wyniki estymacji można porównać z rozkładami otrzymanymi dla przypadków bez stosowania ograniczenia (rysunki 3.1, 3.2 i 3.3).

Dość znaczną różnicę obserwuje się w przypadku zmiennej *czas pobytu*. Dzieje się tak, gdyż spora część próbek znajduje się na granicy ograniczenia (*czas pobytu* = 2 dni). W związku z odbiciem tych części jąder, które znajdują się poza dozwolonym przedziałem, wartość gęstości prawdopodobieństwa dla granicy podwoiła się, osiągając znaczną wartość. Dla zmiennej *wiek* zwiększenie wartości gęstości dla granicy jest zdecydowanie mniejsze i nie ma wpływu na interpretację wyników.

Reasumując, zastosowanie ograniczenia dla zmiennej *wiek* jest korzystne - można się w ten sposób uchronić przed niewygodnymi pytaniami laików o ujemny wiek. Z kolei

ograniczenia takie dla *czasu pobytu* nie wydaje się zbyt celowe, tym bardziej, że brak ograniczenia w tym przypadku nie prowadzi do kontrowersyjnych wniosków.

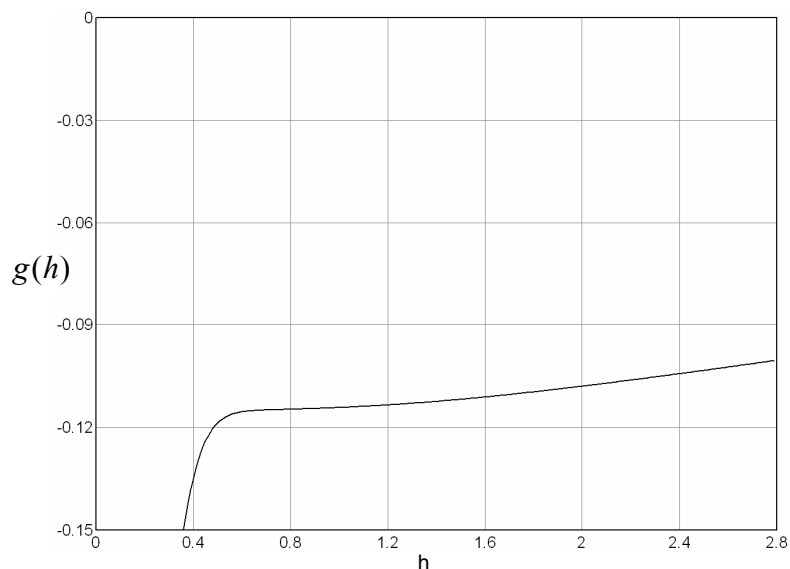
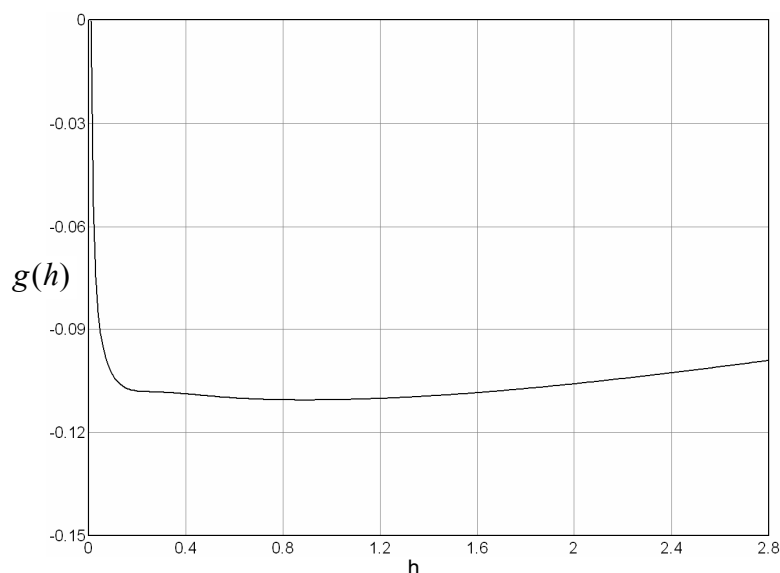
3.1.4. Wpływ randomizacji danych na działanie algorytmu krzyżowego uwiarygodniania

Do wyznaczenia estymatora jądrowego w przypadku *czasu pobytu* w pensjonacie, użyto parametru wygładzania h , obliczonego metodą podstawień. Podjęto również próbę porównania otrzymanego parametru wygładzania z parametrem uzyskanym metodą krzyżowego uwiarygodniania. Przeszkodą w tym okazał się jednak brak minimum funkcji g (rysunek 3.7). Zauważyć można, że w pobliżu zera funkcja ta „załamuje” się i gwałtownie zmierza do $-\infty$. Przypuszczalnie powodem takiego zjawiska jest fakt, iż zmienna losowa przyjmuje tylko wartości całkowite, dodatkowo jedynie z wąskiego zakresu [2, 14].

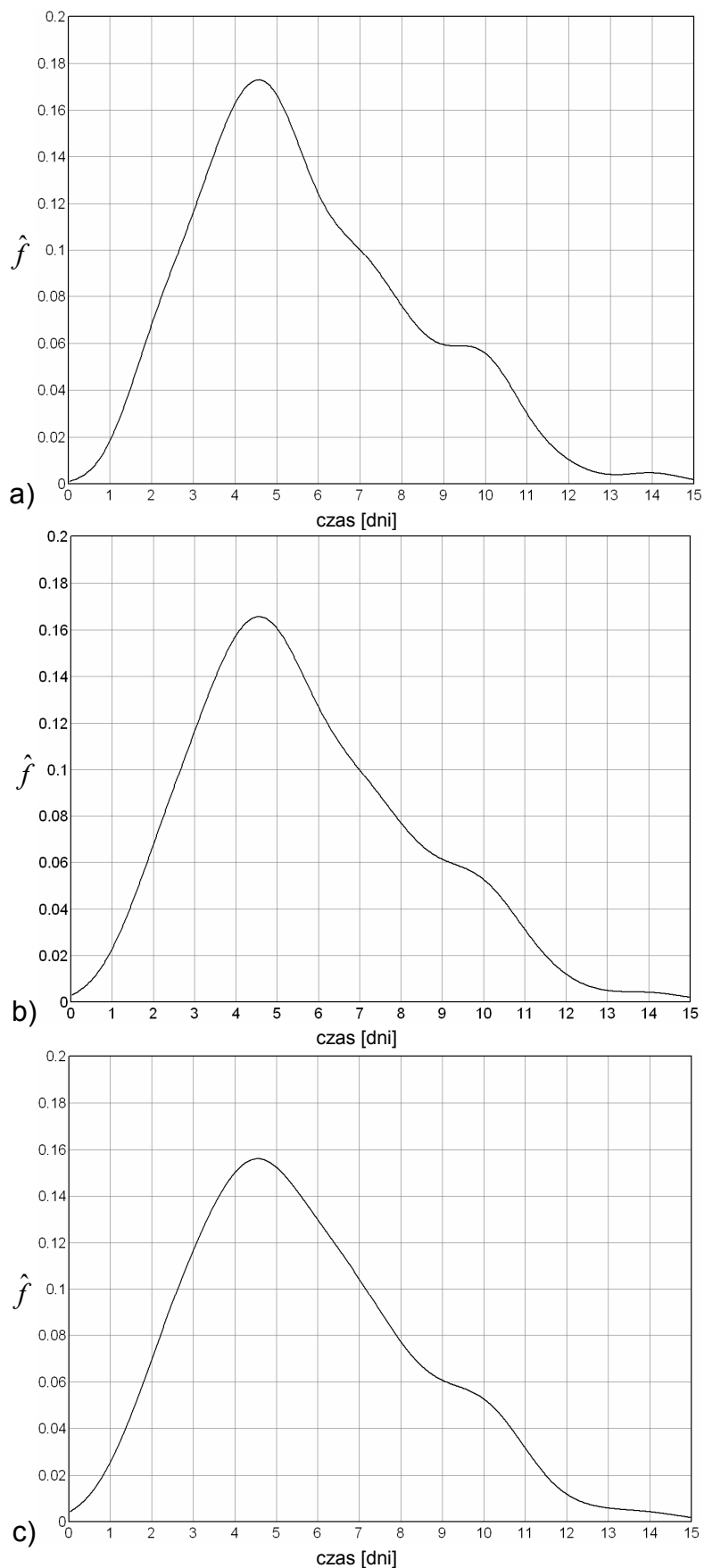
Jednym ze sposobów przeciwdziałania takiej sytuacji jest tzw. *randomizacja danych*. Polega ona na dodaniu do każdego elementu próby losowej wartości z przedziału $[-b, b]$, gdzie b to połowa zakresu pomiędzy poszczególnymi wartościami, jakie przyjmuje oryginalny element próby losowej. Można więc powiedzieć, że jest to proces odwrotny do procesu *kwantyzacji danych*, powszechnie stosowanego np. w akwizycji danych – przy zamianie danych analogowych w cyfrowe.

Kierując się powyższą zasadą, na podstawie danych dotyczących czasu pobytu w pensjonacie, przy użyciu programu *Matlab*, stworzono nową serię danych poprzez dodanie losowych liczb z zakresu $[-0,5; 0,5]$ otrzymanych z generatora rozkładu jednostajnego.

Otrzymany dzięki takiemu zabiegowi wykres funkcji g przedstawia rysunek 3.8. W przypadku randomizacji, minimum funkcji istnieje i wynosi $h = 0,89$, dodatkowo funkcja ta nie wykazuje załamania (przypomina asymetryczną parabolę o ramionach skierowanych w górę).

Rysunek 3.7. Wykres funkcji $g(h)$ bez randomizacji danych, minimum - brakRysunek 3.8. Wykres funkcji $g(h)$ przy randomizacji $[-0,5; 0,5]$, minimum dla $h = 0,89$

Pozostaje jeszcze do rozstrzygnięcia sprawa, jaki wpływ na sam wynik estymacji ma randomizacja danych. Porównanie wyników przedstawiają rysunki 3.10a - c. Pierwszy z rysunków przedstawia wynik estymacji jądrowej dla oryginalnych danych z użyciem $h = 0,74$ (wyznaczonego metodą *plug-in*), drugi – estymację jądrową dla oryginalnych danych wyznaczonej jednak z użyciem $h = 0,89$ (a więc uzyskanego metodą krzyżowego uwiarygodniania po randomizacji danych), zaś trzeci - estymacji jądrowej dla danych zrandomizowanych z użyciem $h = 0,89$.



Rysunek 3.9. Porównanie wyników po randomizacji danych: a) dane oryginalne / $h=0,74$ /, b) dane oryginalne / $h=0,89$ /, c) dane zmodyfikowane / $h = 0,89$ /

Porównując rysunek 3.10a z rysunkiem 3.10b, widać oczywiste (ze względu na większą wartość parametru h) wygładzenie funkcji gęstości. Porównując jednak rysunek 3.10b z rysunkiem 3.11c wyraźnie widać, że sama randomizacja danych nie ma zupełnego wpływu na kształt funkcji gęstości f (oczywiście dla tej samej wartości h).

Podsumowując, randomizacja danych wydaje się przydatnym narzędziem umożliwiającym wyznaczenie parametru wygładzania metodą krzyżowego uwiarygodniania. Jej skuteczność powinna zostać jeszcze sprawdzona dla danych wielowymiarowych. Właśnie dla takich danych stosowanie randomizacji byłoby przydatne, gdyż, jak stwierdzono w części teoretycznej, w takim przypadku niemożliwe jest użycie metody podstawień. Sprawdzenie użyteczności randomizacji w tym zakresie leży jednak poza zakresem niniejszej pracy, jest trudne ze względu na brak prostej metody pozwalającej na zweryfikowanie uzyskanych wyników i zostało jedynie zasygnalizowane jako kolejny krok w badaniu własności estymatorów jądrowych.

3.2. Analiza wyników sondażu PGSS

Przedmiotem rozważań niniejszego podrozdziału jest zastosowanie estymatorów jądrowych w celu uzyskania gęstości rozkładu prawdopodobieństwa zmiennej losowej, której próba pobrana została z ogólnodostępnych danych socjologicznych stanowiących wyniki Polskich Generalnych Sondaży Społecznych (w skrócie PGSS) [4].

Sondaże PGSS, przeprowadzono kolejno w latach 1993, 1994, 1995, 1997, 1999, 2002. Ich obszerne podsumowanie stanowi publikacja [5]. Ogólnym celem omawianego Sondażu jest „systematyczny pomiar trendów i skutków zmian społecznych w Polsce” [5].

W sześciu edycjach PGSS przebadano ogółem 13664 respondentów, zadając im w różnych wariantach ankiet ponad 1300 pytań. W wyniku opracowania wyników uzyskano obszerny zestaw danych do wielorakich zastosowań związanych z badaniem postaw i poglądów Polaków. Zbiór ów zawiera rekordy o długości 1335 pól każdy i dostępny jest w sieci Internet, po zarejestrowaniu, w postaci systemowego pliku skumulowanych danych w języku pakietu statystycznego SPSS (interfejs programu przedstawia rysunek 3.10).

The screenshot shows the SPSS Data Editor window with a data table. The table has columns: pres00_1, who00_1, religrel, relig, attend, pray, reliten, postlife, churhpow, degree, degree1, educ, educ1, schoconti, school. The rows contain individual respondent data, including names like OLECHOWSKI, KRZAKLEWSKI, and KWASNIEWSKI, and their responses to various questions. A 'Variables' dialog box is open, showing a list of variables on the left and 'Variable Information' for 'who00_1' on the right. The 'Variable Information' for 'who00_1' includes: Label: NA KOGO GŁOSOWAŁ W WYB PREZYDEN 2000, Type: F2, Missing Values: -2, -1, 99, Measurement Level: Scale, and Value Labels: -2 ND:PYT NIE ZADANE, -1 ND:NIE GŁOSOWAŁ, 1 GRABOWSKI,TADEUSZ, 2 IKONOWICZ,PIOTR, 3 KALINOWSKI,JAROS, 4 KOFWIN,MIKKE,JANUSZ, 5 KRZAKLEWSKI,MARIAN.

Rysunek 3.10. Interfejs programu SPSS

Do analizy w niniejszej pracy wybrano wyniki uzyskane w roku 2002, a konkretnie próbę dla zmiennej losowej AGE określającą wiek respondenta. Odniesiono ją do wartości przyjmowanych przez zmienną losową WHO00_1. Określa ona odpowiedź respondenta na pytanie: „Na kogo oddał(a) Pan(i) głos w wyborach prezydenckich w 2000 r.?”.

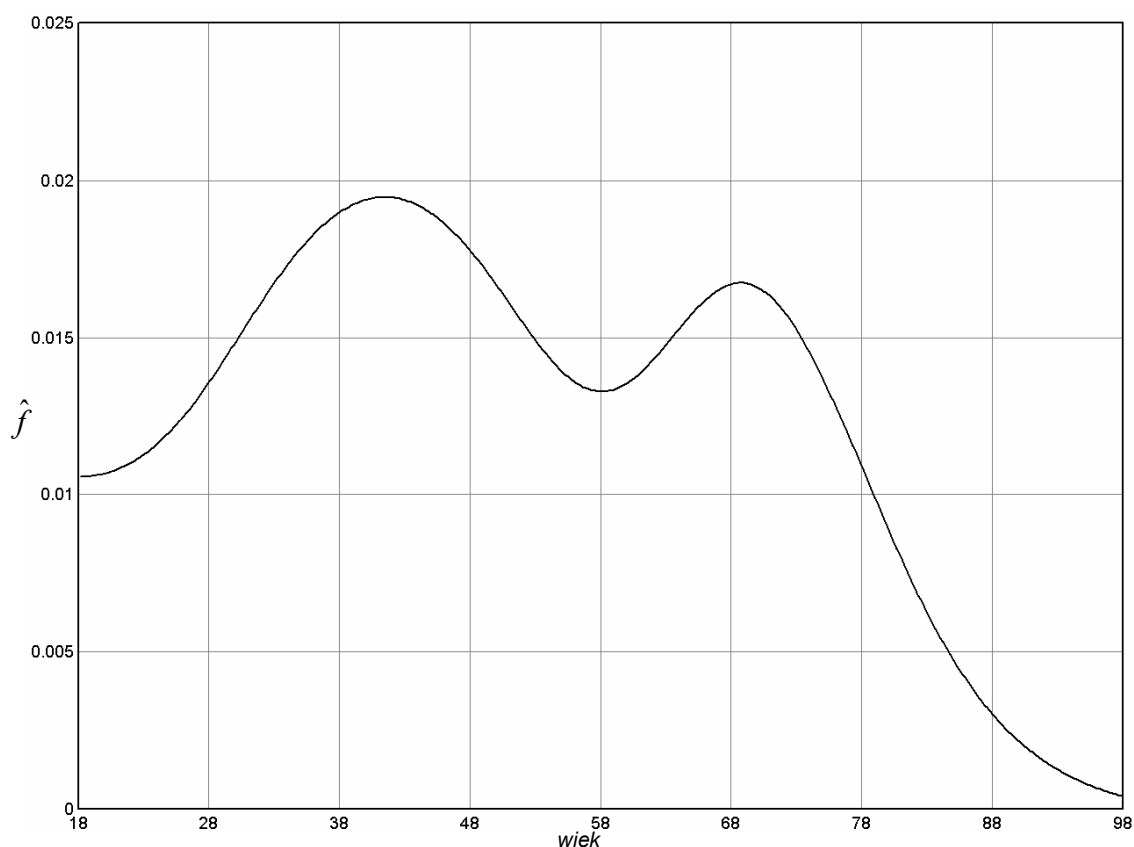
Po przeniesieniu danych z formatu SPSS do postaci obsługiwanej przez program *KDEstim* (podstawy pracy z SPSS obejmuje pozycja [10]) wydzielono klasy odpowiadające poszczególnym kandydatom, na których respondenci oddali swój głos. Każda z klas zawiera próbę losową (jej licznosc podaje tabela 3.1) reprezentującą wiek respondentów (w roku wyborczym) którzy zagłosowali na daną osobę ubiegającą się o urząd Prezydenta Rzeczypospolitej Polskiej w 2000 roku. Próby o licznosci mniejszej niż 70 zostały odrzucone. Dodatkowo uwzględniono również klasę odpowiadającą próbie respondentów, którzy na zadane pytanie odpowiadali „nie pamiętam” (próba ta posiada znaczną licznosc – 113).

Tabela 3.1 Liczność próby losowej dla poszczególnych kandydatów

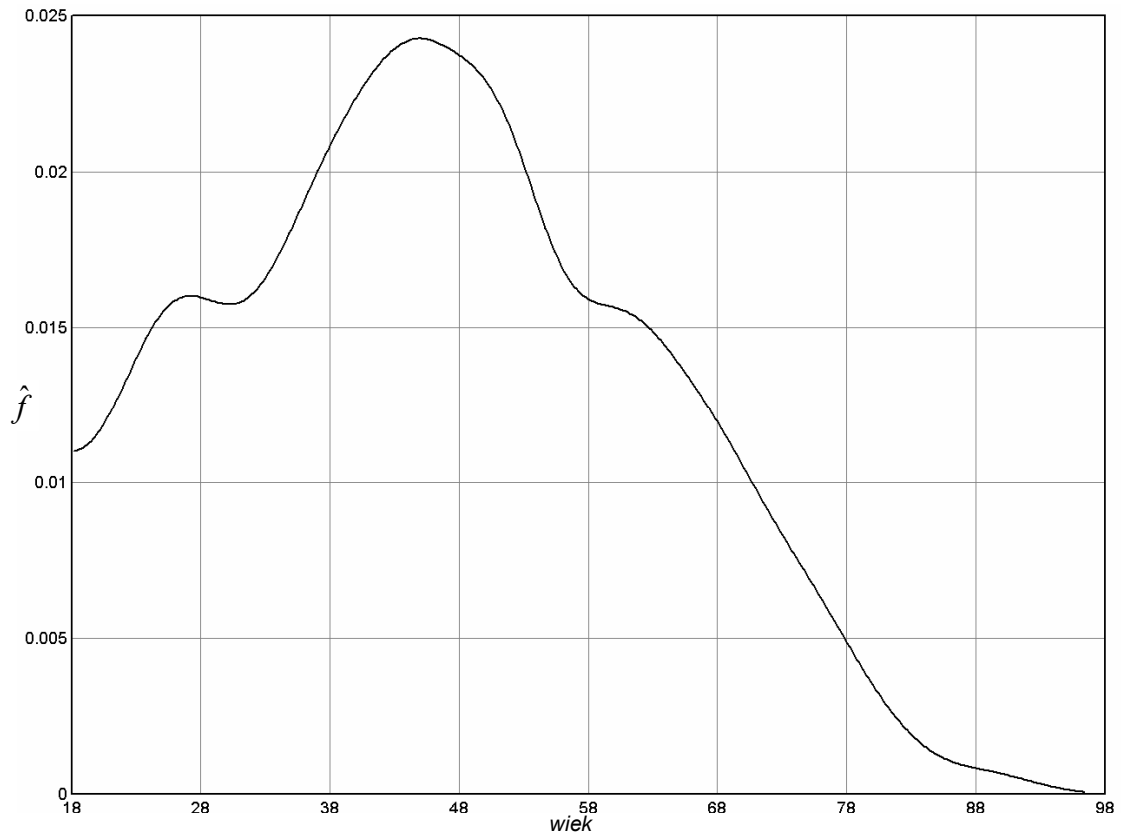
Kandydat	Liczność próby
Marian Krzaklewski	74
Aleksander Kwaśniewski	1271
Andrzej Olechowski	190
Lech Wałęsa	79
„nie pamiętam”	113

Dla każdej z klas wyznaczono gęstość rozkładu prawdopodobieństwa stosując estymatory jądrowe o współczynnikach wygładzania wyznaczonych metodą *plug-in*. Aby oddać rzeczywistą własność populacji zastosowano ograniczenie lewostronnie nośnika zmiennej losowej dla wartości 18 (czynne prawo wyborcze przysługuje w Polsce obywatelom polskim, którzy ukończyli 18 lat).

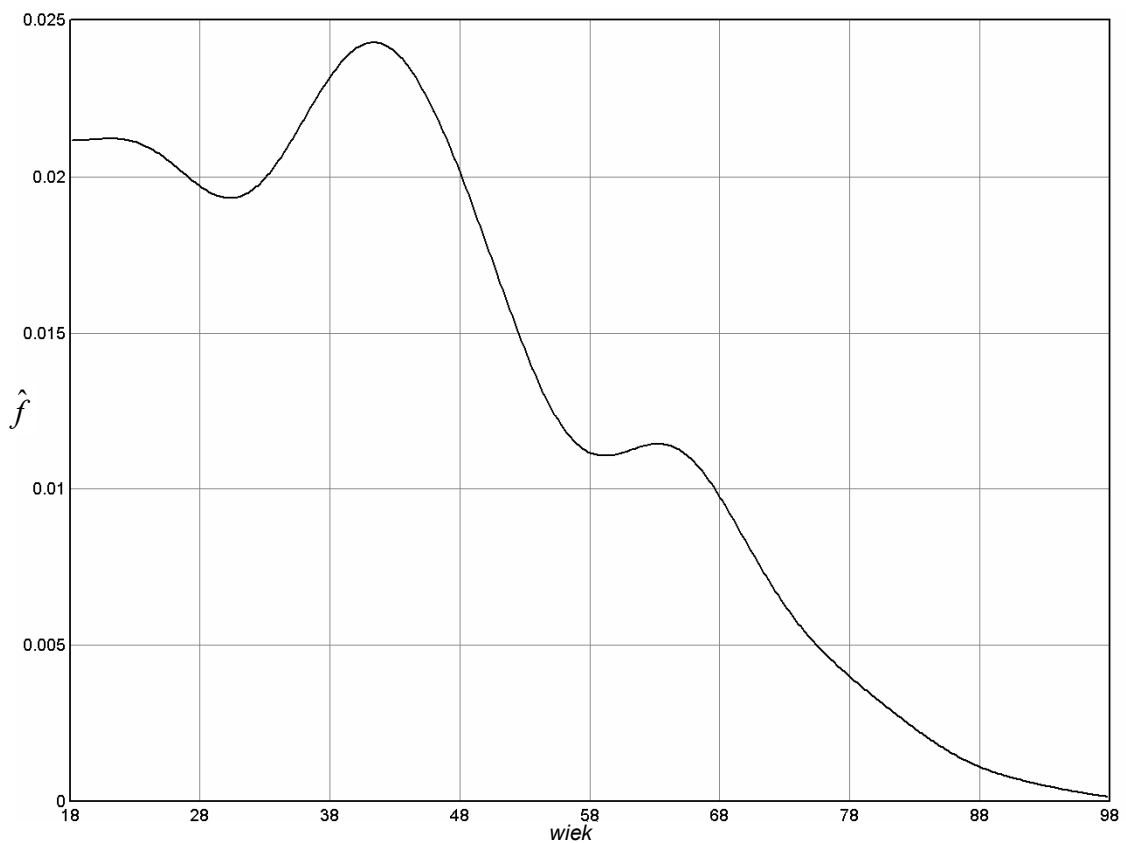
Graficzną prezentację otrzymanych wyników przedstawiają rysunki 3.11-3.15.



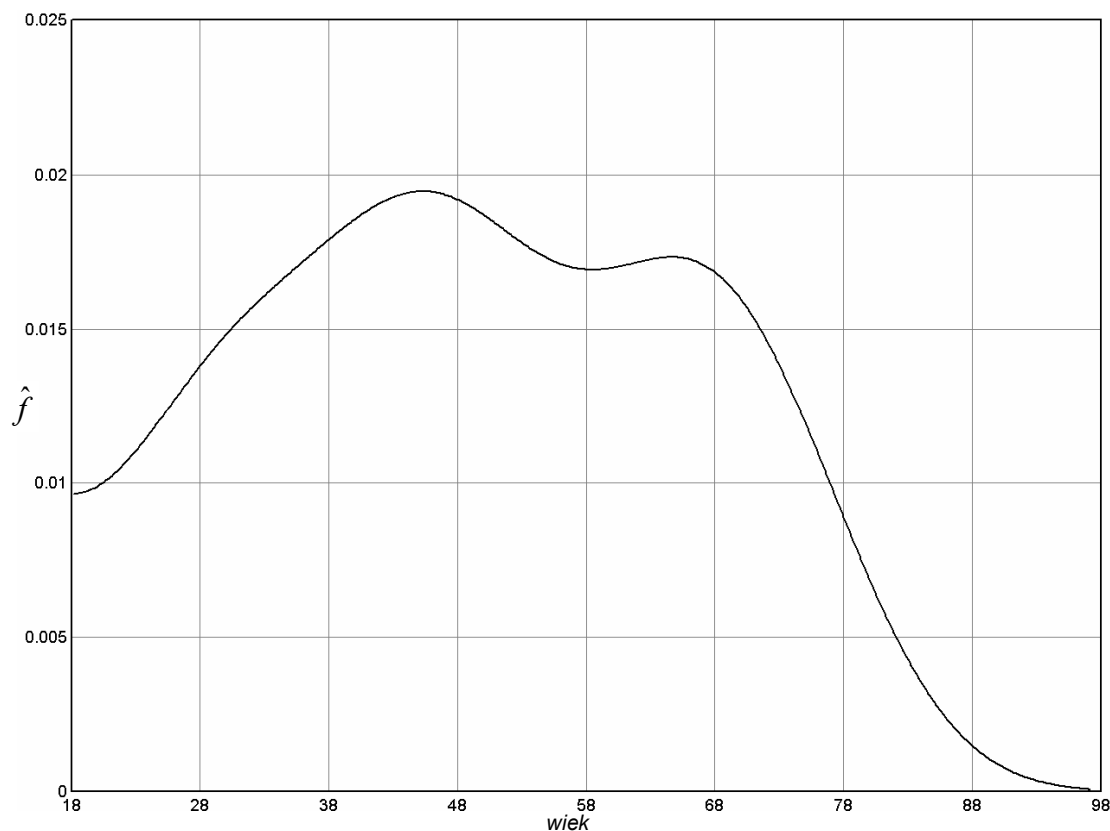
Rysunek 3.11. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej *wiek* – kandydat Marian Krzaklewski / $h = 6,56$ /



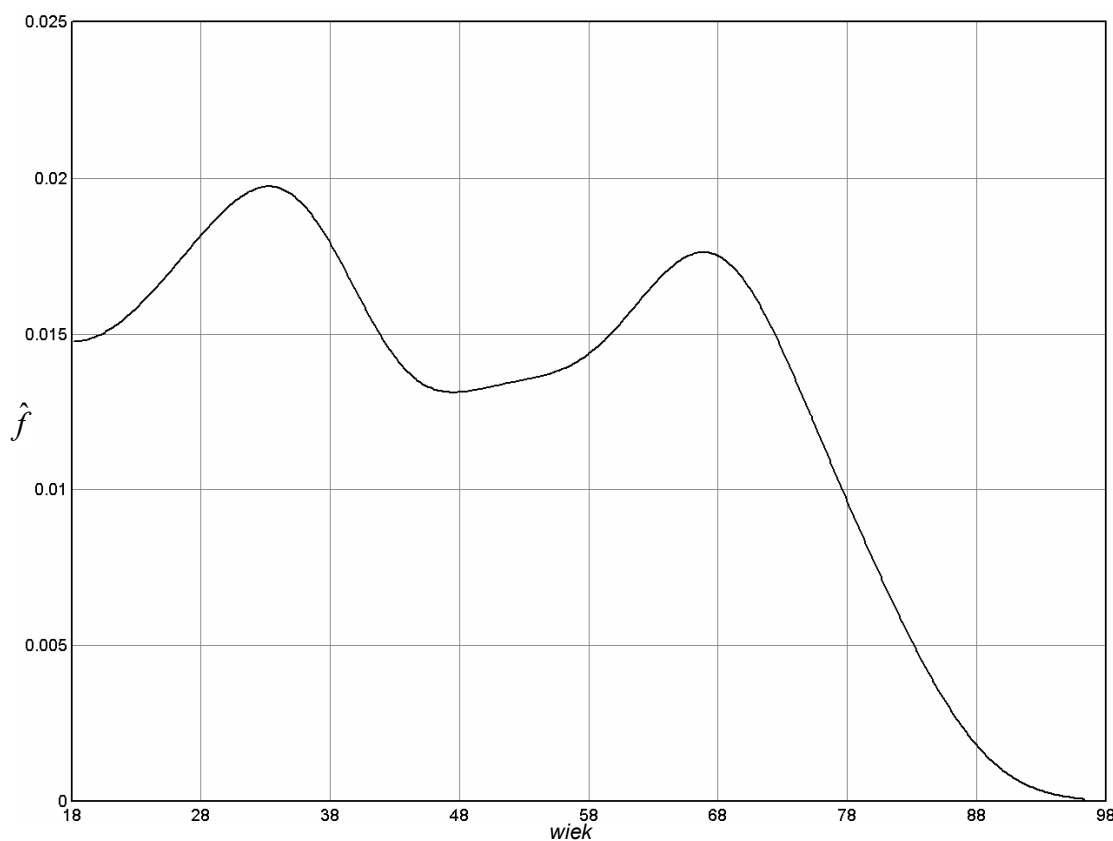
Rysunek 3.12. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej *wiek* – kandydat Aleksander Kwaśniewski / $h = 3,13$ /



Rysunek 3.13. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej *wiek* – kandydat Andrzej Olechowski / $h = 4,74$ /



Rysunek 3.14. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej $wiek$ – kandydat Lech Wałęsa / $h = 6,79$ /



Rysunek 3.15. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej $wiek$ – odpowiedź „nie pamiętam” / $h = 5,64$ /

Analiza otrzymanych wyników prowadzi do interesujących wniosków. W przypadku dwóch kandydatów – Mariana Krzaklewskiego i Lecha Wałęsy można zaobserwować, że relatywnie sporą część elektoratu stanowią ludzie starsi (od 65 roku życia) – dowodem tego jest występowanie w obu przypadkach, w tym zakresie, pojedynczej wartości modalnej rozkładu zmiennej losowej *wiek*.

Rozkład elektoratu Lecha Wałęsy cechuje się największą równomiernością – występują w nim dwa niezbyt wyraźne mody dla wartości zmiennej losowej *wiek* równych 45 i 65.

Struktura wiekowa elektoratu dwóch liderów sondażu, jak i również samych wyborów - Aleksandra Kwaśniewskiego i Andrzeja Olechowskiego jest podobna, z silnym maksimum rozkładu w okolicach 40-45 roku życia. Warto jednak zauważyć przesunięcie tego maksimum w przypadku drugiego z kandydatów w kierunku mniejszych wartości zmiennej losowej *wiek*. Przesunięcie to, wraz ze słabo zaznaczonym modem w okolicach 19 lat świadczy, że wśród głosujących na Andrzeja Olechowskiego występuje większy odsetek ludzi młodych.

W rozkładzie zmiennej losowej *wiek* dla grupy respondentów którzy na pytanie ankietera odpowiedzieli „nie pamiętam” posiada dwie wyraźne wartości modalne: w okolicach 32 i 65 roku życia. Można domniemywać, że odpowiedź taka wynika z niechęci do udzielania tego typu informacji. W przypadku 30-latków powodowane może być generalną niechęcią do polityki, a w przypadku ludzi starszych – nieufnością co do intencji pytającego i zachowania tajności odpowiedzi.

Dowodem na reprezentatywność danych sondażowych oraz skuteczność zastosowanej metody badawczej może być porównanie zaprezentowanych rezultatów z uzyskanymi poprzez tzw. exit-polls czyli “sondażu prowadzonego w czasie wyborów, w którym wyborcy wychodzący z lokali wyborczych (*polling station*) odpowiadają na pytanie na kogo głosowali” [13]. Skrót wyników takiego badania w trakcie wyborów prezydenckich w 2000 roku podaje [15]. Przedstawione są one w tabeli 3.2.

Rezultat tak przeprowadzonej weryfikacji jest pozytywny. Potwierdzone zostaje m.in. twierdzenie o względnie największym odsetku ludzi starszych wśród głosujących na Mariana Krzaklewskiego oraz informacja o strukturze elektoratu Andrzeja Olechowskiego, wskazująca na relatywnie duży w nim udział ludzi młodszych.

Tabela 3.2 Wyniki sondażu wyborczego (na zlecenie portalu *interia.pl*)

Wiek	Marian Krzaklewski	Aleksander Kwaśniewski	Andrzej Olechowski
18-24	10,2	15,8	18,9
25-39	23,0	26,7	31,3
40-59	38,5	40,9	38,4
60 i więcej	28,2	16,6	11,4

3.3. Wykrywanie uszkodzeń silnika asynchronicznego

Jedną z możliwości wykorzystania estymatorów jądrowych jest ich użycie do detekcji i diagnozy uszkodzeń. Niniejszy rozdział poświęcony został badaniu przydatności estymacji jądrowej do wykrycia uszkodzenia i próbie oszacowania jego wielkości w przypadku trójfazowego silnika asynchronicznego.

W pracy wykorzystano zbiory danych wyznaczonych numerycznie w zakresie częstotliwości 0-2500 Hz. Zostały one wygenerowane na podstawie modelu zbudowanego w programie *Matlab*. Opis powyższego modelu można znaleźć w publikacji [14]. Najważniejszy wzór, z którego korzystano m.in. przy wyznaczaniu prądów stojana ma postać:

$$\begin{bmatrix} \vdots \\ [0] \\ [U_\eta^s] \\ [0] \\ \vdots \\ \vdots \\ [0] \\ [0] \\ [0] \\ \vdots \end{bmatrix} = \{diag \begin{bmatrix} \vdots \\ [R^s] \\ [R^s] \\ [R^s] \\ \vdots \\ \vdots \\ [R^r] \\ [R^r] \\ [R^r] \\ \vdots \end{bmatrix} + diag \begin{bmatrix} \vdots \\ j(\eta\omega_0 + \omega)[E_s] \\ j\omega[E_s] \\ j(\eta\omega_0 - \omega)[E_s] \\ \vdots \\ \vdots \\ j(\eta\omega_0 + \omega)[E_r] \\ j\omega[E_r] \\ j(\eta\omega_0 - \omega)[E_r] \\ \vdots \end{bmatrix} \} \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \vdots & [L_0^s] & [L_1^s] & [L_2^s] & \dots & \dots & [M_0] & [M_1] & [M_2] & \dots \\ \vdots & \vdots & [L_{-1}^s] & [L_0^s] & [L_1^s] & \dots & [M_{-1}] & [M_0] & [M_1] & \dots \\ \vdots & \vdots & [L_{-2}^s] & [L_{-1}^s] & [L_0^s] & \dots & [M_{-2}] & [M_{-1}] & [M_0] & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ [I_{\eta,1}^s] \\ [I_{\eta,0}^s] \\ [I_{\eta,-1}^s] \\ \vdots \\ \vdots \\ [I_{\eta,1}^r] \\ [I_{\eta,0}^r] \\ [I_{\eta,-1}^r] \\ \vdots \end{bmatrix}$$

W równaniu tym $[R^{\{s,r\}}]$ oznaczają macierze rezystancji stojana oraz wirnika, $[L_i^{\{s,r\}}]$ – macierze indukcyjności własnych, $[M_l]$ – macierze indukcyjności wzajemnych, $[U_\eta^s]$ – wektory składowych harmonicznego napięcia stojana, $[I_{\eta,i}^{\{s,r\}}]$ – wektory składowych harmonicznego prądu stojana oraz wirnika.

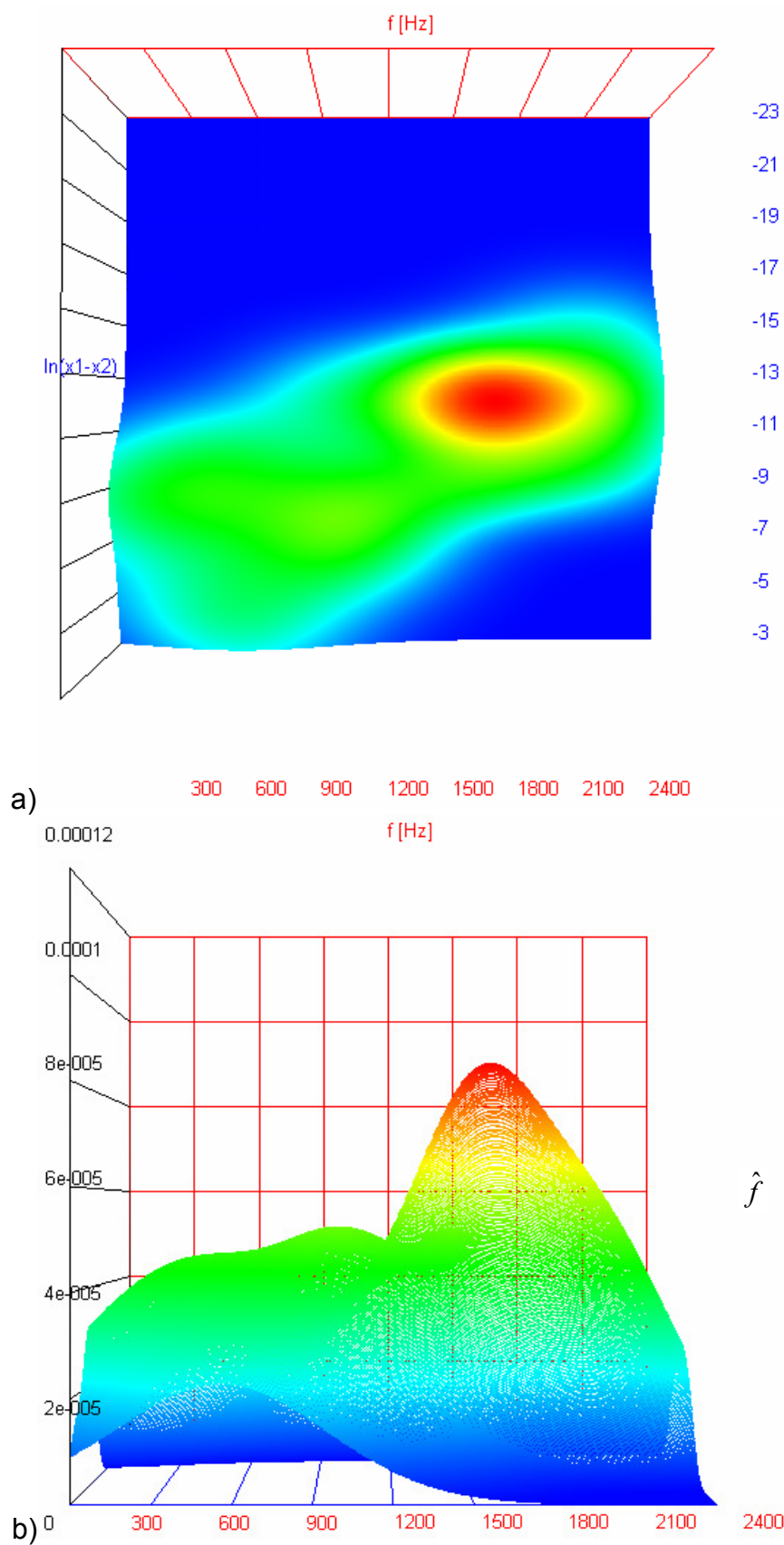
Założonym uszkodzeniem jest zmiana rezystancji jednego z prętów klatki wirnika. Oryginalne dane zawierają wartości składowych harmoniczných prądu dla danej fazy uzwojenia stojana, połączonego w gwiazdę bez przewodu zerowego wraz częstotliwościami tych harmoniczných. Powyższe dane dotyczą symulacji następujących przypadków:

- symetrii elektrycznej klatki wirnika (silnik w pełni sprawny),
- klatki, w której jeden z prętów ma zwiększoną rezystancję o 10% w stosunku do pozostałych (nieznaczne uszkodzenie silnika),
- klatki, w której jeden z prętów ma zwiększoną rezystancję o 100% w stosunku do pozostałych (poważne uszkodzenie silnika),
- klatki, w której jeden z prętów ma zwiększoną rezystancję 20 razy w stosunku do pozostałych (bardzo poważne uszkodzenie silnika).

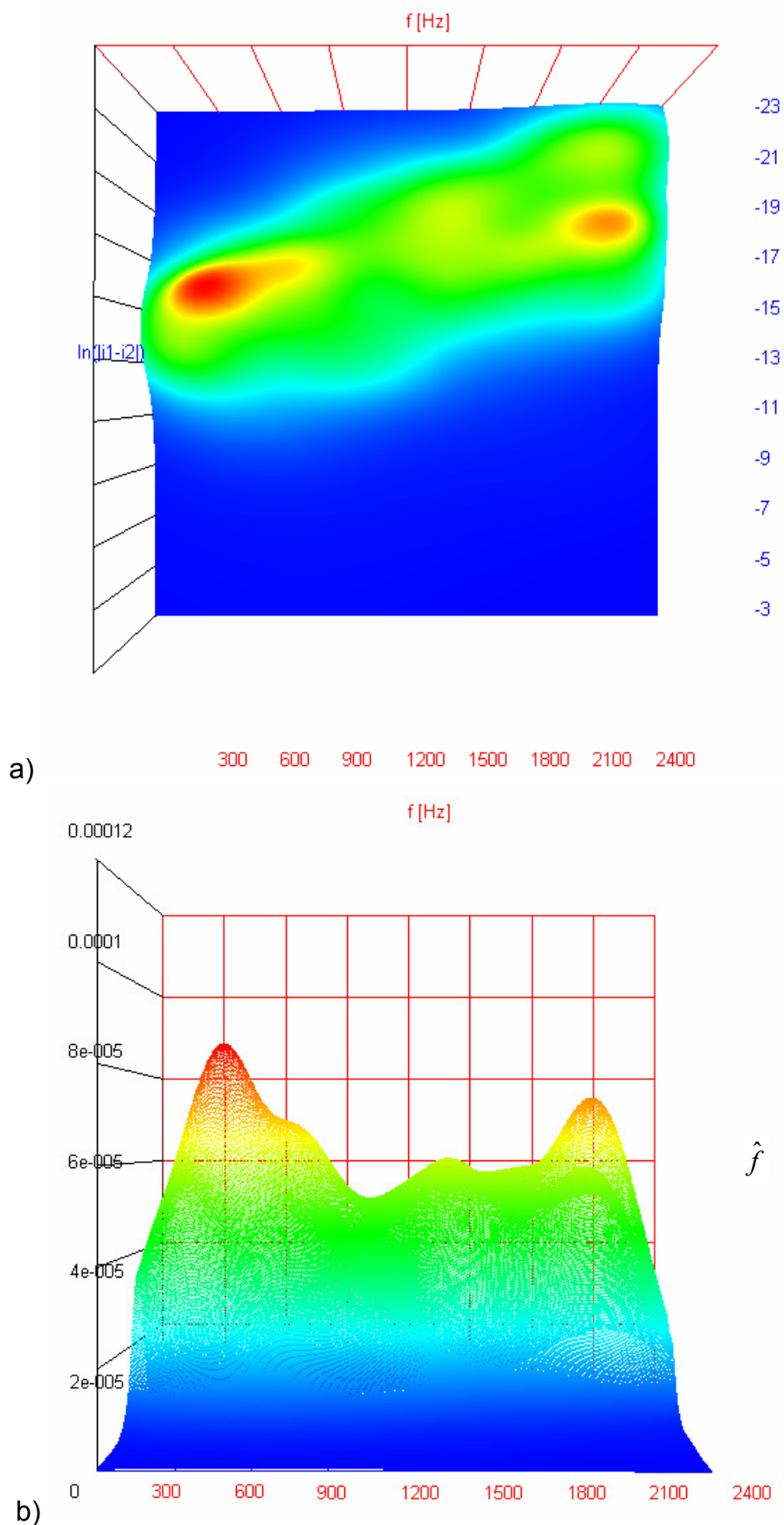
Wszystkie te przypadki zostały rozpatrzone dla poślizgu $s = 0,05$ oraz $s = 0,07$ (poślizg znamionowy).

W celu identyfikacji uszkodzeń silnika asynchronicznego stworzono nowe zmienne, które zostały wykorzystane w programie *KDEstim*. Powyższe dane zostały podzielone na 8 reprezentatywnych grup: po cztery dla poślizgu $s = 0,05$ oraz $s = 0,07$, w zależności od stopnia uszkodzenia silnika. Dla każdej składowej harmoniczných w poszczególnych grupach policzona została wartość bezwzględna z różnicy pomiędzy pierwszą a drugą składową prądu stojana. Ponieważ dla poszczególnych częstotliwości wartości te różniły się o kilka rzędów wielkości, zostały one przekonwertowane do skali logarytmicznej.

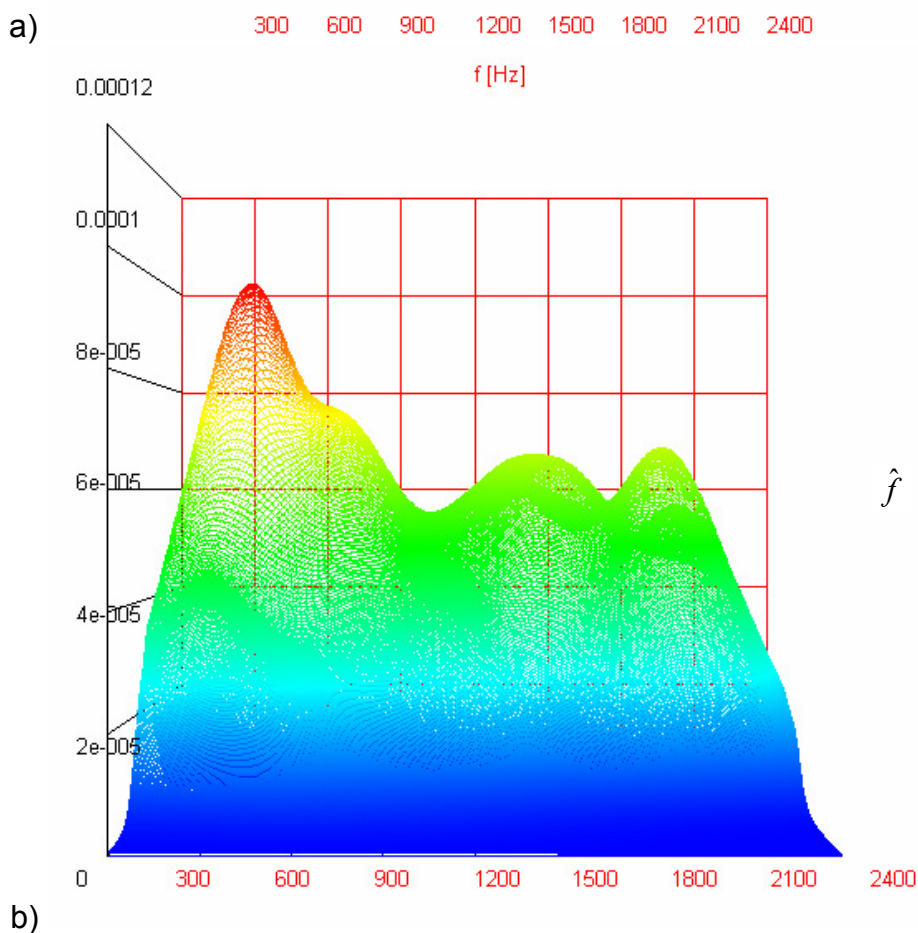
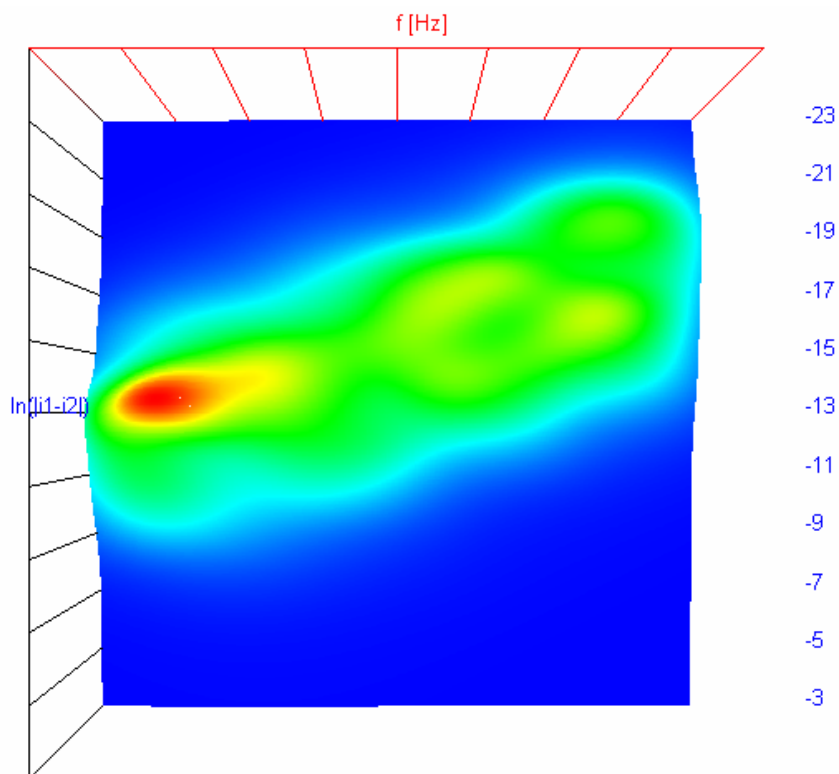
Powstała w ten sposób dwuwymiarowa zmienna losowa, której pierwszą składową jest *logarytm naturalny z wartości bezwzględnej różnicy dwóch harmoniczných składowych prądu uzwojenia stojana* a drugą - *częstotliwość harmoniczných*. Wyniki estymacji dla poślizgu $s = 0,05$ prezentują rysunki 3.16 – 3.19, a dla poślizgu $s = 0,07$ – rysunki 3.20-3.23.



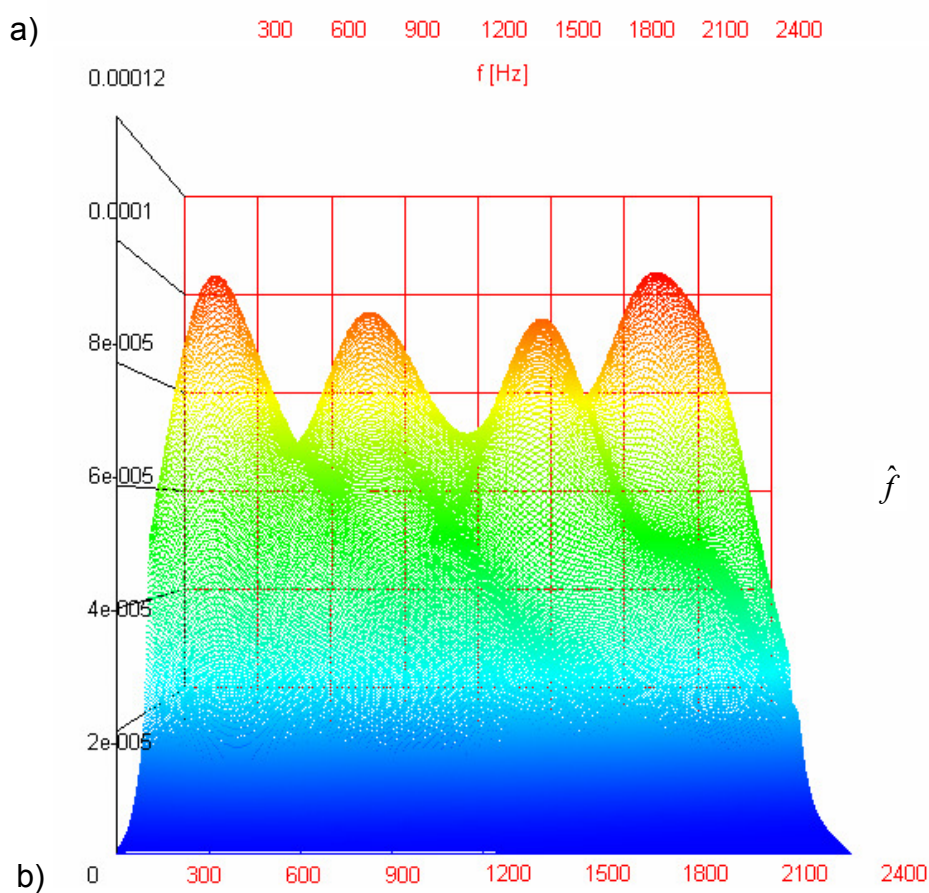
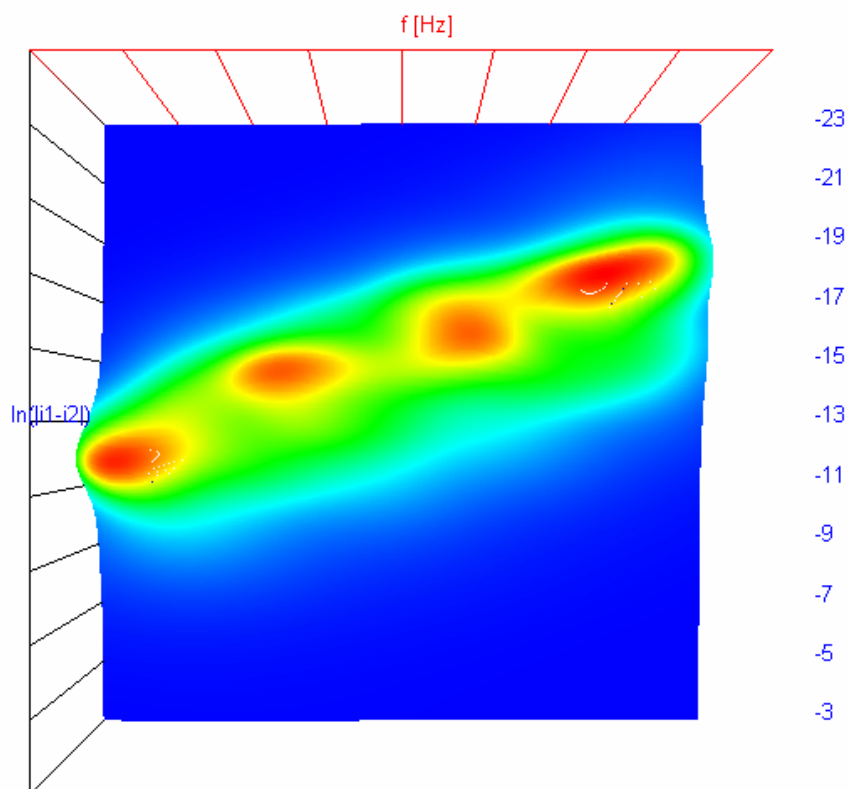
Rysunek 3.16. Wyniki estymacji dla sprawnego silnika (poślizg $s = 0,05$).



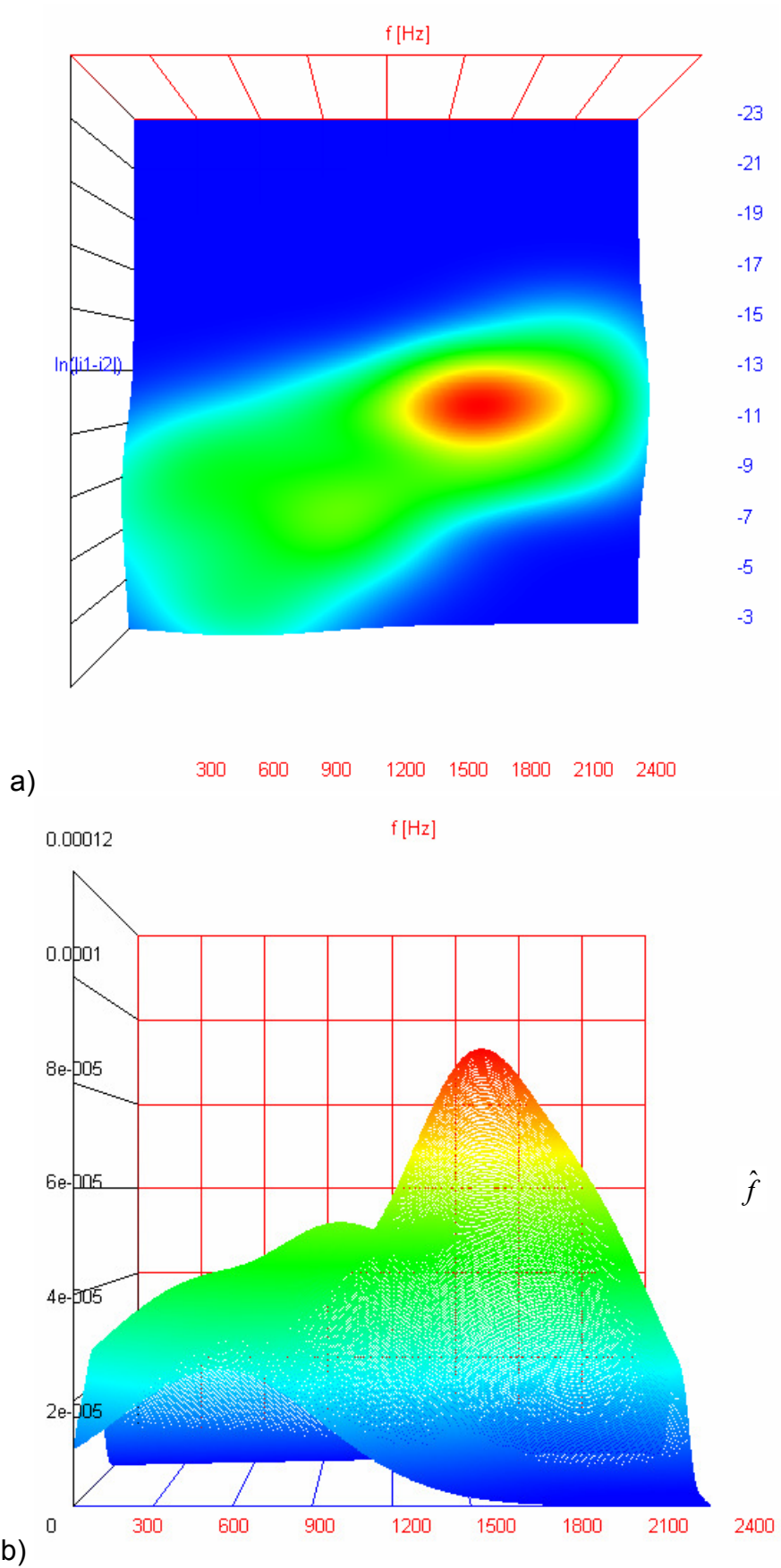
Rysunek 3.17. Wyniki estymacji dla silnika nieznacznie uszkodzonego - rezystancja jednego z uzwojeń jest większa o 10 % (poślizg $s = 0,05$).



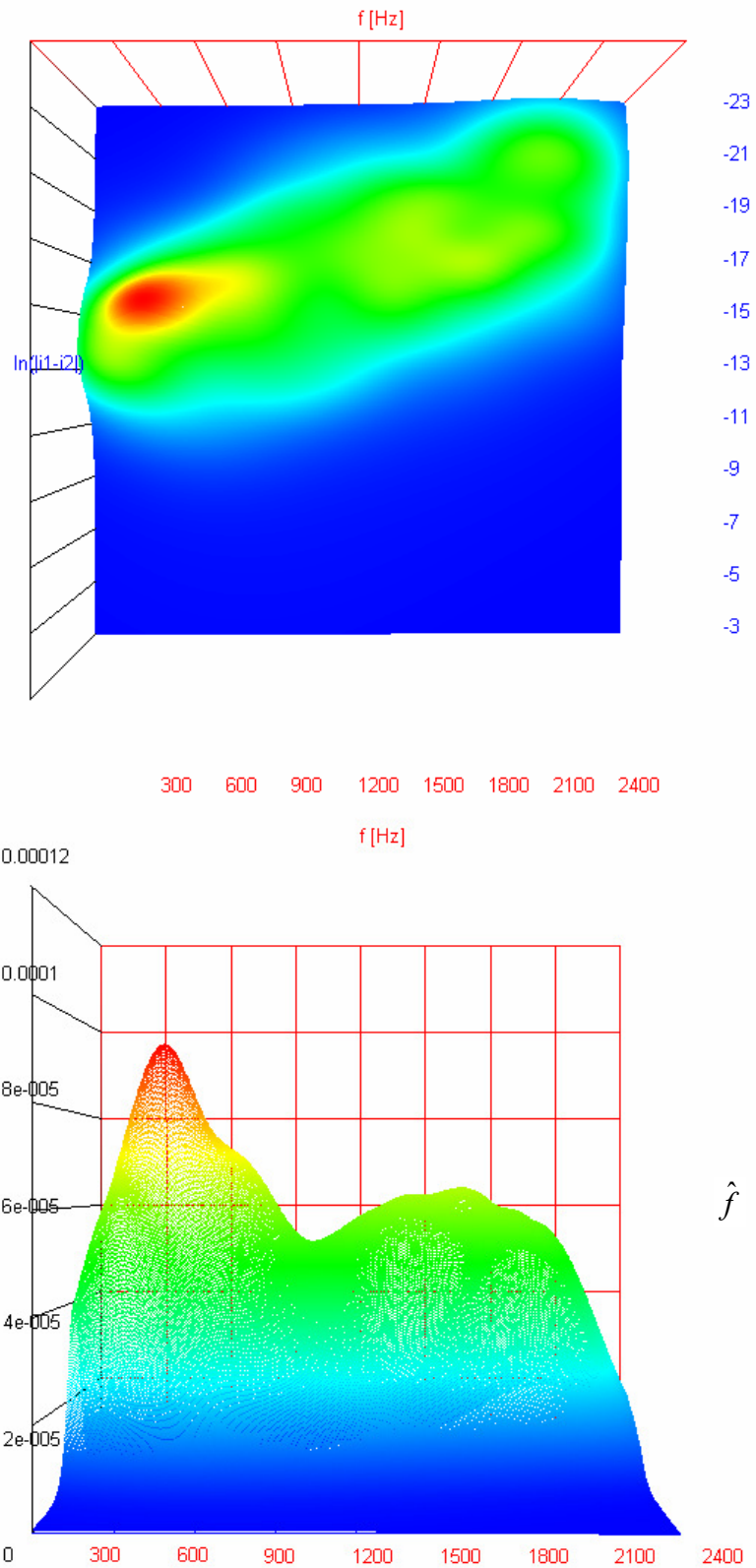
Rysunek 3.18. Wyniki estymacji dla średnio uszkodzonego silnika - rezystancja jednego z uzwojeń jest dwa razy większa (poślizg $s = 0,05$).



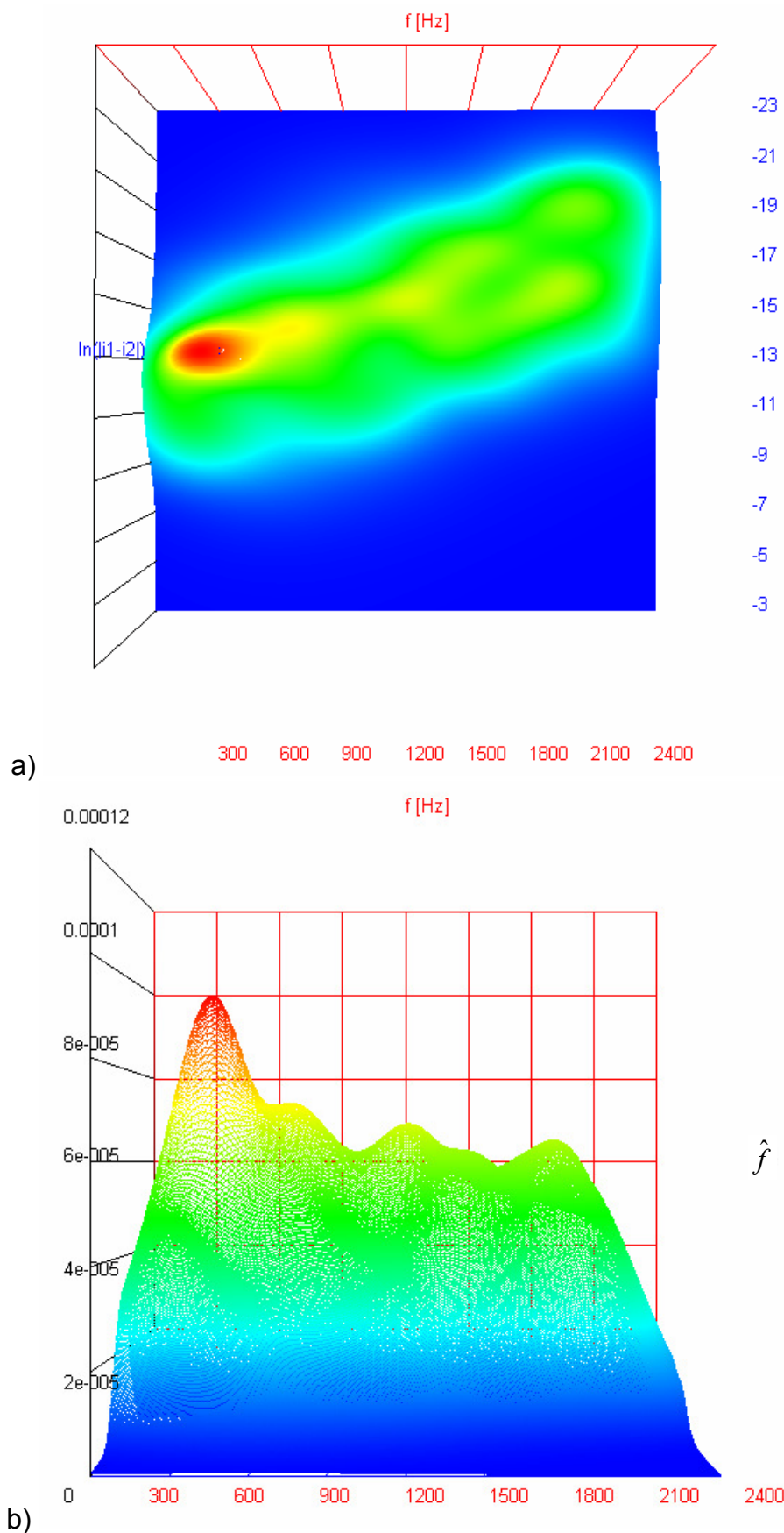
Rysunek 3.19. Wyniki estymacji dla całkowicie uszkodzonego silnika - rezystancja jednego z uzwojeń jest 20 razy większa (poślizg $s = 0,05$).



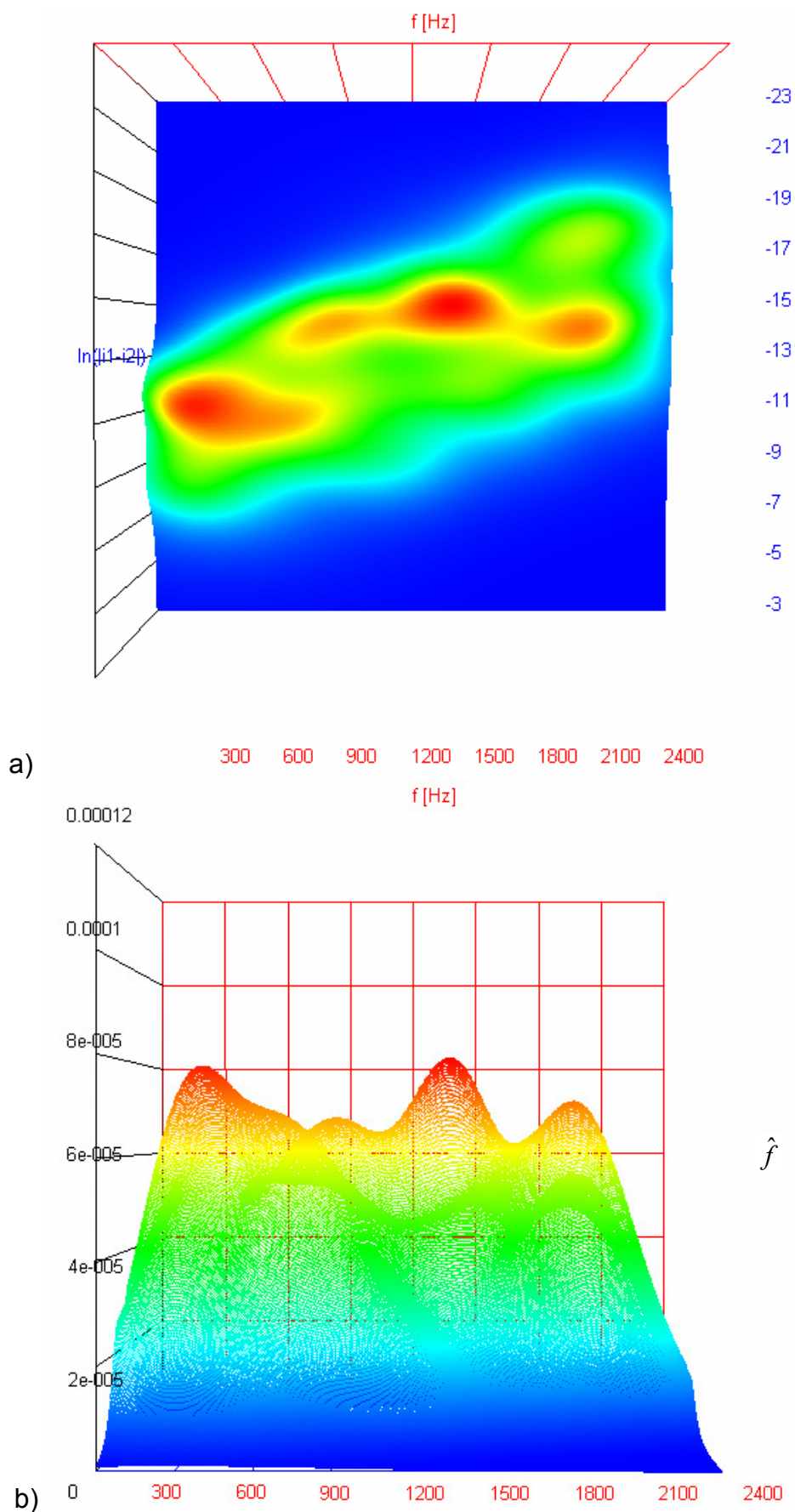
Rysunek 3.20. Wyniki estymacji dla sprawnego silnika (poślizg $s = 0,07$).



Rysunek 3.21. Wyniki estymacji dla silnika nieznacznie uszkodzonego - rezystancja jednego z uzwojeń jest większa o 10 % (poślizg $s = 0,07$).



Rysunek 3.22. Wyniki estymacji dla silnika średnio uszkodzonego - rezystancja jednego z uzwojeń jest dwa razy większa (poślizg $s = 0,07$).



Rysunek 3.23. Wyniki estymacji dla całkowicie uszkodzonego silnika - rezystancja jednego z uzwojeń jest 20 razy większa (poślizg $s = 0,07$).

Pierwszym wnioskiem dotyczącym powyższych wyników jest zdecydowane różnica kształtu funkcji gęstości dla silnika sprawnego, a tej samej funkcji dla silnika uszkodzonego (nawet przy niewielkim defekcie). Dotyczy to zarówno rozkładu maksimumów jak i położenia zakresu najbardziej prawdopodobnych wartości zmiennej losowej.

Dla sprawnego silnika, zarówno dla poślizgu $s = 0,05$ jaki i $s = 0,07$, zauważyć można wyraźne maksimum dla wyższych częstotliwości i różnicy prądów ok. $16 \mu\text{A}$. Natomiast w przypadku silnika uszkodzonego (w niewielkim lub średnim stopniu) powyżej opisane maksimum zanika, pojawia się z kolei inne – dla częstotliwości znacznie mniejszych. Druga współrzędna przyjmuje dla tego maksimum wartości pomiędzy -14 a -16 (po przeliczeniach różnica prądów wynosi od $0,8 \mu\text{A}$ do $0,1 \mu\text{A}$).

Bardzo ciekawy rozkład można zaobserwować dla najbardziej uszkodzonego silnika. Dla poślizgu $s = 0,05$ cechuje się on 4 wyraźnymi modami dla częstotliwości równych około 200, 800, 1400 oraz 1800 Hz. Dla poślizgu $s = 0,07$ rozkład ten znacznie różni się też od pozostałych przypadków uszkodzeń.

Przede wszystkim warto jednak zauważyć przesunięcie „pasma” rozkładu w przypadku silnika uszkodzonego – w kierunku mniejszych różnic prądów stojana.

Na podstawie powyższych obserwacji można stwierdzić, że stosowanie estymatorów jądrowych do detekcji uszkodzeń daje dobre rezultaty. Jednak ich przydatność, w tej postaci, do diagnozy uszkodzeń nie została w pełni potwierdzona (choć proponowana metoda rozróżnia całkowite uszkodzenie silnika). Wprowadzenie dodatkowej zmiennej do analizy prawdopodobnie wpłynęłoby na lepsze określenie wartości uszkodzenia i jego przypuszczalną lokalizację.

3.4. Analiza wybranych rozkładów z zakresu fizyki wysokich energii

Obiektem rozważań w tej części pracy jest identyfikacja rozkładu zmiennych losowych mierzonych w eksperymencie fizyki wysokich energii. Przedstawione próby zostały uzyskane przy pomocy detektora ZEUS na akceleratorze HERA w DESY w Hamburgu. Dane doświadczalne zgromadzono w dedykowanych pomiarach mających na celu precyzyjne wyznaczenie wartości przekroju czynnego na *fotoprodukcję*. Dane symulowane uzyskane przy pomocy metod Monte Carlo dotyczą procesu promieniowania hamowania – *bremsstrahlungu*. Szczegółowe informacje dotyczące strony teoretycznej omawianych zjawisk, jak i eksperymentów akceleratora HERA, w którym były one badane można znaleźć w artykułach [1], [2] oraz [3].

3.4.1. Analiza procesu fotoprodukcji

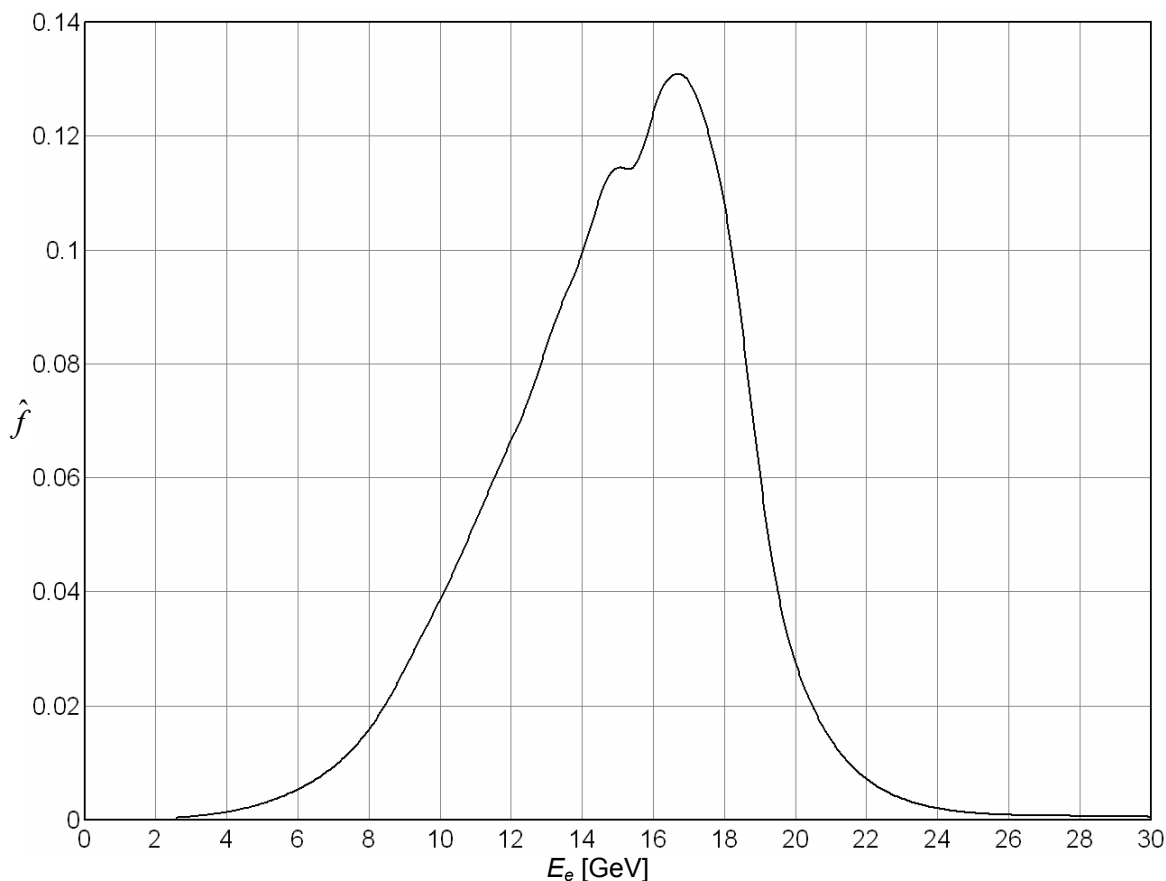
W procesie *fotoprodukcji* rzeczywiste fotony oddziaływując z materią (w tym wypadku z protonami) produkują hadrony w stanie końcowym. Przypadki oddziaływań, dla których obserwowano pozytron (antycząstkę elektronu) w dedykowanym kalorymetrze elektronowym [1] i jednocześnie obserwowano znaczącą aktywność w głównym kalorymetrze detektora ZEUS, były akceptowane i zapisywane na dysku. W czasie trwania eksperymentu wybrano szerszą klasę oddziaływań (spełniających łagodniejsze żądania) do zapisania na dysku. Miało to na celu z jednej strony możliwość kalibracji aparatury doświadczalnej a z drugiej kontrolę procesów składających się na tło. W szczególności dane te zawierały znaczny procent danych pochodzących z procesu *bremsstrahlungu* czy też promieniowania w stanie początkowym (patrz: [3]). Dane te pochodzą z 1996 roku, gdy akcelerator HERA zderzał ze sobą wiązkę pozytronów o energii 27,5 GeV z wiązką protonów o energii 820 GeV.

Z otrzymanych w ten sposób danych wyodrębniono próby następujących zmiennych losowych:

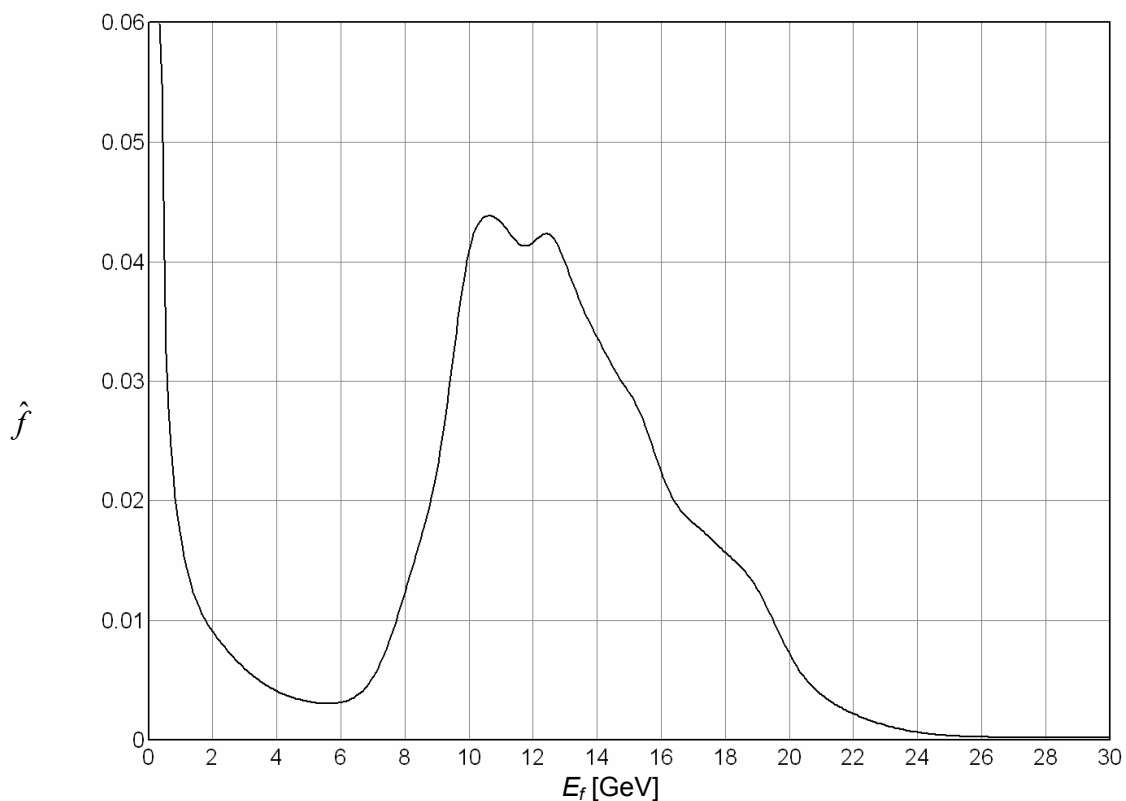
- *elume* – energia elektronu, mierzona przez kalorymetr elektronów, wyrażona w GeV (liczność: 4395),
- *elumg* – energia fotonu, mierzona przez kalorymetr fotonowy [1], wyrażona w GeV (liczność: 4395),

- $elumevsg$ – energia elektronu względem energii fotonu (zmienna dwuwymiarowa o liczności: 4395),
- z_vertex – położenie wierzchołka oddziaływania, w cm (liczność: 4395).

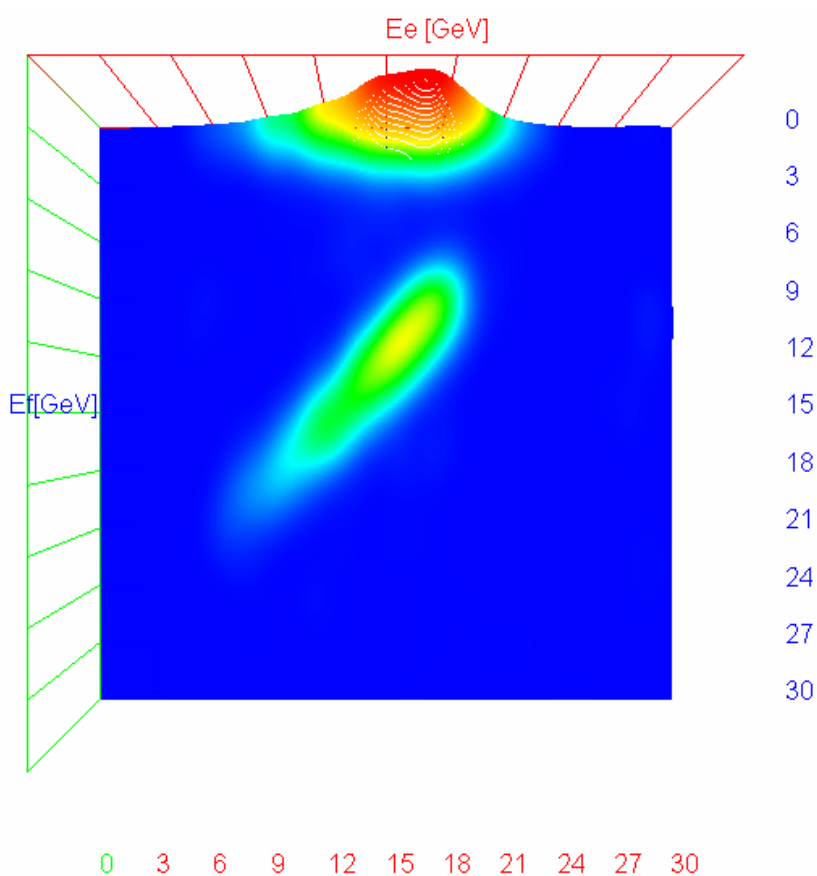
Rysunki 3.24 – 3.27 prezentują rozkłady prawdopodobieństwa wyżej wymienionych zmiennych losowych – uzyskane poprzez zastosowanie estymatorów jądrowych o współczynnikach wygładzania uzyskanych metodą *plug-in* (rozkłady $elume$, $elumg$, z_vertex) oraz przy zastosowaniu wzoru (1.52) (rozkład $elumevsg$). W przypadku rozkładów zmiennych $elume$, $elumg$, $elumevsg$ skorzystano również z lewostronnego ograniczenia nośnika w zerze, (aby oddać rzeczywisty charakter zmiennej losowej, która nie może przyjmować wartości ujemnych). Dla rozkładów $elume$, $elumg$, z_vertex zastosowano także modyfikację parametru wygładzania, natomiast w przypadku $elumevsg$ transformację liniową.



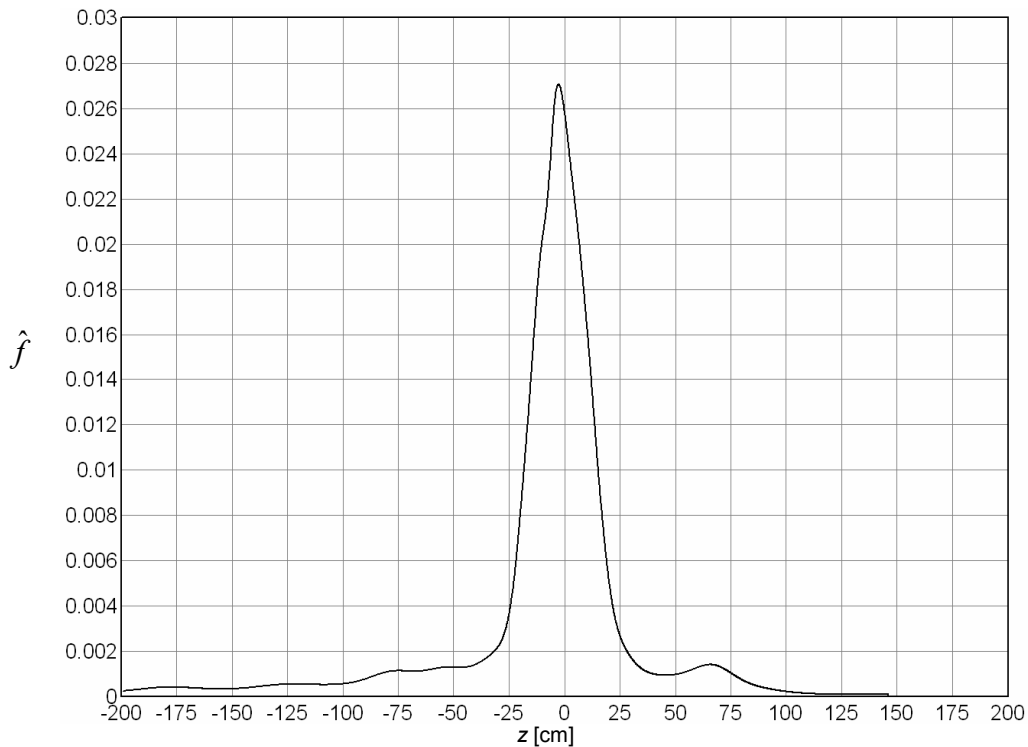
Rysunek 3.24. Rozkład energii elektronu mierzonej przez kalorymetr elektronowy. Dane dotyczące procesu fotoprodukcji $/h = 0,53/$.



Rysunek 3.25. Rozkład energii fotonu mierzony w kalorymetrze fotonowym. Dane dotyczące fotoprodukcji /h = 0,17/.



Rysunek 3.26. Dwuwymiarowy rozkład energii elektronu względem energii fotonu w danych dotyczących fotoprodukcji /h = 0,17/



Rysunek 3.27. Rozkład położenia wierzchołka oddziaływania elektron-proton w zjawisku fotoprodukcji / $h = 3,29$ /

Gęstość rozkładu prawdopodobieństwa energii elektronu posiada jedno wyraźne maksimum (dla ok. 16-17 GeV). Kształt rozkładu jak i położenie maksimum są zgodne z wynikami zaprezentowanymi w pracy [3]. Zakres energii elektronu wynika z własności pól magnetycznych użytych do prowadzenia wiązki elektronowej oraz własności geometrycznych i dynamicznych tej wiązki (na przykład: rozbieżności kątowej).

Rozkład energii fotonu jest ciągły z maksimum w zerze i drugim szerokim maksimum w okolicy 10-13 GeV. Jak wspomniano wyżej dane te zawierają dane z procesu *fotoprodukcji*, promieniowania w stanie początkowym i *bremstrahlungu*. Dla czystej *fotoprodukcji* nie obserwuje się fotonu w *kalorymtrze* fotonowym. Szerokie maksimum w zerze odpowiada przypadkom czystej *fotoprodukcji* i *fotoprodukcji* z promieniowaniem w stanie początkowym lub stowarzyszonym z promieniowaniem synchrotronowym. Szerokie maksimum w okolicy 10-13 GeV pochodzi od przypadków *bremstrahlungu*.

Łączna mierzona energia rozproszonego elektronu i fotonu w przypadku promieniowania hamowania była w przybliżeniu równa energii pierwotnej wiązki elektronów (27,5 GeV), co obrazuje wydłużony mniejszy „garb” rozkładu (rysunek 3.26), zakładając, że obie energie są różne od zera. „Garb” ten obrazuje anty-korelację między energiami elektronu i fotonu.

W przypadku, gdy energia fotonu mierzona przez kalorymetr jest równa lub bliska zeru widmo energii elektronu posiada wyraźne maksimum w 16-17 GeV.

Rozkład położenia wierzchołka oddziaływania ma jedno wyraźne maksimum (w okolicy 0 cm). W kierunku dodatnim krzywa rozkładu zanika do zera dla ok. 110 cm, ujemnym – dla -200 cm. Asymetryczność ta wynika ze struktury wiązek protonowej i elektronowej oraz ze struktury czasowej akceleratora HERA.

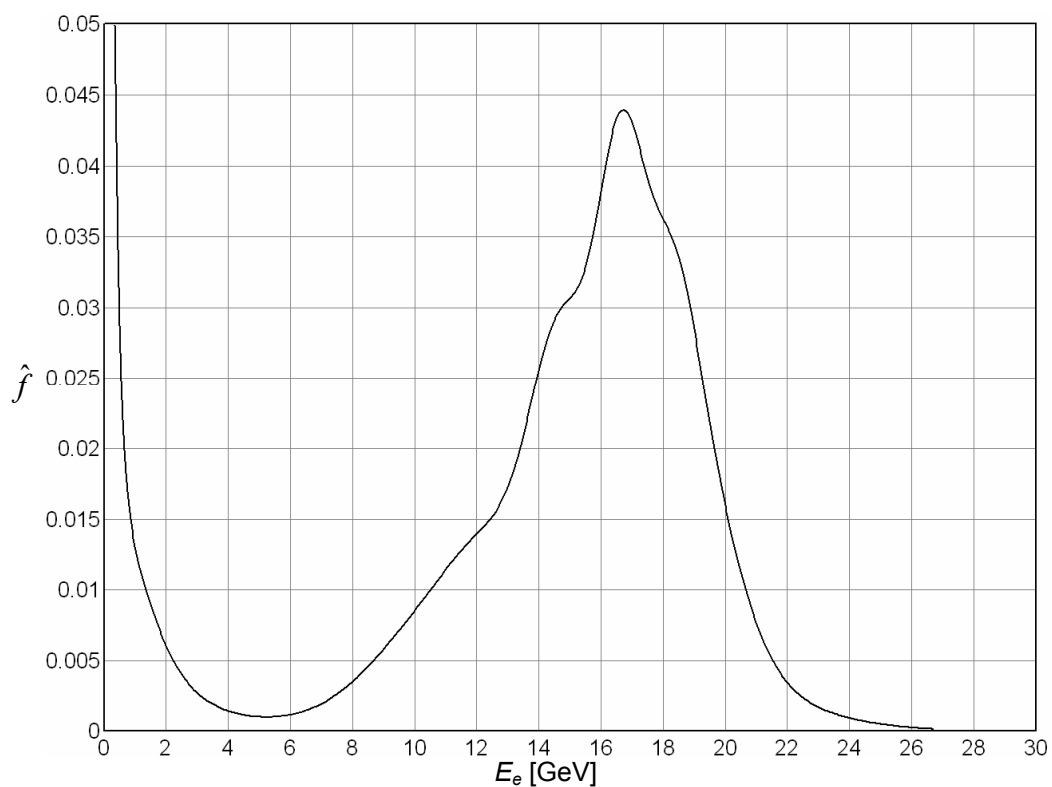
3.4.2. Analiza zjawiska bremsstrahlungu

Próby drugiej grupy zmiennych losowych otrzymano z wyników symulacji Monte-Carlo zjawiska *bremsstrahlungu* podczas zderzeń elektronów o energii 27,5 GeV z protonami o energii 820 GeV. *Bremsstrahlung* (niem. *bremsen* – hamować, *strahlung* – promieniowanie) to promieniowanie hamowania emitowane przez naładowaną cząstkę poruszającą się (wyhamowywaną) w polu elektrycznym.

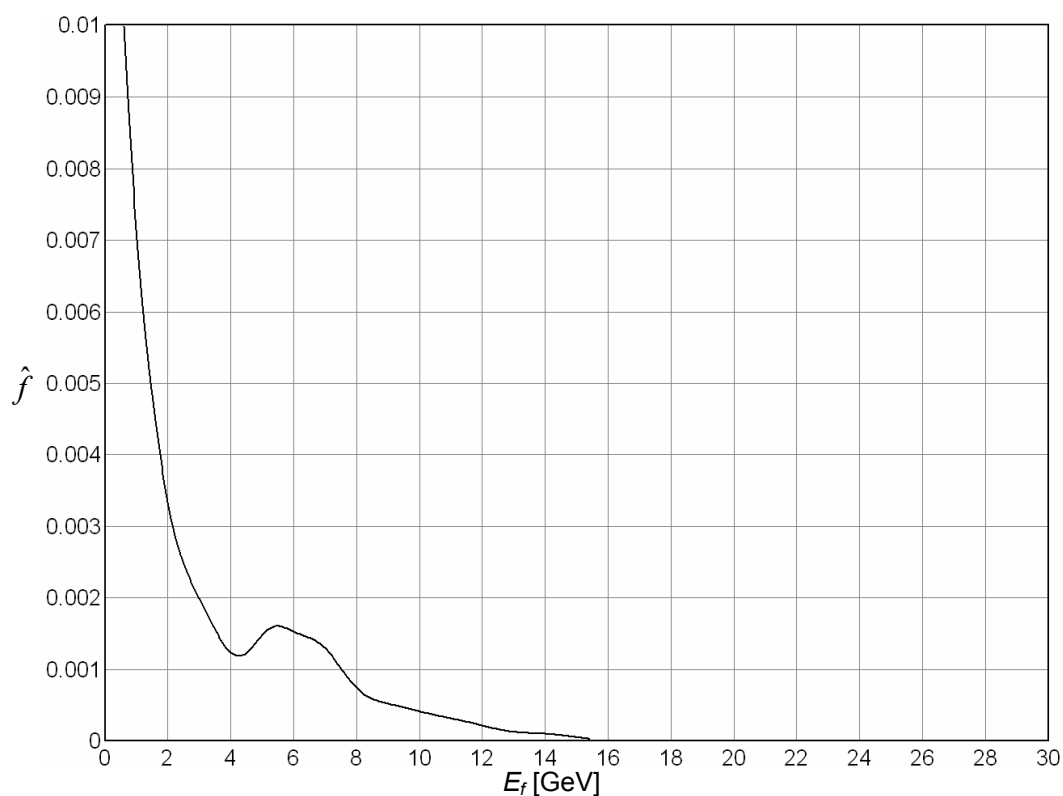
W przypadku tego zjawiska przedmiotem analizy były rozkłady następujących zmiennych:

- *elume* – energia elektronu, wyrażona w GeV (liczność: 4395),
- *elumg* – energia fotonu, wyrażona w GeV (liczność: 4395),
- *elumevsg* – energia elektronu, energia fotonu (zmienna dwuwymiarowa o liczności: 4395),
- *z_vertex* – położenie wierzchołka oddziaływania, w cm (liczność: 4395).

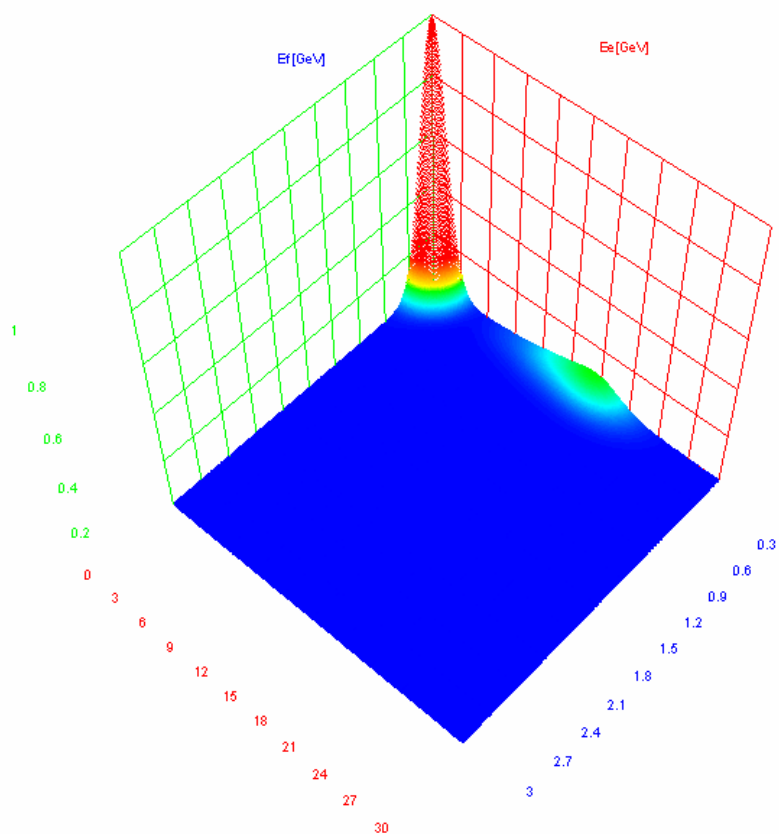
Estymaty rozkładów tychże zmiennych uzyskane przy użyciu tej samej metodologii jak w przypadku rozkładów w zjawisku *fotoprodukcji* przedstawiają rysunki 3.28 – 3.31.



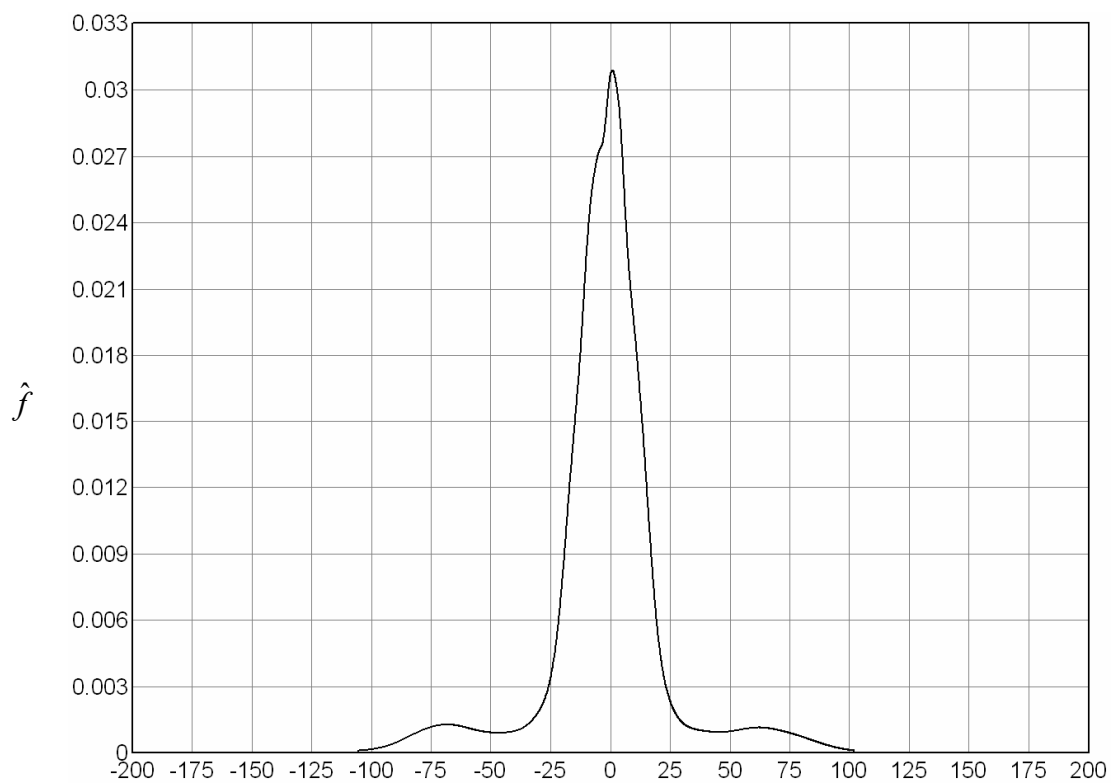
Rysunek 3.28. Rozkład energii elektronu mierzonej przez kalorymetr elektronowy w symulacji zjawiska bremsstrahlungu / $h = 0,17$ /



Rysunek 3.29. Rozkład energii fotonu mierzonej przez kalorymetr fotonowy w symulacji zjawiska bremsstrahlungu / $h = 0,01$ /



Rysunek 3.30. Rozkład energii elektronu względem fotonu mierzonej w symulacji zjawiska bremsstrahlungu / $h = 0,20$ /



Rysunek 3.31. Symulowany rozkład położenia wierzchołka oddziaływania elektron-proton / $h = 2,52$ /

Rozkład energii elektronu posiada szerokie maksimum w zakresie 10-13 GeV oraz silne maksimum w zerze. Własności tego rozkładu są w bardzo dużym stopniu określone poprzez własności geometryczne i dynamiczne wiązki elektronowej oraz własności geometryczne i pomiarowe kalorymetru elektronowego.

Krzywa rozkładu energii fotonu jest zbliżona do idealnej, przewidywanej przez teorię krzywej logarytmicznej. Maksimum w zerze jest związane z budową układu kalorymetru fotonowego, a zwłaszcza z jego akceptancją geometryczną i energetyczną.

Kształt rozkładu *elumevsg* dla niewielkiej energii fotonu (poniżej 3GeV) jest wynikiem skomplikowanego złożenia czynników takich jak: geometria aparatury, wiązki elektronowej i zdolności rozdzielczej przyrządów pomiarowych.

Rozkład wierzchołka oddziaływania jest rozłożony symetrycznie względem zera, w którym posiada wyraźne maksimum, i zanika dla ok. 100 cm.

Przedstawione rezultaty estymacji są zgodne z wynikami wcześniej przeprowadzanych badań w zakresie rozkładów energii w zjawiskach *fotoprodukcji* i *bremsstrahlungu*. Pozwalają również na wyciągnięcie podobnych wniosków dotyczących natury rozważanych zjawisk. Metodyka estymatorów jądrowych znajduje zatem swe efektywne zastosowanie również w złożonej, acz interesującej dziedzinie fizyki wysokich energii.

Rozdział 4. PODSUMOWANIE

W niniejszej pracy przedstawiono problem konstruowania estymatora jądrowego rozkładu prawdopodobieństwa zmiennej losowej. W celu przeprowadzenia numerycznej, eksperymentalnej analizy danych przy użyciu tej metody, stworzono program komputerowy *KDEstim*. Pozwala on na wyznaczenie estymatora jądrowego zmiennej losowej na podstawie danej próby, obliczenie kilkoma metodami współczynnika wygładzania właściwego dla rozważanego problemu, a także wizualizację i zapis otrzymanych wyników.

W programie zaimplementowano trzy procedury pozwalające na poprawę jakości estymatora: transformację liniową, modyfikację parametru wygładzania oraz ograniczenie nośnika zmiennej losowej. Warto również podkreślić, że nie ograniczono się jedynie do przypadku rozkładów jedno- i dwuwymiarowych. Program *KDEstim* pozwala bowiem na rozważanie próby n -wymiarowej – jedynym ograniczeniem są tu: wielkość pamięci i moc obliczeniowa komputera na którym program ów działa. O otwartości przedstawionego narzędzia świadczy także użycie plików tekstowych jako formatu obsługiwanych danych. Są one dogodne do uzyskania – jako źródła prób i odpowiednie do dalszej analizy – jeśli zawierają wyniki estymacji.

W toku pracy poddano analizie kilka rozkładów pochodzących z wybranych systemów rzeczywistych z zakresu inżynierii, nauk przyrodniczych i socjologicznych. Analiza ta, poza dostarczeniem interesujących wniosków (podanych w każdym z rozdziałów jej poświęconych), zweryfikowała pozytywnie skuteczność zastosowanej metodologii oraz jakość stworzonego programu.

Praca niniejsza stanowi również cenny impuls w zakresie przyszłych badań prowadzonych w dziedzinie estymatorów jądrowych. Zauważono problem w zastosowaniu metody krzyżowego uwiarygodniania w przypadku skwantowanych danych oraz

zweryfikowano skuteczność randomizacji jako przykładowego sposobu jego rozwiązania. Również drastycznie wzrastający, z licznością i wymiarem próby, czas obliczeń nasunął pomysł zastosowania metod przetwarzania równoległego w celu wyznaczenia estymatora jądrowego zmiennej losowej wielowymiarowej.

Praca, poza wartością naukowo-dydaktyczną, posiada również znaczenie popularyzatorskie, przedstawiając zastosowania metody estymatorów jądrowych, która jest w Polsce stosunkowo nieznana i rzadko stosowana. W tym celu dołączono do niej także krótki opis implementacji metody estymacji jądrowej w istniejących narzędziach statystycznych (patrz: dodatek B). Jest to o tyle cenne, że analizując wyniki otrzymane w wyniku użycia tej metody oraz duży potencjał możliwości, jakie ona posiada, można zalecać jej częstsze stosowanie w wielu problemach z dziedziny szeroko rozumianej analizy i eksploracji danych.

DODATEK A. OPIS ZAWARTOŚCI PŁYTY CD

Do niniejszej pracy dołączono nośnik CD zawierający:

- program *KDEstim* (w wersji instalacyjnej i wykonywalnej) oraz jego kod źródłowy
– w katalogu *\KDEstim*,
- próby badanych rozkładów oraz wyniki estymacji
– w katalogu *\data*,
- kod źródłowy programu w języku Visual Basic pakietu Statistica wyznaczającego estymator jądrowy gęstości rozkładu prawdopodobieństwa zmiennej losowej wielowymiarowej (patrz: dodatek B)
– w katalogu *\Statistica*,
- dokument Microsoft Word zawierający tekst pracy
– w katalogu *\documentation*.

DODATEK B. ESTYMATORY JĄDROWE W DOSTĘPNYCH NARZĘDZIACH STATYSTYCZNYCH

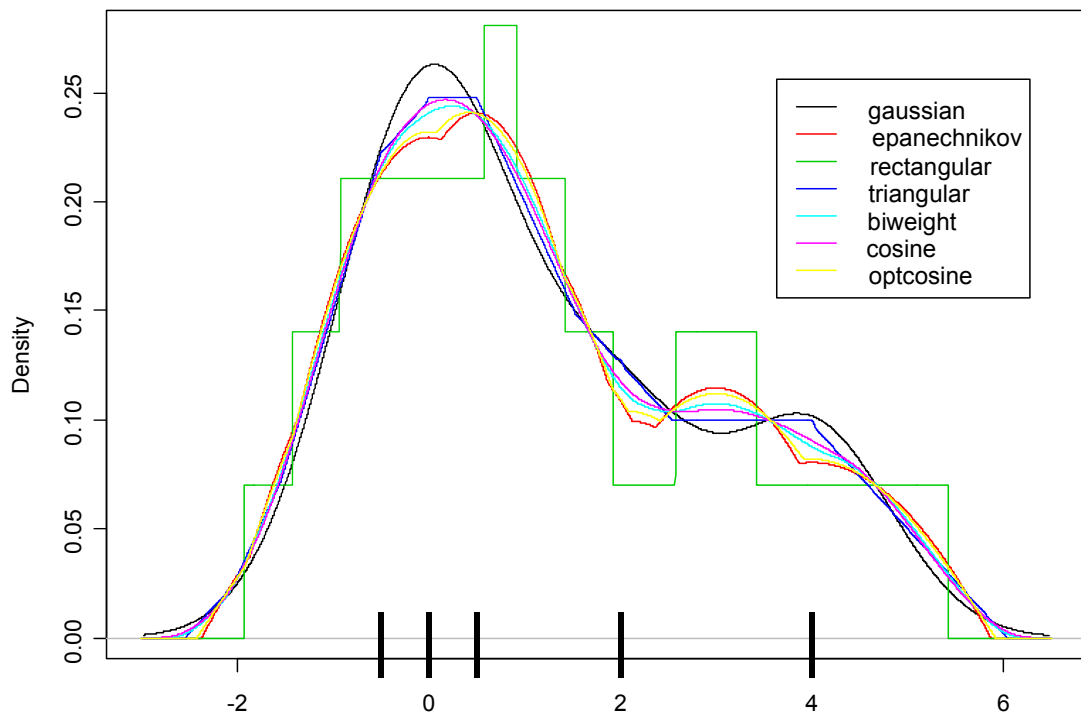
Skuteczną i szybką analizę statystyczną można przeprowadzić przy użyciu wielu istniejących na rynku programów ku temu celu przeznaczonych. Celem niniejszego rozdziału jest przedstawienie metodologii procesu estymacji przy użyciu estymatorów jądrowych w dwóch liczących się środowiskach służących do obliczeń statystycznych: w darmowym programie *R* (<http://www.r-project.org/>) i w rozwijanym przez firmę Statsoft pakiecie *Statistica* (<http://www.statsoft.pl/>)

Jak podaje internetowa encyklopedia Wikipedia „*GNU R* jest językiem programowania i środowiskiem do obliczeń statystycznych i wizualizacji wyników. Jest to projekt GNU podobny do języka i środowiska *S* stworzonego w Bell Laboratories (dawniejsze AT&T, obecnie Lucent Technologies) przez Johna Chambersa i jego współpracowników. *R* może być traktowane jako implementacja języka i całego środowiska *S*” [16].

W *R* proces estymacji gęstości rozkładu prawdopodobieństwa zmiennej losowej jedno- lub dwuwymiarowej można przeprowadzić przy użyciu odpowiednich funkcji z pakietów *stats* i *MASS*. I tak: funkcja *density* (pakiet *stats*) pozwala na uzyskanie estymatora gęstości dla zmiennej jednowymiarowej, z określonym jądrem (w wersji 2.1.1 pakietu *stats* dostępne są jądra Epanecznikowa, jednostajne, dwuwagowe, normalne, trójkątne i cosinusowe) i podanym parametrem wygładzania (możliwy jest również dobór tego parametru przy pomocy metod opisanych w rozdziale 1.5). Natomiast estymator rozkładu zmiennej losowej dwuwymiarowej otrzymuje się w *R* przy użyciu funkcji *kde2d* z pakietu *MASS*.

Rysunek B.1 został wykonany w programie R. Przedstawia on porównanie wyników estymacji dla wyżej wymienionych typów jąder. Kod napisany w języku R ma następującą postać:

```
u=c(-0.5,0,0.5,2,4)
(kernels <- eval(formals(density.default)$kernel))
h.f <- sapply(kernels, function(k)density(kern = k, give.Rkern = TRUE))
(h.f <- (h.f["gaussian"] / h.f)^ .2)
bw <- bw.SJ(u) ## sensible automatic choice
plot(density(u, bw = bw, n = 2^13))
for(i in 2:length(kernels))
  lines(density(u, bw = bw, adjust = h.f[i], kern = kernels[i],
              n = 2^13), col = i)
legend(55, 0.035, legend = kernels, col = seq(kernels), lty = 1)
points(u,0*u,pch="|",cex=3)
```



Rysunek B.1. Porównanie kilku typów jąder (program R)

Statistica jest uniwersalnym, zintegrowanym systemem służącym do statystycznej analizy danych, tworzenia wykresów, operowania na bazach danych, wykonywania transformacji danych i tworzenia aplikacji. W skład systemu wchodzi wszechstronny zestaw zaawansowanych procedur analitycznych, stosowanych w nauce, biznesie, technice oraz zgłębianiu danych” [15].

Pakiet *Statistica* nie zawiera, w swej podstawowej wersji, procedur pozwalających na sprawne wyznaczenie estymatora gęstości prawdopodobieństwa metodyką estymatorów jądrowych. Pozwala jednakże na tworzenie własnych programów w języku Visual Basic – nic nie stoi zatem na przeszkodzie by zaimplementować tą metodę estymacji. Poniżej przedstawiono listing prostego programu realizującego estymację gęstości rozkładu

prawdopodobieństwa n -wymiarowej zmiennej losowej (której próba zawarta jest w aktywnym arkuszu *Statistica*). Program wymaga podania współczynnika wygładzania - dodanie funkcjonalności doboru tego współczynnika na podstawie metod opisanych w rozdziale 1.5 nie stanowiłoby jednak większej trudności. Zastosowano w nim estymator o jądrze normalnym radialnym.

```
Sub Main
    Dim s As Spreadsheet
    Dim res_s As New Spreadsheet
    Dim tempstring As String
    Dim
h,Minx(),Maxx(),Stepx(),Tempx(),pierw,data_length,kernel_value,eksp As
Double
    Dim dimension,results_no,i,j,p,k,l,m As Integer
    ' first we should "load data" from active spreadsheet
    Set s = ActiveSpreadsheet
    If s Is Nothing Then GoTo Leave
    ' check dimension
    dimension = s.Variables.Count
    data_length = s.Cases.Count
    ReDim Minx(dimension)
    ReDim Maxx(dimension)
    ReDim Stepx(dimension)
    ReDim Tempx(dimension)
    ' get estimation ranges:
    results_no=1
    For i=1 To dimension
        ' minimum
        tempstring = InputBox("Enter x"+CStr(i)+"_min:")
        If tempstring = vbNullString Then
            MsgBox "No value entered"
            GoTo Leave
        End If
        Minx(i)=Cdbl(tempstring)

        ' maximum
        tempstring = InputBox("Enter x"+CStr(i)+"_max:")
        If tempstring = vbNullString Then
            MsgBox "No value entered"
            GoTo Leave
        End If
        Maxx(i)=Cdbl(tempstring)

        ' step
        tempstring = InputBox("Enter x"+CStr(i)+"_step:")
        If tempstring = vbNullString Then
            MsgBox "No value entered"
            GoTo Leave
        End If
        Stepx(i)=Cdbl(tempstring)

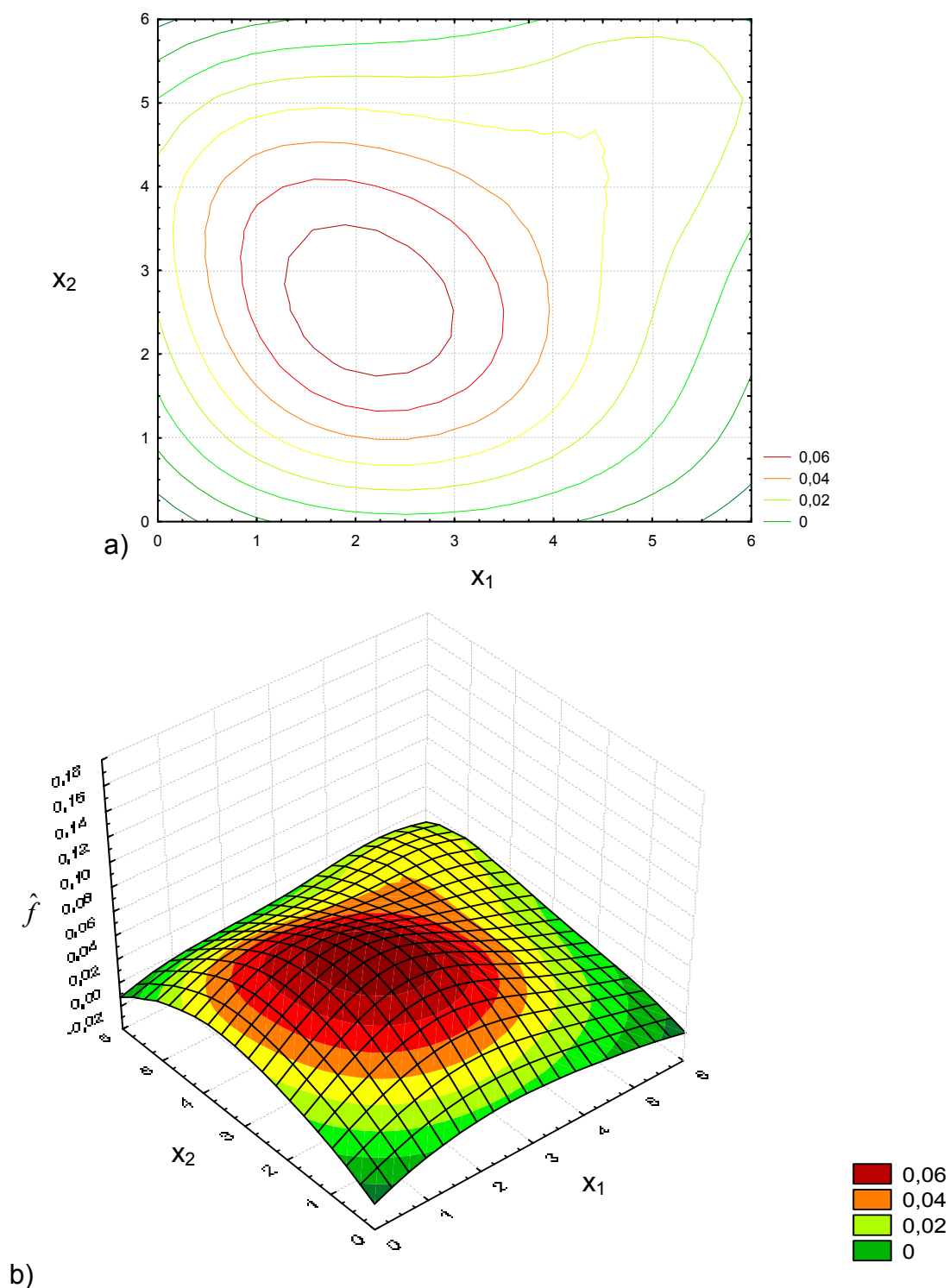
        ' count overall number of points
        results_no=results_no*((Maxx(i)-Minx(i))/Stepx(i)+1)
        Tempx(i)=Minx(i)
    Next i
    ' get bandwidth
    tempstring = InputBox("Enter bandwidth:")
    If tempstring = vbNullString Then
```

```

        MsgBox "No value entered"
        GoTo Leave
    End If
    h=Cdbl(tempstring)
    ' calculate kernel (result saved into a new spreadsheet)
    ' first prepare clean spreadsheet
    ' delete old variables
    ' add new ones
    res_s.AddVariables("Results",0,dimension)
    res_s.AddCases(0,results_no)
    res_s.SetSize(results_no,dimension+1)
    ' then perform calculations
    For i=1 To results_no
        ' loop through data set
        kernel_value=0
        For k=1 To data_length
            pierw=0
            ' loop through dimensions
            For l=1 To dimension
                pierw=pierw+(Tempx(l)-s.Value(k,l))^2
            Next l
            ' underflow protection - this part is optional
            ' (when working without this protection on complex data
            ' "invalid floating point operation" may occur)
            eksp=Exp(-pierw/2/h^2)
            If Abs(eksp)<1E-308 Then
                kernel_value=0
            Else
                kernel_value=kernel_value+eksp
            End If
        Next k
        ' normalize & save kernel value
        res_s.Value(i,dimension+1)=((kernel_value*(2*PI)^(-
dimension/2))/data_length)/(h^dimension)
        ' save point coordinates
        For m=1 To dimension
            res_s.Value(i,m)=Tempx(m)
        Next m
        ' if last one was calculated - leave
        If i=results_no Then
            GoTo SaveAndLeave
        End If
        ' generate next estimated point
        j=dimension
        While (Abs(Tempx(j) - Maxx(j))<1E-6)
            j=j-1
        Wend
        Tempx(j)=Tempx(j)+Stepx(j)
        If Tempx(j)>=Maxx(j) Then
            Tempx(j)=Maxx(j)
        End If
        For p=j+1 To dimension
            Tempx(p)=Minx(p)
        Next p
    Next i
    ' here save and leave
    SaveAndLeave:
    res_s.Activate
    res_s.Visible=True
    ' do not save - just leave (some error?)
    Leave:
End Sub

```

Uzyskane w programie *Statistica* przykładowe wizualizacje rozkładu zmiennej dwuwymiarowej o próbie 5 elementowej: $[1, 4; 3, 2; 2, 2; 5, 5; 2, 3]$ przedstawia rysunek B.2 (użyto współczynnika wygładzania $h = 0,5$).



Rysunek B.2. Estymacja jądrowa w programie Statistica

SPIS RYSUNKÓW I TABEL

Rysunek 1.1. Gęstość prawdopodobieństwa rozkładu normalnego	9
Rysunek 1.2. Wpływ doboru szerokości przedziałów h na własności histogramu: a) zbyt mała, b) zbyt duża, c) optymalna wartość	12
Rysunek 1.3. Interpretacja graficzna jednowymiarowego jądrowego estymatora gęstości prawdopodobieństwa $/m = 5/$	15
Rysunek 1.4. Wpływ parametru wygładzania h na estymator jądrowy: a) zbyt mała wartość, b) zbyt duża wartość, c) optymalna wartość parametru wygładzania.....	20
Rysunek 1.5. Interpretacja transformacji liniowej: a) brak transformacji, b) postać diagonalna, c) postać pełna.....	28
Rysunek 1.6. Koncepcja zmodyfikowanego parametru wygładzania.....	30
Rysunek 1.7. Lewostronne ograniczenie nośnika estymatora jądrowego	33
Rysunek 2.1. Okno główne programu KDEstim.....	36
Rysunek 2.2. Okno wizualizacji funkcji $g(h)$	37
Rysunek 2.3. Okno postępu.....	38
Rysunek 2.4. Okno wizualizacji wyników (dla danych jednowymiarowych)	38
Rysunek 2.5. Okno wizualizacji wyników (dla danych dwuwymiarowych)	39
Rysunek 2.6. Okno właściwości wykresu	39
Rysunek 2.7. Okno właściwości osi	40
Rysunek 2.8. Okno kolorów wykresu	40
Rysunek 3.1. Wynik estymacji dla wieku gości pensjonatu $/h = 4,19/$	43
Rysunek 3.2. Wynik estymacji dla czasu pobytu gości w pensjonacie $/h = 0,74/$	44
Rysunek 3.3. Wynik estymacji (a, b, c) dla zmiennej [wiek gości, czas pobytu] $/h = 0,32/45$	
Rysunek 3.4. Wynik estymacji dla wieku gości pensjonatu (z ograniczeniem lewostronnym dla 0 lat)	47
Rysunek 3.5. Wynik estymacji dla czasu pobytu w pensjonacie (z ograniczeniem	

lewostronnym dla 2 dni).....	47
Rysunek 3.6. Wynik estymacji dla zależności czasu pobytu w pensjonacie od wieku gości w przypadku ograniczenia nośnika / $h = 0,32$ /	48
Rysunek 3.7. Wykres funkcji $g(h)$ bez randomizacji danych, minimum - brak.....	50
Rysunek 3.8. Wykres funkcji $g(h)$ przy randomizacji $[-0,5; 0,5]$, minimum dla $h = 0,89$..	50
Rysunek 3.9. Porównanie wyników po randomizacji danych: a) dane oryginalne / $h=0,74$ /, b) dane oryginalne / $h=0,89$ /, c) dane zmodyfikowane / $h = 0,89$ /.....	51
Rysunek 3.10. Interfejs programu SPSS	53
Rysunek 3.11. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej <i>wiek</i> – kandydat Marian Krzaklewski / $h = 6,56$ /	54
Rysunek 3.12. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej <i>wiek</i> – kandydat Aleksander Kwaśniewski / $h = 3,13$ /	55
Rysunek 3.13. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej <i>wiek</i> – kandydat Andrzej Olechowski / $h = 4,74$ /	55
Rysunek 3.14. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej <i>wiek</i> – kandydat Lech Wałęsa / $h = 6,79$ /	56
Rysunek 3.15. Gęstość rozkładu prawdopodobieństwa dla zmiennej losowej <i>wiek</i> – odpowiedź „nie pamiętam” / $h = 5,64$ /	56
Rysunek 3.16. Wyniki estymacji dla sprawnego silnika (poślizg $s = 0,05$).....	60
Rysunek 3.17. Wyniki estymacji dla silnika nieznacznie uszkodzonego - rezystancja jednego z uzwojeń jest większa o 10 % (poślizg $s = 0,05$).	61
Rysunek 3.18. Wyniki estymacji dla średnio uszkodzonego silnika - rezystancja jednego z uzwojeń jest dwa razy większa (poślizg $s = 0,05$).....	62
Rysunek 3.19. Wyniki estymacji dla całkowicie uszkodzonego silnika - rezystancja jednego z uzwojeń jest 20 razy większa (poślizg $s = 0,05$).....	63
Rysunek 3.20. Wyniki estymacji dla sprawnego silnika (poślizg $s = 0,07$).....	64
Rysunek 3.21. Wyniki estymacji dla silnika nieznacznie uszkodzonego - rezystancja jednego z uzwojeń jest większa o 10 % (poślizg $s = 0,07$).	65
Rysunek 3.22. Wyniki estymacji dla silnika średnio uszkodzonego - rezystancja jednego z uzwojeń jest dwa razy większa (poślizg $s = 0,07$).....	66
Rysunek 3.23. Wyniki estymacji dla całkowicie uszkodzonego silnika - rezystancja jednego z uzwojeń jest 20 razy większa (poślizg $s = 0,07$).....	67
Rysunek 3.24. Rozkład energii elektronu mierzonej przez kalorymetr elektronowy. Dane dotyczące procesu fotoprodukcji / $h = 0,53$ /	70
Rysunek 3.25. Rozkład energii fotonu mierzony w kalorymetrze fotonowym. Dane	

dotyczące fotoprodukcji $/h = 0,17/$	71
Rysunek 3.26. Dwuwymiarowy rozkład energii elektronu względem energii fotonu w danych dotyczących fotoprodukcji $/h = 0,17/$	71
Rysunek 3.27. Rozkład położenia wierzchołka oddziaływania elektron-proton w zjawisku fotoprodukcji $/h = 3,29/$	72
Rysunek 3.28. Rozkład energii elektronu mierzonej przez kalorymetr elektronowy w symulacji zjawiska bremsstrahlungu $/h = 0,17/$	74
Rysunek 3.29. Rozkład energii fotonu mierzonej przez kalorymetr fotonowy w symulacji zjawiska bremsstrahlungu $/h = 0,01/$	74
Rysunek 3.30. Rozkład energii elektronu względem fotonu mierzonej w symulacji zjawiska bremsstrahlungu $/h = 0,20/$	75
Rysunek 3.31. Symulowany rozkład położenia wierzchołka oddziaływania elektron-proton $/h = 2,52/$	75
Rysunek B.1. Porównanie kilku typów jąder (program <i>R</i>)	81
Rysunek B.2. Estymacja jądrowa w programie Statistica	84
Tabela 1.1. Najczęściej stosowane typy jąder	16
Tabela 3.1 Liczność próby losowej dla poszczególnych kandydatów	54
Tabela 3.2 Wyniki sondażu wyborczego (na zlecenie portalu <i>interia.pl</i>)	58

BIBLIOGRAFIA

- [1] Andruszków J. et al.: „Luminosity measurement in the ZEUS experiment”, Acta Phys. Polonica vol. B32, pp. 2025-2057, 2001;
- [2] Chwastowski J., Figiel J.: „Photoproduction at HERA”, Phys. Part. Nucl. vol. 35, pp. 619-632, 2004;
- [3] Chwastowski J. (praca habilitacyjna): „Energy Evolution of the Total Cross Sections – Significance of the HERA γp Measurements”, Report no 1940/PH, Instytut Fizyki Jądrowej im. Henryka Niewodniczańskiego w Krakowie, 2004;
- [4] Cichomski B. (kierownik programu), Jerzyński T., Zieliński M.: „Polskie Generalne Sondaze Społeczne: skumulowany komputerowy zbiór danych 1992-2002”, Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa 2003;
- [5] Cichomski B. (kierownik programu), Jerzyński T., Zieliński M.: „Polskie Generalne Sondaze Społeczne: struktura skumulowanych wyników badań 1992-2002”, Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa 2003;
- [6] Code Project, <http://www.codeproject.com>;
- [7] Gajek L., Kałuszka M.: „Wnioskowanie statystyczne. Modele i metody”, WNT, Warszawa 1999;
- [8] Kulczycki P.: „Wykrywanie uszkodzeń w systemach zautomatyzowanych metodami statystycznymi z elementami losowego sterowania czasooptymalnego”, Wydawnictwo ALFA, Warszawa 1998;
- [9] Kulczycki P.: „Estymatory jądrowe w analizie systemowej”, WNT, Warszawa 2005;
- [10] Pavkov T., Pierce K.: „Do biegu, gotowi – start! Wprowadzenie do SPSS dla Windows”, Gdańskie Wydawnictwo Psychologiczne, Gdańsk 2005;

- [11] Paleczek W.: „Metody analizy danych”, Wydawnictwa Politechniki Częstochowskiej, Częstochowa 2004;
- [12] Polska Norma, PN-ISO 3534-1: „Statystyka. Terminologia i symbole. Część 1: Ogólne terminy z zakresu rachunku prawdopodobieństwa i statystyki”, Polski Komitet Normalizacyjny, Warszawa 2002;
- [13] Serwis Wybory 2000, portal internetowy *interia.pl*,
<http://wybory2000.interia.pl/id/info/www/artukul&numer=42496&kat=>;
- [14] Sobczyk T. J., Weinreb K., Węgiel T., Sułowicz M.: „Theoretical Study of Effects Due to Rotor Eccentricities in Induction Motors”, IEEE International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED'99), pp.289-295, Gijon, 1999;
- [15] Statistica – Przewodnik, Statsoft, Kraków, 2005;
- [16] Wikipedia – Wolna Encyklopedia, [http://pl.wikipedia.org/wiki/Exit_poll](http://pl.wikipedia.org/wiki/Exit_poll;).;
- [17] Wikipedia – Wolna Encyklopedia, [http://pl.wikipedia.org/wiki/R_\(program\)](http://pl.wikipedia.org/wiki/R_(program)).