Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 227 · 2015

Joanna Trzęsiok

Uniwersytet Ekonomiczny w Katowicach
Wydział Zarządzania
Katedra Analiz Gospodarczych i Finansowych
joanna.trzesiok@ue.katowice.pl

O ODPORNOŚCI NA OBSERWACJE ODSTAJĄCE WYBRANYCH NIEPARAMETRYCZNYCH MODELI REGRESJI

Streszczenie: Artykuł jest poświęcony zagadnieniu odporności metod regresji na obserwacje odstające występujące w zbiorze danych. W pierwszej części przedstawiono wybrane metody identyfikacji obserwacji nietypowych. Następnie badano odporność trzech nieparametrycznych metod regresji: PPR, POLYMARS i RANDOM FORESTS. Analiz dokonano za pomocą procedur symulacyjnych na rzeczywistym zbiorze danych Mieszkania, w którym wykryto obserwacje odstające. Pomimo dosyć powszechnych przekonań o odporności regresji nieparametrycznej, okazało się, że modele zbudowane na całym zbiorze danych mają istotnie mniejsze zdolności predykcyjne niż modele uzyskane na zbiorze, z którego usunięto obserwacje nietypowe.

Słowa kluczowe: obserwacje odstające, odporność, regresja nieparametryczna.

Wprowadzenie

Problem odporności jest bardzo ważnym zagadnieniem w modelowaniu zjawisk ekonomicznych. Budowanie modelu statystycznego na zbiorze danych, w którym wartości cech są zakłócone np. błędami pomiaru, brakiem losowości próby czy występowaniem wartości odstających, może doprowadzić do niekorzystnych konsekwencji. Wykorzystanie metod nieodpornych na zaburzenia danych może skutkować zbudowaniem modelu, który nie będzie odzwierciedlał głównych mechanizmów regulujących zachowanie badanego zjawiska. Jest wtedy wysoce prawdopodobne, że wnioskowanie, predykcja i podejmowanie decyzji na podstawie takiego modelu będzie obarczone dużymi błędami.

Problem odporności jest również złożonym zagadnieniem. W najbardziej ogólnym rozumieniu, zastosowanie odpornej metody regresji oznacza, że mamy do czynienia z modelem, który wskazuje tendencję reprezentowaną przez większość obserwacji. Odporność można jednak rozpatrywać w kilku aspektach, jako np. niewrażliwość na występowanie w zbiorze uczącym wartości odstających, zmiennych nieistotnych, braków wartości niektórych zmiennych czy losowych zakłóceń wartości cech.

W tym artykule przedmiotem badania jest odporność wybranych nieparametrycznych metod regresji: PPR, POLYMARS i RANDOM FORESTS na występowanie w zbiorze danych wartości odstających. Celem zaś pracy – zbadanie czy metody te prowadzą do uzyskania modeli odpornych, czyli takich, dla których wartości miar dokładności predykcji nie zmienią się istotnie po usunięciu obserwacji nietypowych.

1. Odporność na obserwacje odstające

Jednym z podstawowych założeń występujących również w wielowymiarowych metodach regresji jest założenie o jednorodności zbioru danych [Jajuga, 1993, s. 78]. Jego przyjęcie oznacza, że dane wykorzystane do analizy są traktowane jako zbiór obserwacji pochodzących z tej samej populacji, pomijając fakt występowania wartości oddalonych, nieprzystających do reszty. Obserwacje takie często pojawiają się w rzeczywistych zbiorach danych i wymagają szczególnej uwagi, ponieważ mogą mieć istotny wpływ na wyniki analizy.

Mówiąc o odporności regresji, na ogół mamy na myśli niewrażliwość modelu na jakość danych, czyli przede wszystkim na obecność w zbiorze uczącym obserwacji odstających (nietypowych). W kontekście założenia o jednorodności zbioru danych, można jednak rozpatrywać ten problem w najbardziej ogólnym przypadku, jako odporność metody regresji na niespełnienie części założeń wymaganych dla prawidłowego działania danej metody [Jajuga, 1993, s. 81].

1.1. Identyfikacja obserwacji odstających

Ze względu na przyczyny powstawania, obserwacje odstające można podzielić na:

– obserwacje nietypowe wynikające z różnego rodzaju błędów: z błędnego pomiaru, błędów przy gromadzeniu i wprowadzaniu danych, zamierzonej nieuczciwości w sprawozdawczości, jak również źle dobranej metodologii badań, źle dobranej próby badawczej czy błędnych założeń;

– obserwacje nietypowe, pochodzące z tzw. ogona rozkładu;
– obserwacje wpływowe, które mają istotny wpływ na postać modelu i mogą prowadzić do uzyskania ciekawych hipotez.

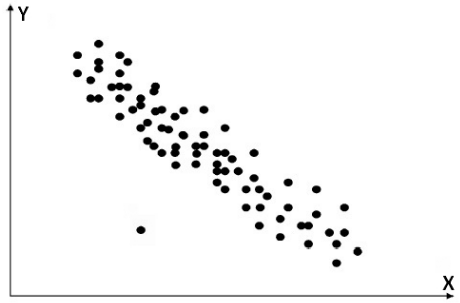
Identyfikacja obserwacji odstających oraz sposoby radzenia sobie z nimi są ważnymi zagadnieniami związanymi z pojęciem odporności w statystyce [Trzpiot, red., 2013]. W pracy tej zostaną przedstawione, jak również zastosowane w badaniu, trzy wybrane metody wykrywania obserwacji nietypowych.

- **Jednowymiarowe kryterium kwartylowe**, które jest wykorzystywane m.in. do budowy wykresów pudełkowych wprowadzonych przez Tukeya [1977]. Zgodnie z tym kryterium wartość pojedynczej zmiennej zostaje uznana za odstającą, jeśli znajduje się poza przedziałem:

$$\langle Q_1 - \frac{3(Q_3-Q_1)}{2}, Q_3 + \frac{3(Q_3-Q_1)}{2} \rangle, \tag{1}$$

gdzie Q_1 to pierwszy kwartyl, zaś Q_3 – trzeci kwartyl.

Metoda ta jest wykorzystywana do wstępnej analizy zbioru danych, jednak w przypadkach danych wielowymiarowych na ogół nie jest skuteczna w wykrywaniu obserwacji odstających. Przykład takiej sytuacji przedstawia rys. 1, na którym zaznaczono obserwację oddaloną, która ze względu na zmienne X, jak i Y z osobna, nie odbiega znacząco od mediany.



Rys. 1. Przykład zbioru z obserwacją oddaloną, której nie można zidentyfikować za pomocą kryterium kwartylowego

Źródło: Na podstawie [Jajuga, 1993].

- **Kryterium opierające się na odległości Cooka** [1977] jest popularną metodą stosowaną do wykrywania obserwacji odstających w analizie regresji wielorakiej. W metodzie tej porównuje się stopień dopasowania do danych dwóch modeli:
  – pełnego, który uwzględnia wszystkie obserwacje ze zbioru uczącego,
  – zbudowanego na zbiorze, z którego usunięto jedną, wybraną obserwację o numerze i.

Odległość Cooka można zapisać wzorem:

$$D_i = \frac{\sum_{j=1}^{m}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{m \cdot MSE}, \tag{2}$$

gdzie $\hat{Y}_j$ to prognozowana wielkość zmiennej Y dla obserwacji o numerze j w modelu pełnym,
$\hat{Y}_{j(i)}$ – prognozowana wartość zmiennej Y dla obserwacji o numerze j w modelu zbudowanym na zbiorze, z którego usunięto obserwację o numerze i,
m – liczba parametrów modelu,
zaś MSE – błąd średniokwadratowy modelu.

Obserwację i uznajemy za odstającą, zgodnie z powyższym kryterium, jeśli odpowiadająca jej odległość D_i jest większa od wartości granicznej:

$$\frac{4}{n-m-2}, \tag{3}$$

gdzie n jest liczbą obserwacji.

- **Kryterium opierające się na odległości Mahalanobisa** [Healy, 1968]:

$$MD(\mathbf{x}) = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}})^T}, \tag{4}$$

gdzie $\hat{\boldsymbol{\mu}}$ jest wartością przeciętną, zaś $\hat{\boldsymbol{\Sigma}}$ – macierzą wariancji i kowariancji:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x} - \hat{\boldsymbol{\mu}})^T(\mathbf{x} - \hat{\boldsymbol{\mu}}), \tag{5}$$

Na mocy tego kryterium, obserwacje traktujemy jako odstające, jeśli odpowiadają im duże wartości $MD(\mathbf{x})$ w porównaniu do wartości krytycznych odczytanych z tablic rozkładu χ^2.

Podstawową wadą tej metody jest to, że opiera się na statystykach klasycznych, które są bardzo wrażliwe na występowanie obserwacji odstających i tym samym wartości miary MD nie zawsze można uważać za wiarygodne. Z tego też powodu w literaturze zaproponowano wiele modyfikacji odległości Mahalanobisa. Jedną z nich jest podejście zaproponowane w 2008 r. przez Filzmosera, Maronnę i Wernera, wykorzystujące do identyfikacji obserwacji odstających analizę głównych składowych. Metoda ta jest szerzej opisana w pracy [Filzmoser, Maronna, Werner, 2008].

2. Metody regresji wykorzystane w badaniu

Problem odporności nabiera szczególnego znaczenia w przypadku nieparametrycznych modeli regresji, które charakteryzują się dużą elastycznością i zdolnością do adaptacyjnego, dokładnego dopasowania się do danych, uwzględniając

również zmienność wynikającą z zakłóceń. Pojawia się pytanie, jak zachowują się modele nieparametryczne budowane na zbiorach uczących zaburzonych wartościami odstającymi.

W świetle powyższej uwagi, metody nieparametryczne mogą generować modele nieodporne na występowanie w zbiorze uczącym wartości odstających, które mają niewielkie zdolności predykcyjne i tym samym małą wartość poznawczą dla badacza. Z drugiej jednak strony, wiele z tych metod ma wbudowany mechanizm regularyzacji, który pozwala ograniczyć problem nadmiernego dopasowania modelu do danych ze zbioru uczącego. Mechanizm ten polega na przyjęciu pewnego kompromisu pomiędzy dopasowaniem modelu a jego złożonością [Trzęsiok, 2011], co prowadzi do zwiększenia zdolności predykcyjnych modelu. Zachodzi jednak pytanie, w jakim stopniu mechanizm ten jest skuteczny, a omawiane metody są rzeczywiście odporne na wartości oddalone?

W badaniu zostały wykorzystane trzy wybrane metody nieparametryczne, często występujące w badaniach porównawczych i charakteryzujące się dobrymi własnościami predykcyjnymi [Meyer, Leisch, Hornik, 2003]:

- metoda rzutowania PPR [Friedman, Stuetzle 1981],
- wielowymiarowa metoda krzywych sklejanych POLYMARS [Kooperberg, Bose, Stone, 1997],
- metoda zagregowanych drzew Breimana – RANDOM FORESTS [Breiman, 2001].

3. Badanie odporności

Ze względu na odmienne mechanizmy działania nieparametrycznych metod regresji, niemożliwe jest analityczne porównanie generowanych przez nie modeli. Z tego względu badania porównawcze przeprowadzono za pomocą procedur symulacyjnych, na zbiorze danych rzeczywistych *Mieszkania*, który utworzono na podstawie informacji o zrealizowanych transakcjach sprzedaży mieszkań, udostępnianych przez serwis internetowy www.oferty.net.

Dane dotyczą transakcji sprzedaży mieszkań zrealizowanych od czerwca 2007 r. do września 2009 r. Zbiór *Mieszkania* zawiera 747 obserwacji opisywanych przez 8 zmiennych objaśniających (z których jedna jest mierzona na skali porządkowej, dwie to zmienne nominalne, a pozostałe są mierzone na skalach mocnych):

X_1 – powierzchnia mieszkania [w m^2],
X_2 – lokalizacja (nazwa dzielnicy),
X_3 – odległość mieszkania od centrum [w km],

X_4 – liczba pokoi,
X_5 – piętro,
X_6 – rok wybudowania (oddania do użytku) budynku, w którym znajduje się mieszkanie,
X_7 – typ własności (mieszkanie: spółdzielcze, własnościowe, hipoteczne, spółdzielczo-własnościowe),
X_8 – stan mieszkania (5 – bardzo dobry, 4 – dobry, 3 – do wykończenia, 2 – do remontu).

Zmienną zależną jest Y – cena transakcyjna mieszkania [w tys. zł].

W tab. 1 przedstawiono wybrane statystyki opisowe dla zmiennej zależnej (cena transakcyjna). Bardzo silne zróżnicowanie cen mieszkań (68%), jak i silna asymetria prawostronna (zestandaryzowany moment centralny trzeciego rzędu równy 3,2) nie stanowią problemu dla nieparametrycznych metod regresji, wykorzystanych do modelowania. Metody te nie wymagają transformacji zmiennych poprzez przekształcenia monotoniczne, co występuje np. w przypadku klasycznej metody liniowej.

Tab. 1. Charakterystyki opisowe zmiennej zależnej Y w zbiorze danych *Mieszkania*

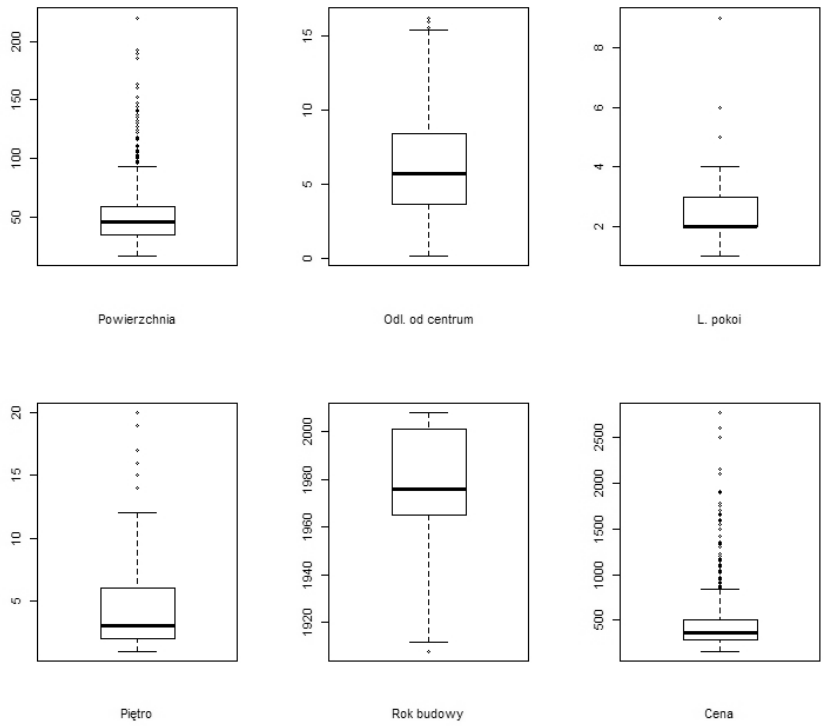
Średnia	Współczynnik zmienności	Współczynnik asymetrii
460 278 zł	68%	3,2
Minimum	Mediana	Maksimum
160 000 zł	366 000 zł	2 770 000 zł

3.1. Obserwacje odstające w zbiorze *Mieszkania*

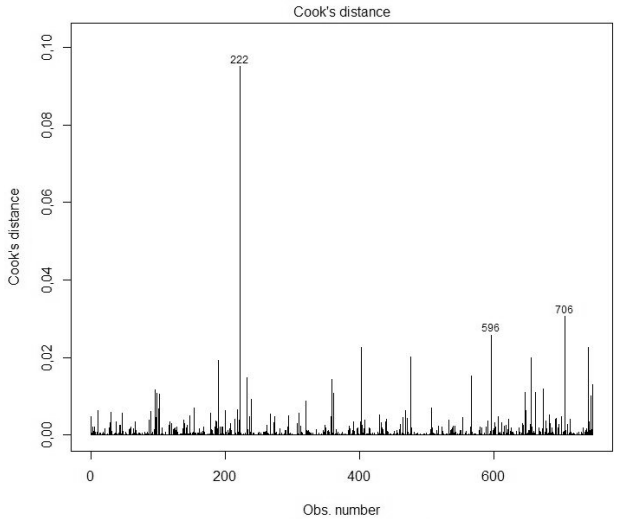
Do identyfikacji obserwacji odstających wykorzystano: jednowymiarową metodę kwartylową, kryterium opierające się na odległości Cooka oraz metodę wykorzystującą odległość Mahalanobisa z poprawką zaproponowaną przez Filzmosera i in. (ozn. MD^*).

Tak jak już wspomniano, wykresy pudełkowe stanowią jedynie wstępną analizę zbioru danych (zob. rys. 2).

Za pomocą odległości Cooka zidentyfikowano 35 obserwacji odstających, z których trzy największe zaznaczono na wykresie zaprezentowanym na rys. 3. W tym przypadku wartość graniczna przedstawiona we wzorze (3) jest równa w przybliżeniu 0,0056.



Rys. 2. Wykresy pudełkowe dla sześciu zmiennych z zaznaczonymi obserwacjami odstającymi, wyznaczonymi poprzez kryterium kwartylowe



Rys. 3. Wynik identyfikacji obserwacjami odstającymi za pomocą odległości Cooka

Kryterium oparte na odległości Mahalanobisa MD^* z poprawką Filzmosera wykrywa natomiast aż 68 obserwacji odstających. Kilka wybranych, przykładowych obserwacji nietypowych przedstawiono w tab. 2.

Tab. 2. Wybrane, przykładowe obserwacje odstające w zbiorze danych *Mieszkania*, otrzymane za pomocą kryterium Mahalanobisa

Nr obserwacji odstającej	Powierzchnia	Dzielnica	Odległość od centrum	Liczba pokoi	Piętro	Rok budowy	Typ własności	Stan mieszkania	Cena
98	122	Ursynów	10,8	5	3	2007	hipoteczne	bdb	1100
99	141	Śródmieście	3,6	4	8	2007	hipoteczne	do wykończenia	1198
222	220	Żoliborz	5,1	9	1	1928	hipoteczne	bdb	2770
403	193	Śródmieście	1,3	6	5	1921	hipoteczne	bdb	2600
435	107	Mokotów	6,7	4	3	2002	spółdzielcze	bdb	820
583	102	Białołęka	14,7	4	3	1998	hipoteczne	bdb	777,6
708	100	Śródmieście	1,7	3	4	1999	spółdzielcze	bdb	1020

Jak łatwo zauważyć, obserwacje odstające, przedstawione w tab. 2, mają znacznie wyższą cenę transakcyjną w porównaniu do średniej (por. tab. 1); są to również mieszkania o dużej powierzchni.

3.2. Analiza porównawcza

Identyfikacja obserwacji odstających była tylko pierwszym etapem badania. Oczywiście dla rzeczywistego zbioru danych niemożliwe jest jednoznaczne zidentyfikowanie takich obserwacji. Opisane metody nie dają nam żadnej gwarancji na to, że wyznaczono wszystkie obserwacje nietypowe, może bowiem zachodzić zjawisko maskowania się większej liczby obserwacji odstających leżących blisko siebie. Mimo to, przyjmując jako obserwacje odstające te, które zostały wyznaczone za pomocą odpornego kryterium Mahalanobisa MD^*, w kolejnym kroku sprawdzano, czy wybrane nieparametryczne modele regresji są odporne na występowanie tych obserwacji w zbiorze danych.

W tym celu zbudowano, za pomocą każdej z wymienionych metod, dwa typy modeli:

– na całym oryginalnym zbiorze danych,
– na zbiorze danych, z którego usunięto 68 obserwacji odstających.

Dla tak wyznaczonych modeli obliczono metodą sprawdzania krzyżowego błąd średniokwadratowy MSE_{CV} (z podziałem zbioru danych na 10 części). Otrzymane wyniki przedstawiono w tab. 3.

Tab. 3. Wartości błędu średniokwadratowego MSE_{CV} obliczone na całym zbiorze danych *Mieszkania* oraz na zbiorze *Mieszkania*, z którego usunięto obserwacje odstające

Metoda regresji	Wartości MSE_{CV}	
	cały zbiór danych	zbiór bez obserwacji odstających
PPR	11 320,5	3 566,2
POLYMARS	10 348,2	3 275,4
R. FORESTS	8 036,7	1 803,7

Wstępna analiza wyników przedstawionych w tab. 3 pokazuje, że modele zbudowane (dla każdej z metod) na całym zbiorze danych mają znacznie mniejsze zdolności predykcyjne (wyższe wartości MSE_{CV}) niż na zbiorze, z którego wyeliminowano wartości odstające (znacznie niższe wartości MSE_{CV}). Wynik ten potwierdzono badając istotność różnic między wspomnianymi wartościami błędu średniokwadratowego MSE_{CV} z wykorzystaniem testu Manna–Whitneya–Wilcoxona (szczegółowo procedurę tę przedstawiono w pracy [Trzęsiok, 2013]).

Podsumowanie

W pracy przedstawiono wybrane metody identyfikacji obserwacji odstających, które pozwalają na wstępną analizę zbioru danych, a tym samym mogą zwrócić uwagę badacza na pewne anomalie występujące w tym zbiorze. Nie ma jednak żadnej gwarancji, że wśród rzeczywistych danych metody te wykryją wszystkie obserwacje odstające.

Warto również zwrócić uwagę na to, że występowanie obserwacji odstających nie oznacza konieczności usunięcia ich ze zbioru danych. Mogą one bowiem mieć istotny, ale pozytywny wpływ na zbudowany model. Dobrym wyjściem z sytuacji jest zastosowanie do analizy takiego zbioru danych metod odpornych. W tej pracy sprawdzano, czy trzy nieparametryczne metody regresji: PPR, POLYMARS i RANDOM FORESTS można uznać za odporne.

Wyniki przeprowadzonych badań pokazują jednak jednoznacznie, że wybrane metody regresji mają znacznie mniejsze wartości błędów średniokwadratowych MSE_{CV} po usunięciu ze zbioru danych obserwacji nietypowych. Być może jest to wynikiem specyfiki badanego zbioru *Mieszkania*, niemniej nie można uznać tych metod za odporne na występowanie w zbiorze uczącym wartości odstających.

Literatura

Breiman L. (2001), *Random Forests*, „Machine Learning", No. 45, s. 5-32.

Cook R.D. (1977), *Detection of Influential Observations in Linear Regression*, „Technometrics", No. 19 (1), s. 15-18.

Filzmoser P., Maronna R.A., Werner M. (2008), *Outlier Identification in High Dimensions*, „Computational Statistics & Data Analysis", no. 52, s. 1694-1711.

Friedman J., Stuetzle W. (1981), *Projection Pursuit Regression*, „Journal of the American Statistical Association", No. 76, s. 817-823.

Healy M. J. R. (1968), *Multivariate Normal Plotting*, „Applied Statistics", No. 17, s. 157-161.

Jajuga K. (1993), *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN, Warszawa.

Kooperberg C., Bose S., Stone C. (1997), *Polychotomous Regression*, „Journal of the American Statistical Association", No. 92, s. 117-127.

Meyer D., Leisch F., Hornik K. (2003), *The Support Vector Machine under Test*, „Neurocomputing", Vol. 1-2, No. 55, s. 169-186.

Trzęsiok J. (2011), *Przegląd metod regularyzacji w zagadnieniach regresji nieparametrycznej* [w:] Jajuga K., Walesiak M., red., *Taksonomia 18. Klasyfikacja i analiza danych*, Wydawnictwo UE, Wrocław, s. 330-339.

Trzęsiok J. (2013), *Wybrane symulacyjne techniki porównywania nieparametrycznych metod regresji* [w:] Jajuga K., Walesiak M., red. *Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania*, Wydawnictwo UE, Wrocław, s. 197-205.

Trzpiot G. , red. (2013), *Wybrane elementy statystyki odpornej*, Wydawnictwo UE, Katowice.

Tukey J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley.

ROBUSTNESS FOR OUTLIERS OF SELECTED NONPARAMETRIC REGRESSION MODELS

Summary: The paper presents an important problem of robustness for outliers in regression. In the first part selected outliers detection techniques are described. Moreover, we empirically examine the robustness of the following methods: PPR, POLYMARS and RANDOM FORESTS on real world dataset. We show, that after removing outliers the prediction abilities of the models increase.

Keywords: outliers, robust, nonparametric regression.