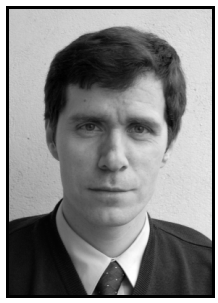
Jarosław GRAMACKI, Artur GRAMACKI
UNIWERSYTET ZIELONOGÓRSKI, INSTYTUT INFORMATYKI I ELEKTRONIKI

Estymacja nieparametryczna wybranych parametrów bloku gazowo-parowego

Dr inż. Jarosław GRAMACKI

Pracuje w Instytucie Informatyki i Elektroniki Uniwersytetu Zielonogórskiego na stanowisku adiunkta. Jego zainteresowania koncentrują się wokół zagadnień związanych z bazami danych, eksploracją danych oraz ich praktycznymi zastosowaniami.



e-mail: j.gramacki@iie.uz.zgora.pl

Dr inż. Artur GRAMACKI

Pracuje w Instytucie Informatyki i Elektroniki Uniwersytetu Zielonogórskiego na stanowisku adiunkta. Jego zainteresowania koncentrują się wokół zagadnień związanych z bazami danych, eksploracją danych oraz ich praktycznymi zastosowaniami.

e-mail: a.gramacki@iie.uz.zgora.pl

Streszczenie

W pracy pokazano przykład użycia nieparametrycznej estymacji danych. Z pomocą tej techniki dokonano oszacowania emisji tlenków azotu (NOx) na podstawie danych eksploatacyjnych zbieranych podczas normalnej pracy Elektrociepłowni w Zielonej Górze. Na wstępnie dokonano krótkiego przeglądu najbardziej popularnych technik estymacji parametrycznej i porównano je z technikami nieparametrycznymi. Następnie na prostym przykładzie pokazano istotę działania estymacji nieparametrycznej. Pracę kończy rozdział, w którym krótko omówiono uzyskane wyniki symulacyjne.

Słowa kluczowe: estymacja nieparametryczna, estymacja jądrowa, Elektrociepłownia Zielona Góra.

Nonparametric estimation of selected parameters of steam and gas power plant

Abstract

In the paper there are shown some practical examples of using nonparametric estimation. Using this technique there were estimated the nitrogen oxides (NOx) emissions based on the data taken from a real industry plant (gas and steam combined heat and power (CHP) plant in Zielona Góra, Poland). This work can be treated as a continuation of the paper [2]. In the first section there is given a short overview of estimation methods, including the linear and nonlinear regression, and comparison of them with nonparametric ones. In the second section there is briefly presented the nonparametric estimation technique and there is given a simple illustrative example. The third paragraph is dedicated to presenting the experimental results. Basing on the data from the CHP plant, the NOx emission was estimated and the satisfactory results (in comparison, for example, with the results obtained from the linear regression estimator) were obtained. All calculations were carried out using *np* package for R-project environment which implements a variety of nonparametric (and also semiparametric) kernel-based estimators.

Keywords: nonparametric estimation, kernel estimation, combined heat and power plant, CHP.

1. Wstęp

Analiza regresji jest bardzo popularną i chętnie stosowaną techniką statystyczną pozwalająca opisywać związki zachodzące pomiędzy zmiennymi wejściowymi (objaśniającymi) a wyjściowymi (objaśnianymi). Innymi słowy dokonujemy *estymacji* jednych danych, korzystając z innych (dokładniej: estymujemy parametry modelu). Technika ta rozwija się bardzo intensywnie od początku wieku XX i obecnie jest już bardzo wnikliwie opracowana. Pozycje [3, 5] należy traktować jako subiektywny wybór autorów spośród bardzo bogatej literatury przedmiotu.

Istnieje wiele różnych technik regresji. Niewątpliwie najpopularniejszą jest regresja liniowa. Jak wskazuje nazwa zakłada ona, że pomiędzy zmiennymi objaśniającymi i objaśnianymi istnieje mniej lub bardziej wyrazista zależność liniowa. Ponadto muszą być dodatkowo spełnione pewne inne założenia, których szczegółów jednak tu nie podajemy, odsyłając czytelnika do dostępnej literatury. Pewnymi wariantami podstawowej regresji liniowej są: regresja grzbietowa (ang. *ridge regression*), regresja odporna (ang. *robust regression*), regresja logistyczna i inne.

Jeżeli w analizowanych danych spodziewamy się nieliniowych zależności między zmiennymi objaśniającymi i objaśnianymi to możemy postąpić na dwa sposoby. Możemy mianowicie próbować dokonać transformacji zmiennych, tak aby w pewnym stopniu „uliniowić" model, lub też rozważyć zastosowanie regresji nieliniowej. Można spróbować też zastosować pewne bardziej zaawansowane i specjalistyczne modele regresji, jak np. regresja metodą składowych głównych (ang. *principal component regression, PCR*), regresja metodą częściowych najmniejszych kwadratów (ang. *partial least squares regression PLS*) uogólniona regresja nieliniowa (ang. *generalized non linear least squares, GNLS*), odporna regresja lokalnie ważona Lowess (ang. *locally weighted scatterplot smoothing*) i inne.

Cechą wspólną wszystkich wyżej wymienionych rodzajów regresji jest to, że musimy znać a priori (lub założyć na wstępie analizy) jakąś matematyczną zależność wiążącą zmienne objaśniające i objaśniane. W związku z tym, wszystkie wymienione wyżej rodzaje regresji możemy nazwać regresjami *parametrycznymi*.

Zadanie znalezienia wspomnianej postaci matematycznej nie zawsze jest łatwe do wykonania, a czasami wręcz niewykonalne (gdy analizujemy dane o bardzo złożonych zależnościach). W takich sytuacjach warto rozważyć użycie jednej z metod *nieparametrycznych*, która nie wymaga w żadnym miejscu przyjmowania założeń co do postaci funkcji wiążącej dane wejściowe i wyjściowe. Argumentem za użyciem metod nieparametrycznych jest ich prostota. Często jest bowiem tak, że nawet jeśli użycie metod parametrycznych jest uzasadnione, metody nieparametryczne po prostu łatwiej jest zastosować. Ponadto stosowanie metod parametrycznych jest obwarowane koniecznością brania pod uwagę wielu założeń, szczególnie względem *rozkładu populacji*. Gdy te nie są spełnione, albo są spełnione „słabo", uzyskane wyniki mogą być niepewne. Z drugiej strony, gdy założenia co do rozkładu populacji są dobrze spełnione metody parametryczne dają lepsze wyniki niż metody nieparametryczne – generują mniejszy błąd i bardziej istotne statystycznie wyniki, a testy mają większą moc.

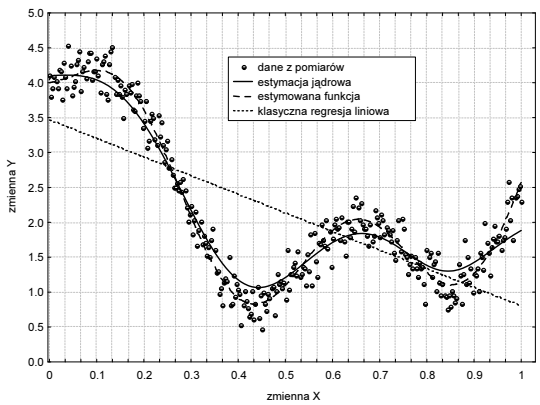
W pracy zademonstrowano celowość użycia takiego właśnie nieparametrycznego podejścia do celów estymacji pewnych wybranych parametrów bloku gazow-parowego w Elektrociepłowni w Zielonej Górze [10]. Niniejsza praca jest w pewnym sensie kontynuacją i uzupełnieniem pracy [2][1] i dlatego też pewnych zawartych tam szczegółowych informacji na temat badanego obiektu obecnie nie zamieszczamy, odsyłając zainteresowanego czytelnika do wymienionego źródła.

2. Wprowadzenie do estymacji nieparametrycznej

Aby uzmysłowić sobie problemy, jakie mogą pojawić się w trakcie estymacji danych, rozważmy prosty przykład pokazany

[1] Praca ta jest obecnie w trakcie publikacji. Osoby zainteresowane prosimy o kontakt osobisty z autorami celem uzyskania szczegółowych danych bibliograficznych.

na rysunku 1. Mamy tu pewne dane pomiarowe (zmienna *x*, objaśniająca oraz zmienna *y*, objaśniana). Widać, że zależność jest na pewno daleka od liniowej. Zastosowanie klasycznej regresji liniowej jest absolutnie niedopuszczalne. Z drugiej strony zastosowanie tzw. nieparametrycznej *estymacji jądrowej* daje nam bardzo ładne dopasowanie krzywej do estymowanej funkcji[2].



Rys. 1. Istota estymacji nieparametrycznej
Fig. 1. The idea of nonparametric estimation

Aby dobrze zrozumieć istotę działania estymacji nieparametrycznej wykonajmy następujące rozumowanie [6][3]. Niech będą dane dwie jednowymiarowe zmienne losowe *X* oraz *Y* oraz uzyskana z nich próba losowa $(x_1,y_1),\ldots,(x_n,y_n)$, gdzie *n* oznacza ilość obserwacji. Poszukujemy funkcji *f*, która najlepiej charakteryzowałaby ew. zależność między zmiennymi losowymi *X* oraz *Y*. Jej estymator $\hat{m}(x)$ wyznacza się tak, aby zminimalizować wyrażenie

$$\sum_{i=1}^{n}\left[y_i - m(x_i)\right]^2. \tag{1}$$

Zmienne losowe są powiązane ze sobą relacją $y_i=m(x_i)+e_i$, gdzie e_i to biały szum o pewnych określonych właściwościach statystycznych. Funkcja $m(x)$ opisuje warunkową wartość oczekiwaną zmiennej losowej *Y* pod warunkiem, że zmienna *X* przyjęła zadaną wartość *x*, czyli $m(x)=E(Y|X=x)$. W odróżnieniu od metod estymacji parametrycznej, estymator $\hat{m}(x)$ nie ma postaci funkcji analitycznej a jest po prostu pewną gładką, ale o nieznanej postaci, funkcją.

Załóżmy, iż chcemy wyznaczyć wartość $\hat{m}(x)$ biorąc pod uwagę punkty z jego otoczenia $D_x=[x-b,\ x+b]$, $b>0$. Wówczas $\hat{m}(x)$ można estymować następującym wyrażeniem:

$$\hat{m}(x) = \frac{\sum_{x_i \in D_x} y_i}{\#\{x_i \in D_x\}}, \tag{2}$$

gdzie $\#\{x_i \in D_x\}$ jest liczbą elementów wewnątrz D_x. Zauważmy, że $m(x_t)$ można zapisać w postaci

$$\hat{m}(x_t) = \frac{\sum_{i=1}^{n} w_{ix} y_i}{\sum_{i=1}^{n} w_{ix}}, \tag{3}$$

gdzie w_{ix}= 0 lub 1 zwane jest wagą. Równanie (3) jest niczym innym jak wzorem na średnią ważoną. Dane, którym przypisano większe wagi mają większy udział w określeniu średniej ważonej niż dane, którym przypisano mniejsze wagi.

Powyższa postać estymatora ma znaczenie tylko teoretyczne, gdyż uzyskiwane estymaty nie są gładkie ani dokładne. Niemniej jednak (3) dobrze obrazuje istotę zagadnienia. W praktyce wagi w_{ix} zastępowane są przez funkcją o postaci

$$w_{ix} = \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \equiv K_h(x - x_i), \tag{4}$$

gdzie $K(x)$ jest określane mianem jądra estymatora i ma najczęściej postać funkcji gaussowskiej (normalnej)

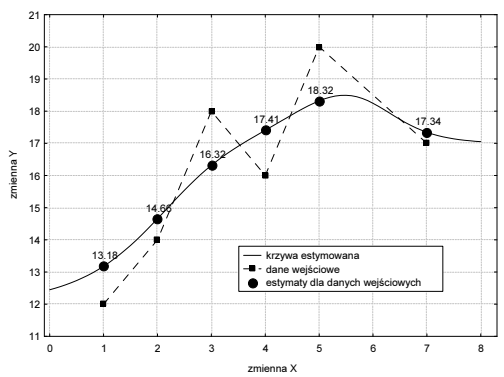
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \tag{5}$$

natomiast *h* jest pewnym ustalonym, dodatnim współczynnikiem wygładzania (ang. *bandwidth*). Otrzymujemy więc podstawową postać jądrowego estymatora regresji zwanego estymatorem Nadaraya-Watsona [7, 9].

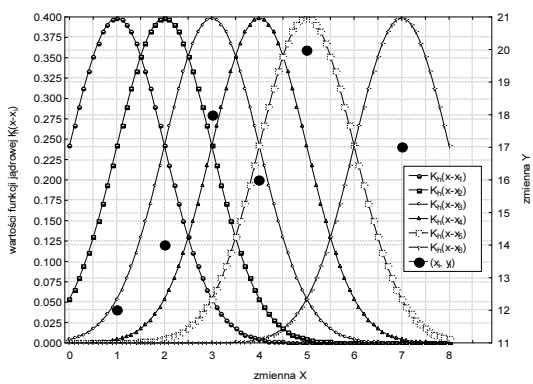
$$\hat{m}(x_t) = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} \equiv \sum_{i=1}^{n} w_i y_i\ . \tag{6}$$

Bardzo istotne znaczenie dla kształtu linii regresji ma właściwy dobór współczynnika wygładzania *h*. Opracowano wiele algorytmów jego automatycznego doboru. Mniej istotna jest natomiast postać funkcji jądrowej. W literaturze zaproponowano kilka innych jej „zastępników" [6]. Najpopularniejsze to jądra: Epanecznikowa, jednostajne, dwuwagowe, trójkątne.

Na rysunku 2 pokazano wyniki estymacji bardzo prostego przykłady (zbioru danych) składającego się zaledwie z 6 punktów *(1,12), (2,14), (3,18), (4,16), (5,20), (7,17)*. Pokazano na nim punkty wejściowe, wartości estymat w tych punktach wyliczone według wzoru (6) oraz całą krzywą regresji (współczynnik wygładzania *h=1*). Widać, że nawet dla tak prostego przykładu estymacja jest całkiem dobra. Zaznaczono również wartości liczbowe $m(x_i)$ wyliczone wprost z równania (6).



Rys. 2. Jądrowy estymator funkcji regresji: wyniki estymacji
Fig. 2. Kernel estimator of the regression function: estimation results



Rys. 3. Jądrowy estymator funkcji regresji: przebiegi funkcji jądrowej
Fig. 3. Kernel estimator of the regression function: kernel function plots

[2] Powyższy przykład jest trochę sztuczny, gdyż powstał w ten sposób, że najpierw założyliśmy sobie jakąś bardzo nieliniową postać funkcji (linia przerywana), następnie losowo zaburzyliśmy ją (kółeczka na wykresie) i następnie próbowaliśmy estymatorem jądrowym dopasować się możliwie jak najlepiej do tej funkcji.

[3] Ograniczamy się do niezbyt formalnego pokazania istoty estymacji nieparametrycznej. Stricte matematyczne uzasadnienie jest dość złożone i osoby zainteresowane mogą sięgnąć do dostępnej literatury, np. [4-6]

Na rysunku 3 (ma on dwie osie Y) pokazano dokładne przebiegi funkcji jądrowej (5) w otoczeniu 6. punktów wejściowych wraz z tymi punktami.

3. Wyniki eksperymentów

Omówioną w poprzednim rozdziale technikę regresji nieparametrycznej zastosowano do analizy danych rzeczywistego obiektu przemysłowego. Jest nim Elektrociepłownia w Zielonej Górze [10]. Podjęto próbę oszacowania (estymacji) wielkości emisji tlenków azotu (NOx,) na podstawie danych eksploatacyjnych zbieranych podczas normalnej pracy Elektrociepłowni w Zielonej Górze. Dokładniejsze informacje na temat obiektu, sposobu wyboru zmiennych objaśniających i objaśnianych oraz wstępnego przygotowania danych zamieszczono w pracy [2]. W tym miejscu nie będziemy więc ich powtarzać. Wspomnijmy tylko, że wybrano 3 zmienne objaśniające: (a) temperatura spalin przed tzw. przegrzewaczem wysokiego ciśnienia w kotle odzysknicowym KO, (b) temperatura sprężonego powietrza podawanego do komory spalania turbiny gazowej oraz (c) wilgotność powietrza. Zmienną objaśnianą jest wielkość emisji tlenków azotu NOx.

Eksperymenty wykonano stosując, oprócz estymatora Nadaraya-Watsona (6), również nieco bardziej złożony estymator jądrowy, który nosi nazwę liniowej lokalnej regresji jądrowej (ang. *local linear kernel regression*) i daje, przynajmniej teoretycznie, nieco lepsze wyniki. Obliczenia wykonywano używając systemu R (dedykowanego głównie do obliczeń statystycznych) i pakietu o nazwie *np* [12]. W pakiecie tym zaimplementowano m.in. dwie różne wersje estymatorów jądrowych: wspomniany już estymator Nadaraya-Watsona oraz liniowej lokalnej regresji jądrowej oraz dwie metody automatycznego doboru współczynnika wygładzania *h*: metodę krzyżowego uwiarygodniania najmniejszych kwadratów (ang. *least-squares cross-validation*) oraz metodę krzyżowego uwiarygodniania Kullbacka-Leiblera [12]. W materiałach referencyjnych do tego pakietu można znaleźć odnośniki do pozycji literaturowych dotyczących algorytmów wyboru wartości parametru *h* (ang. *bandwidth selection*).

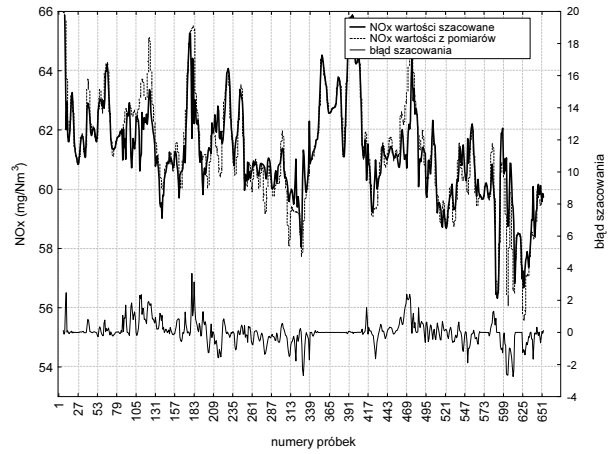
W tabeli 1 zebrano najważniejsze wyniki eksperymentów. Pogrubiono najlepsze rezultaty. Porównując otrzymane wyniki z tymi, które otrzymano dla prostej regresji liniowej [2] można stwierdzić, że regresja nieparametryczna pozwala uzyskiwać dużo lepsze wyniki. W przypadku regresji liniowej najlepszy wynik dla współczynnika determinacji R^2 jaki udało się uzyskać nieznacznie tylko przekraczał wartość 0.4. Natomiast błąd *MSE* był kilka razy wiekszy. Zwróćmy również uwagę, że teoretycznie gorszy estymator Nadaraya-Watsona w tym konkretnym przypadku okazał się lepszy niż liniowy estymator lokalnej regresji jądrowej. Potwierdza to znany fakt, że ostatecznego wyboru konkretnej metody nieparametrycznej należy dokonać analizując uzyskane wyniki dla konkretnego zadania. Podobnie sytuacja ma się metodą doboru parametru wygładzania *h*. W naszym eksperymencie lepsza okazała się metoda krzyżowego uwiarygodniania najmniejszych kwadratów, która jest mniej złożona niż metoda Kullbacka-Leiblera a mimo to daje lepsze wyniki.

Na rysunki 4 pokazano przebieg oryginalnej zmiennej NOx oraz jej estymację nieparametryczną, jak również wartości błędów szacowania.

Tab. 1. Wyniki estymacji nieparametrycznej dla różnych postaci estymatorów oraz różnych metod wyboru parametru *h*
Tab. 1. Results of nonparametric estimation for different types of kernel regression estimators and different methods of selecting bandwidths

typ estymatora jądrowego	metoda wyboru parametru *h*	współczynnik determinacji R^2	błąd *MSE* (ang. *mean square error*)
NW	KL	0.732	0.969
NW	LS	**0.856**	**0.532**
LLRJ	KL	0.632	1.322
LLRJ	LS	0.843	0.577

NW – estymator Nadaraya-Watsona
LLRJ – liniowa lokalna regresja jądrowa
KL – Kullback-Leibler cross-validation
LS – least-squares cross-validation



Rys. 4. Przebieg oryginalnej zmiennej *NOx* oraz jej estymacja
Fig. 4. Original values of *NOx* and its estimates

4. Wnioski

W artykule zademonstrowano użycie nieparametrycznej metody estymacji danych z rzeczywistego obiektu przemysłowego (Elektrociepłownia w Zielonej Górze. Uzyskano bardzo dobre wyniki, znacznie przewyższające te, które są możliwe do uzyskania przy zastosowaniu bardziej klasycznej metody estymacji, jaką jest regresja liniowa.

Celowym byłoby wykonanie podobnych eksperymentów na innych fragmentach badanego obiektu i porównanie wyników. Warto by również przeprowadzić dokładniejsze konsultacje z pracownikami ECZG, odnośnie wyboru do symulacji takiego fragmentu, który byłby ważny z praktycznego punktu widzenia. Wstępnie wydaje się, że analiza obiegu wody chłodzącej (skraplacz, chłodnia kominowa) byłaby bardzo pożądana.

Celem przybliżenia czytelnikowi istoty estymacji nieparametrycznej zamieszczony również bardzo prosty przykład numeryczny. Na danych składających się tylko z 6 próbek zademonstrowano stosowne wyniki obliczeń.

5. Literatura

[1] Bowman A.W., Azzalini A.: Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations, New York: Oxford University Press, 1997.
[2] Gramacki A., Gramacki J.: Szacowanie emisji tlenków azotu (NOx) na podstawie danych eksploatacyjnych rzeczywistego obiektu przemysłowego (w trakcie publikacji).
[3] Klonecki W.: Elementy statystyki dla inżynierów, Oficyna wydawnicza Politechniki Wrocławskiej, 1996.
[4] Koronacki J., Ćwik J.: Statystyczne systemy uczące się, Akademicka oficyna wydawnicza EXIT, 2008.
[5] Koronacki J., Mielniczuk J.: Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT, Warszawa, 2006.
[6] Kulczyki P.: Estymatory jądrowe w analizie systemowej, WNT, Warszawa, 2005.
[7] Nadaraya, E. A.: On Estimating Regression, Theory Probab. Appl., 10, 186–190, 1964.
[8] Simonoff J. S.: Smoothing Methods in Statistics (Springer Series in Statistics), Springer,1996.
[9] Watson, G. S.: Smooth Regression Analysis, Sankhaya Ser. A, 21, 101–116, 1964.
[10] http://www.ec.zgora.pl
[11] http://www.stat.nus.edu.sg/~staxyc/
[12] http://CRAN.R-project.org/package=np

Artykuł recenzowany