EKONOMETRIA ECONOMETRICS 4(46) • 2014

ISSN 1507-3866

**Marcin Pełka**
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: marcin.pelka@ue.wroc.pl

# PODEJŚCIE WIELOMODELOWE W REGRESJI DANYCH SYMBOLICZNYCH INTERWAŁOWYCH

**Streszczenie:** Podejście wielomodelowe, które polega na zastosowaniu wielu modeli (zamiast jednego), może z powodzeniem znaleźć zastosowanie w analizie danych klasycznych. Celem artykułu jest wskazanie przydatności zastosowania podejścia wielomodelowego z wykorzystaniem metody bagging w regresji danych symbolicznych interwałowych. W artykule zaprezentowano podstawowe pojęcia związane z analizą danych symbolicznych, adaptację metody największych kwadratów na potrzeby danych symbolicznych interwałowych oraz ideę metody bagging. W części empirycznej artykułu przedstawiono wyniki badań z zastosowaniem sztucznych oraz rzeczywistych zbiorów danych dla metody środków oraz metody środków i promieni. Przeprowadzone badania symulacyjne z zastosowaniem różnej liczby modeli bazowych wskazują, że podejście wielomodelowe z zastosowaniem metody bagging pozwala na poprawę dokładności otrzymanych wyników zarówno dla metody środków, jak i metody środków i promieni.

**Słowa kluczowe:** podejście wielomodelowe, regresja danych symbolicznych, dane interwałowe.

DOI: 10.15611/ekt.2014.4.18

## 1. Wstęp

Podejście wielomodelowe polega na agregacji, czyli łączeniu, wyników otrzymanych z wielu modeli bazowych $(D_1,\ldots,D_M)$ w jeden model połączony (zagregowany) $D^*$ (por. np. [Kuncheva 2004, s. 6-7; Gatnar 2008, s. 62]). Celem zastosowania podejścia wielomodelowego, zamiast modelu pojedynczego, jest zmniejszenie błędu predykcji. Wynika to z faktu, że model zagregowany jest zwykle bardziej dokładny niż którykolwiek z modeli, które wchodzą w jego skład (zob. [Gatnar 2008, s. 62]).

Jedną z podstawowych technik łączenia modeli bazowych w jeden model złożony jest metoda agregacji bootstrapowej, którą zaproponował Beiman w 1996 r. Metoda ta jest znana jako bagging (por. [Gatnar 2008, s. 140; Breiman 1996, s. 123]).

Celem artykułu jest zaproponowanie zastosowania metody agregacji bootstrapowej na użytek agregacji modeli regresji danych symbolicznych na przykładzie da-

nych symbolicznych interwałowych. W części empirycznej artykułu przedstawiono wyniki badań symulacyjnych z zastosowaniem rzeczywistych i sztucznych zbiorów danych.

## 2. Dane symboliczne

Obiekty symboliczne, w przeciwieństwie do obiektów w ujęciu klasycznym, mogą być opisywane przez następujące rodzaje zmiennych [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006, s. 7-30; Dudek 2013, s. 35-36]:

- zmienne nominalne,
- zmienne porządkowe,
- zmienne przedziałowe,
- zmienne ilorazowe,
- zmienne interwałowe – czyli przedziały liczbowe,
- zmienne wielowariantowe – czyli listy kategorii lub wartości,
- zmienne wielowariantowe z wagami – czyli listy kategorii z wagami,
- zmienne histogramowe – czyli listy wartości z wagami.

Przykładowe realizacje zmiennych symbolicznych różnego typu zawarto w tab. 1.

Szerzej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi piszą m.in.: [Bock, Diday (red.) (2000), s. 2-8; Dudek 2013, s. 42-43; 2004; Billard, Diday 2006, s. 7-66; Noirhomme-Fraiture, Brito 2011; Diday, Noirhomme-Fraiture 2008, s. 3-30].

**Tabela 1.** Przykłady realizacji zmiennych symbolicznych

| Zmienna symboliczna | Realizacje zmiennej | Typ zmiennej symbolicznej |
|---|---|---|
| Ilość zużytego paliwa w ciągu miesiąca (w l) | <5, 10>; <7, 15>; <12, 30>; <20, 50>; <10, 45> | interwałowa (przedziały nierozłączne) |
| Pojemność silnika (w $cm^3$) | <1000, 1200>; <1300, 1400> <1500, 1600>, <1600, 1800> | interwałowa (przedziały rozłączne) |
| Preferowany kolor samochodu | {czerwony, zielony, niebieski} {żółty, zielony, czarny} | wielowariantowa |
| Preferowany sposób podróży | {samochód (0,8); pociąg (0,2)} {pociąg (0,6); autobus (0,3)} {samochód (0,9); rower (0,1)} | wielowariantowa z wagami |
| Czas dojazdu do pracy (w min.) | {<0, 15) (0,3); <15, 30) (0,7)} {<0, 15) (0,75); <15, 30) (0,15)} {<15, 30) (1,00)} | histogramowa |

Źródło: opracowanie własne na podstawie: [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006; s. 7-30; Noirhomme-Fraiture, Brito 2011].

## 3. Regresja danych symbolicznych interwałowych

Ponieważ metoda najmniejszych kwadratów jest szeroko i dokładnie opisywana w literaturze przedmiotu (zob. np. [Welfe 2003; Kufel 2011; Sobczyk 2013; Jajuga (red.) 1999; Dziechciarz (red.) 2002], w tej części artykułu zostaną przedstawione najważniejsze podstawy związane z metodą najmniejszych kwadratów, tak aby później przedstawić jej rozszerzenie dla danych symbolicznych interwałowych.

Liniowy model regresji dla wielu zmiennych przedstawia się za pomocą następującego równania:

$$Y_t = b_0 X_{0t} + b_1 X_{1t} + \ldots + b_m X_{mt} + e_t = \sum_{j=0}^{m} b_j X_j + e_t, \tag{1}$$

gdzie: $Y$ – zmienna objaśniana (regresant), $X_0, X_1, \ldots, X_m$ – zmienne objaśniające (regresyjne), $b_0, b_1, \ldots, b_m$ – parametry strukturalne modelu, $e$ – składnik losowy, $t = 1 \ldots, T$ – numer obserwacji, $j = 0,1,\ldots,m$ – numer zmiennej objaśniającej.

W zapisie macierzowym model wyrażony równaniem (1) przyjmuje postać:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}. \tag{2}$$

Model ten wymaga spełnienia wielu różnorodnych założeń, które są jego ograniczeniami i które dobrze opisano w literaturze przedmiotu (por. np. [Walesiak, Gatnar (red.) 2004, s. 83-84; Lattin, Carroll, Green 2003, s. 43-47; Welfe 2003, s. 29-32]). Do szacowania wartości parametrów tego modelu wykorzystuje się metodę najmniejszych kwadratów, gdzie poszukuje się takiego ich estymatora $\left(\hat{\mathbf{b}}\right)$, który będzie minimalizował sumę kwadratów odchyleń wartości empirycznych zmiennej zależnej od jej wartości teoretycznych:

$$S = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \rightarrow \min. \tag{3}$$

Po wyznaczeniu pochodnych tej funkcji względem $\hat{\mathbf{b}}$, przyrównaniu do zera oraz rozwiązaniu dla $\hat{\mathbf{b}}$ otrzymujemy:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{4}$$

Ponieważ w przypadku danych symbolicznych interwałowych, zamiast z pojedynczą liczbą, mamy do czynienia z przedziałem liczbowym, w literaturze przedmiotu opracowano następujące rozwiązania [Lima-Neto, de Carvalho 2008, s. 1500-1515; 2010, s. 333-347; Billard, Diday, 2006, s. 198-201; Diday, Noirhomme-Fraiture 2008, s. 360-361][1]:

1. Metodę środków (*center method*), gdzie elementy macierzy **X** oraz **Y** są zastępowane przez środki przedziałów dla tych zmiennych. Wówczas wzór (3) przyjmuje postać:

[1] W dalszej części artykułu zaprezentowane zostaną metoda środków oraz metoda środków i promieni.

$$\hat{\mathbf{b}} = \left( \left( \mathbf{X}^c \right)^T \left( \mathbf{X}^c \right) \right)^{-1} \left( \mathbf{X}^c \right)^T \mathbf{y}^c, \tag{5}$$

gdzie: $\mathbf{X}^c$ – macierz środków zmiennych objaśniających, $\boldsymbol{y}^c$ – macierz środków zmiennej objaśnianej.

Oszacowane wartości teoretyczne zmiennej objaśnianej oblicza się odrębnie dla krańców dolnych (oznaczanych indeksem *L*) i górnych (oznaczanych indeksem *U*) tej zmiennej, zgodnie ze wzorem:

$$\hat{\mathbf{y}}_L = \left( \mathbf{X}_L \right)^T \hat{\mathbf{b}} \text{ oraz } \hat{\mathbf{y}}_U = \left( \mathbf{X}_U \right)^T \hat{\mathbf{b}}. \tag{6}$$

2. Metodę środków i promieni (*center and range method*), gdzie elementy macierzy **X** oraz **Y** są zastępowane przez środki przedziałów oraz połowę rozpiętości przedziału (promienie) dla tych zmiennych. Wówczas wzór (3) przyjmuje postać:

$$\begin{aligned} \hat{\mathbf{b}}^C &= \left( \left( \mathbf{X}^c \right)^T \left( \mathbf{X}^c \right) \right)^{-1} \left( \mathbf{X}^c \right)^T \mathbf{y}^c \\ \hat{\mathbf{b}}^r &= \left( \left( \mathbf{X}^r \right)^T \left( \mathbf{X}^r \right) \right)^{-1} \left( \mathbf{X}^r \right)^T \mathbf{y}^r \end{aligned}. \tag{7}$$

Oznacza to, że *de facto* osobno oszacowywane są parametry środków i promieni dla zmiennych. Ostatecznie wartości teoretyczne zmiennej objaśnianej oblicza się odrębnie dla krańców dolnych (oznaczanych indeksem *L*) i górnych (oznaczanych indeksem *U*) zgodnie ze wzorem:

$$\hat{\mathbf{y}}_L = \hat{\mathbf{y}}^C - \hat{\mathbf{y}}^r \text{ oraz } \hat{\mathbf{y}}_U = \hat{\mathbf{y}}^C + \hat{\mathbf{y}}^r, \tag{8}$$

gdzie: $\hat{\mathbf{y}}^C = \left( \mathbf{X}^C \right)^T \hat{\mathbf{b}}^C$ – oszacowane wartości dla środków,

$\hat{\mathbf{y}}^r = \left( \mathbf{X}^r \right)^T \hat{\mathbf{b}}^r$ – oszacowane wartości dla promieni.

3. Rozszerzenia metod regresji z funkcją kary, tj. regresji grzbietowej (*ridge regression*), regresji lasso (*lasso regression*) oraz sieci elastycznych (*elastic net model*)[2].

W przypadku danych symbolicznych interwałowych miarą dopasowania modelu do danych są dwa współczynniki dopasowania $R^2$. Jeden, oznaczany jako $R_L^2$, wskazuje na dopasowanie modelu do danych, biorąc pod uwagę dolne krańce przedziałów zmiennych symbolicznych. Drugi oznaczany jest jako $R_U^2$ i wskazuje na dopasowanie modelu do danych względem górnych krańców przedziału zmiennych symbolicznych.

W przypadku pozostałych typów danych symbolicznych dla każdego z nich opracowano rozwiązania pozwalające na zastosowanie metody najmniejszych kwa-

[2] Szerzej o metodach regresji dla danych symbolicznych interwałowych z zastosowaniem funkcji kary piszą m.in.: [Lima-Neto, de Carvalho 2010, s. 333-347; 2008, s. 1500-1515].

dratów dla danego typu zmiennej (zob. [Billard, Diday 2006, s. 189-227; Diday, Noirhomme-Fraiture 2008, s. 359-372]).

W przypadku zarówno zmiennych symbolicznych interwałowych, jak i innych typów zmiennych problematyczna wydaje się możliwość prawidłowej weryfikacji założeń metody najmniejszych kwadratów.

## 4. Metoda bagging

Jak wspomniano we wstępie, metoda bagging jest jedną z bardziej znanych metod agregacji modeli bazowych (por. [Gatnar 2008, s. 140; Breiman 1996, s. 123; Kuncheva 2004, s. 203]). Metoda bagging polega na zbudowaniu $M$ modeli bazowych na podstawie prób uczących $U_1, \ldots, U_M$ losowanych ze zwracaniem ze zbioru uczącego. Próby te nazywa się próbami bootstrapowymi [Gatnar 2008, s. 140]. Zwykle około 37% obiektów ze zbioru danych nie trafia do żadnej z prób uczących. Obiekty te tworzą dodatkowy zbiór danych, tzw. OOB (*Out-Of-Bag*), który często stosowany jest jako dodatkowy zbiór testowy [Gatnar 1998, s. 140].

Algorytm metody bagging można przedstawić za pomocą następujących kroków [Polikar 2007, s. 60-61; Gatnar 2008, s. 140; Kuncheva 2004, s. 204]:

1. Ustalenie liczby modeli bazowych.
2. Dla każdego z modeli bazowych wykonywane są następujące kroki:

a) wylosowanie próby bootstrapowej,

b) budowa modelu bazowego na podstawie próby bootstrapowej,

c) predykcja na podstawie modelu zagregowanego, zbudowanego na bazie modeli bazowych, dokonywana jest z zastosowaniem metody głosowania większościowego w przypadku dyskryminacji lub przez uśrednienie wyników w przypadku regresji.

Ponieważ w części empirycznej wykorzystana zostanie regresja liniowa danych symbolicznych, wyniki dla modelu zagregowanego zostaną obliczone jako średnia arytmetyczna z modeli bazowych.

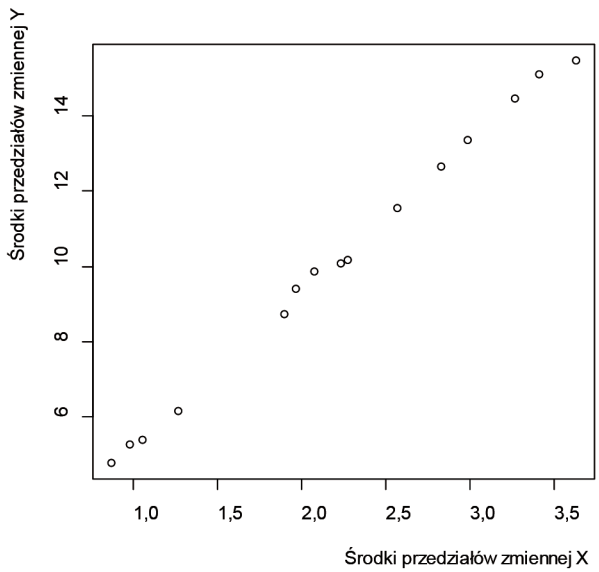
## 5. Wyniki badań symulacyjnych

Na potrzeby badań symulacyjnych przygotowano w programie R dwa sztuczne zbiory danych oraz trzy zbiory danych rzeczywistych. Pierwszy zbiór danych rzeczywistych zawiera obiekty symboliczne drugiego rzędu, które są wynikiem agregacji danych na temat zbiorów pszenicy w latach 1999-2013 w zależności od zużycia nawozów NPK w przekroju województw. Drugi zbiór danych rzeczywistych zawiera dane pacjentów kardiologicznych – zawiera on dane na temat 11 pacjentów. Trzeci zbiór danych rzeczywistych zawiera dane dotyczące przestępczości na terenie USA – zbiór ten zawiera dane na temat 46 stanów USA. Drugi i trzeci zbiór danych rzeczywistych pochodzi z pakietu RSDA programu R.

Pierwszy sztuczny zbiór danych zawiera jedną interwałową zmienną symboliczną objaśnianą oraz jedną interwałową zmienną symboliczną objaśnianą. W zbiorze danych zawarto szesnaście obserwacji. Na rysunku 1 przedstawiono wykres korelacyjny dla środków przedziałów zmiennej objaśnianej i objaśniającej.

Drugi sztuczny zbiór danych zawiera cztery interwałowe zmienne symboliczne objaśniające oraz jedną interwałową zmienną symboliczną objaśnianą. W zbiorze danych zawarto czternaście obserwacji.

W pierwszym kroku zbudowano z zastosowaniem programu R[3] pojedynczy model regresji z zastosowaniem metody środków oraz metody środków i promieni. Wyniki dopasowania modeli do danych oraz średni absolutny błąd procentowy (MAPE) dopasowania prognoz do danych zawarto w tab. 2. Na potrzeby podejścia wielomodelowego zbudowano różną liczbę modeli bazowych dla każdego ze zbiorów. Wyniki zagregowane otrzymano, uśredniając wyniki ze wszystkich modeli [Gatnar 2008, s. 140]. Rezultaty zastosowania podejścia wielomodelowego zawarto w tab. 3.




**Rys. 1.** Wykres korelacyjny dla środków przedziałów zmiennej Y oraz X z pierwszego (sztucznego) zbioru danych

Źródło: opracowanie własne z wykorzystaniem programu R.

Z danych zawartych w tab. 2 i 3 wynika, że w przypadku pojedynczego modelu metoda środków i promieni uzyskuje zwykle lepsze wyniki niż metoda środków.

[3] Zarówno w przypadku modelu pojedynczego, jak i podejścia wielomodelowego przygotowano autorskie procedury w programie R.

**Tabela 2.** Wyniki otrzymane dla pojedynczych modeli regresji

| Zbiór danych | Metoda | $R_L^2$ | $R_U^2$ | $MAPE_L$ | $MAPE_U$ |
|---|---|---|---|---|---|
| Sztuczny zbiór danych I | metoda środków | 0,89 | 0,91 | 7,28% | 4,98% |
| | metoda środków i promieni | 0,99 | 0,98 | 2,54% | 1,72% |
| Sztuczny zbiór danych II | metoda środków | 0,83 | 0,95 | 5,72% | 3,63% |
| | metoda środków i promieni | 0,94 | 0,99 | 3,94% | 0,03% |
| Rzeczywisty zbiór danych – zbiory pszenicy | metoda środków | 0,43 | 0,28 | 13% | 11,93% |
| | metoda środków i promieni | 0,75 | 0,65 | 11,28% | 8,94% |
| Rzeczywisty zbiór danych – dane medyczne | metoda środków | 0,60 | 0,65 | 10,11% | 11,09% |
| | metoda środków i promieni | 0,67 | 0,74 | 10% | 11% |
| Rzeczywisty zbiór danych – przestępstwa na terenie USA | metoda środków | 0,57 | 0,59 | 12,12% | 12,15% |
| | metoda środków i promieni | 0,60 | 0,65 | 11,9% | 11,61% |

$MAPE_L$ i $MAPE_U$ oznaczają odpowiednio średni absolutny błąd procentowy dla krańca dolnego i górnego przedziału zmiennych symbolicznych interwałowych.

Źródło: opracowanie własne na podstawie z zastosowaniem autorskich procedur programu R.

**Tabela 3.** Wyniki otrzymane dla podejścia zagregowanego

| Zbiór danych | Liczba modeli | Metoda | $R_L^2$ | $R_U^2$ | $MAPE_L$ | $MAPE_U$ |
|---|---|---|---|---|---|---|
| Sztuczny zbiór danych I | 20 | metoda środków | 0,891 | 0,92 | 7,23% | 4,92% |
| | | metoda środków i promieni | 0,996 | 0,997 | 2,51% | 1,67% |
| Sztuczny zbiór danych II | 30 | metoda środków | 0,93 | 0,99 | 4,97% | 2,54% |
| | | metoda środków i promieni | 0,997 | 0,999 | 2,65% | 0% |
| Rzeczywisty zbiór danych | 30 | metoda środków | 0,51 | 0,58 | 13,72% | 8,60% |
| | | metoda środków i promieni | 0,82 | 0,81 | 10,26% | 8,53% |
| Rzeczywisty zbiór danych – dane medyczne | 20 | metoda środków | 0,70 | 0,72 | 9,09% | 9% |
| | | metoda środków i promieni | 0,72 | 0,77 | 8,97% | 8,99% |
| Rzeczywisty zbiór danych – przestępstwa na terenie USA | 50 | metoda środków | 0,61 | 0,76 | 10% | 9,78% |
| | | metoda środków i promieni | 0,87 | 0,89 | 10,01% | 9,89% |

Źródło: opracowanie własne na podstawie z zastosowaniem autorskich procedur programu R.

Natomiast podejście wielomodelowe w regresji danych symbolicznych pozwala na uzyskanie znacznie lepszego dopasowania modelu do danych (w sensie miar $R_L^2$ oraz $R_U^2$, a także średniego absolutnego błędu procentowego – MAPE). Dodatkowo, podobnie jak w przypadku pojedynczego modelu, wyniki otrzymane z zastosowaniem metody środków i promieni są bardziej dokładne niż te otrzymane za pomocą metody środków.

## 6. Zakończenie

Podejście wielomodelowe może znaleźć z powodzeniem zastosowanie w analizie regresji danych symbolicznych interwałowych, a także danych symbolicznych innych typów.

Podobnie jak w przypadku regresji liniowej dla danych klasycznych, tak w przypadku regresji liniowej danych symbolicznych interwałowych w wyniku zastosowania podejścia wielomodelowego badania symulacyjne wskazują, że podejście to pozwala otrzymać lepsze dopasowanie modeli do danych (w sensie miar $R_L^2$ oraz $R_U^2$, a także średniego absolutnego błędu procentowego – MAPE).

Wyniki badań empirycznych otrzymane dla metody środków oraz metody promieni pozwalają wskazać, że metoda promieni pozwala uzyskać nieco lepsze dopasowanie modelu do danych, niż ma to miejsce w przypadku metody środków. Wynika to z faktu, że metoda środków i promieni bierze pod uwagę, oprócz samego środka przedziału zmiennej symbolicznej, także jej promień, dzięki czemu zbudowany model regresji liniowej jest lepiej dopasowany do danych niż w przypadku zastosowania metody środków. Dodatkowo w każdym przypadku dopasowanie modeli jest lepsze w przypadku sztucznych zbiorów danych, niż ma to miejsce w przypadku rzeczywistych zbiorów danych. Wynika to z dwóch faktów. Po pierwsze, w przypadku sztucznych zbiorów danych przedziały zmiennych symbolicznych były nieco krótsze (miały mniejszy promień) niż w przypadku zbiorów danych klasycznych. Po drugie, w przypadku rzeczywistych zbiorów danych możemy mieć do czynienia ze zmiennymi, które zakłócają istniejącą zależność między zmiennymi. Dodatkowo w przypadku sztucznych zbiorów danych dane przygotowano w taki sposób, aby korelacje między zmienną objaśnianą a zmiennymi objaśniającymi były jak największe, a korelacje między zmiennymi objaśniającymi jak najmniejsze.

Warto także zwrócić uwagę, że podejście wielomodelowe ma także swoje ograniczenia, które dość dobrze zostały opisane w literaturze przedmiotu (zob. np. [Kuncheva 2004]).

Celem dalszych badań będzie analiza porównawcza wszystkich proponowanych podejść w zakresie regresji danych symbolicznych interwałowych z zastosowaniem sztucznych i rzeczywistych zbiorów danych różnego typu.

## Literatura


Bock H.-H., Diday E. (red.), 2000, *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg.

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

Breiman L., 1996, *Bagging predictors*, Machine Learning, vol. 24, s. 123-140.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu, Wrocław.

Dudek A., 2004, *Tworzenie obiektów symbolicznych z baz danych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021, s. 107-114.

Dziechciarz J. (red.), 2002, *Ekonometria. Metody, przykłady, zadania*, Wyd. Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Hastie T., Tibshirani R., Friedman J., 2008, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York.

Jajuga K. (red.), 1999, *Ekonometria. Metody i analiza problemów ekonomicznych*, Wyd. Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Kuncheva L., 2004, *Combining Pattern Classifiers. Methods and Algorithms*, Wiley, New Jersey.

Kufel T., 2011, *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*, PWN, Warszawa.

Lattin J., Carroll J.D., Green P.E., 2003, *Analyzing Multivariate Data*, Brooks/Cole – Thomson Learning, Pacific Grove.

Lima-Neto E.A., de Carvalho F.A.T., 2008, *Centre and range method to fitting a linear regression model on symbolic interval data*, Computational Statistics and Data Analysis, vol. 52, s. 1500-1515.

Lima-Neto E.A., de Carvalho F.A.T., 2010, *Constrained linear regression models for symbolic interval-valued variables*, Computational Statistics and Data Analysis, vol. 54, s. 333-347.

Noirhomme-Fraiture M., Brito P., 2011, *Far beyond the classical data models: symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, issue 2, s. 157-170.

Oldemar R., 2014, *The* `RSDA` *package of R software*, http://www.r-project.org.

Polikar R., 2007, *Bootstrap inspired techniques in computational intelligence*, IEEE Signal Processing Magazine, vol. 24, no. 4, s. 56-72.

Sobczyk M., 2013, *Ekonometria*, C.H. Beck, Warszawa.

Walesiak M., Gatnar E. (red.), 2004, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wyd. Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Welfe A., 2013, *Ekonometria*, PWN, Warszawa.

# ENSEMBLE LEARNING IN REGRESSION MODEL OF SYMBOLIC INTERVAL DATA

**Summary:** Ensemble learning, which consist in using a lot of models (instead one single model) can be used in classical data analysis. The aim of the paper is to present an adaptation of ensemble learning with the use of bagging for regression analysis of symbolic interval-valued data. The article presents basic concepts concerning symbolic data analysis, the adaptation of ordinary least squares model for symbolic interval-valued data and the idea of bagging approach in ensemble learning. The empirical part contains the results of simulation studies obtained with the application of real and artificial data sets for centers and centers and range methods. The results show that both methods reach usually better results when using bagging than in case of a single model.

**Keywords:** ensemble learning, regression of symbolic data, interval-valued data.