PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu nr 309

RESEARCH PAPERS
of Wrocław University of Economics No. 309

Społeczno-gospodarcze aspekty statystyki

Redaktorzy naukowi

Zofia Rusnak
Edyta Mazurek

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Joanna Szynal

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-398-4

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp 9

Tadeusz Bednarski: Rola Jerzego Spławy-Neymana w kształtowaniu metod statystycznej analizy przyczynowości 11

Filip Borowicz: Ocena możliwości uzupełnienia danych BAEL informacjami ze źródeł administracyjnych w celu dokładniejszej analizy danych o bezrobociu 19

Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Przydatność testów nieparametrycznych Kruskala-Wallisa i mediany w długoterminowej ocenie parametrów kruszyw melafirowych 27

Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Karty kontrolne w ocenie jakości kruszyw dla budownictwa drogowego 42

Czesław Domański: Uwagi o procedurach weryfikacji hipotez z brakującą informacją 54

Stanisław Heilpern: Zależne procesy ryzyka 62

Artur Lipieta, Barbara Pawełek, Jadwiga Kostrzewska: Badanie struktury wydatków w ramach wspólnej polityki UE z wykorzystaniem analizy korespondencji 78

Agnieszka Marciniuk: Dwa sposoby modelowania stopy procentowej w ubezpieczeniach życiowych 90

Beata Bieszk-Stolorz, Iwona Markowicz: Model nieproporcjonalnej intensywności Coxa w analizie bezrobocia 114

Edyta Mazurek: Statystyczna analiza podatku dochodowego od osób fizycznych 127

Katarzyna Ostasiewicz: Awersja do nierówności w modelowaniu użytkowania dóbr wspólnych 159

Piotr Peternek: Porównanie kart kontrolnych indywidualnych pomiarów uzyskanych z wykorzystaniem uogólnionego rozkładu lambda oraz krzywych Johnsona 179

Małgorzata Podogrodzka: Starzenie się ludności a płodność w Polsce w latach 1991-2010 – ujęcie regionalne 192

Renata Rasińska, Iwona Nowakowska: Jakość życia studentów w aspekcie znajomości wskaźników zrównoważonego rozwoju 203

Maria Rosienkiewicz, Jerzy Detyna: Analiza efektywności metod wyboru zmiennych objaśniających do budowy modelu regresyjnego 214
Jerzy Śleszyński: National Welfare Index – ocena nowego miernika rozwoju trwałego i zrównoważonego 236
Maria Szmuksta-Zawadzka, Jan Zawadzki: Wykorzystanie oszczędnych modeli harmonicznych w prognozowaniu na podstawie szeregów czasowych o wysokiej częstotliwości w warunkach braku pełnej informacji..... 261
Anna Zięba: O możliwościach wykorzystania metod statystycznych w badaniach nad stresem 278

Summaries

Tadeusz Bednarski: Role of Jerzy Spława-Neyman in statistical inference for causality 18
Filip Borowicz: Assessing the possibility of supplementing the Polish LFS data with register records for more detailed unemployment data analysis. 26
Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Usefulness of nonparametric Kruskal-Wallis and median tests in long-term parameters assessment of melaphyre crushed rocks 41
Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Control charts in the assessment of aggregates quality for road construction..... 53
Czesław Domański: Some remarks on the procedures of the verification of hypotheses under incomplete information..... 61
Stanisław Heilpern: Dependent risk processes 77
Artur Lipieta, Barbara Pawełek, Jadwiga Kostrzewska: Study of the structure of expenditure under the EU's common policy using correspondence analysis 89
Agnieszka Marciniuk: Two ways of stochastic modelling of interest rate in life insurances 113
Beata Bieszk-Stolorz, Iwona Markowicz: The Cox non-proportional hazards model in the analysis of unemployment..... 126
Edyta Mazurek: Statistical assessment of Personal Income Tax 158
Katarzyna Ostasiewicz: Inequality aversion in modeling the use of common pool resources 178
Piotr Peternek: Comparison of control charts of individual measurements based on general Lambda distribution and Johnson curves..... 191
Małgorzata Podogrodzka: The ageing of the population and fertility in Poland in the years 1991-2010 by voivodeships..... 202
Renata Rasińska, Iwona Nowakowska: Students' life quality in terms of knowledge of sustainable development indicators 213

Maria Rosienkiewicz, Jerzy Detyna: Efficiency analysis of chosen methods of explanatory variables selection within the scope of regression model construction 235
Jerzy Śleszyński: *National Welfare Index* – assessment of a new measure of sustainable development 260
Maria Szmuksta-Zawadzka, Jan Zawadzki: The application of harmonic models in forecasting based on high frequency time series in condition of lack of full information 277
Anna Zięba: About statistical methods in the study on stress 284

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 309 • 2013
Społeczno-gospodarcze aspekty statystyki ISSN 1899-3192

Maria Rosienkiewicz, Jerzy Detyna*
Politechnika Wrocławska

ANALIZA EFEKTYWNOŚCI METOD WYBORU ZMIENNYCH OBJAŚNIAJĄCYCH DO BUDOWY MODELU REGRESYJNEGO

Streszczenie: Podstawowym celem pracy jest zbadanie oraz porównanie kryterium informacyjnego Akaike, kryterium informacyjnego Schwarza, metody entropii krzyżowej i metody wskaźników pojemności informacji Hellwiga pod kątem efektywności konstrukcji modeli regresyjnych. Badanie przeprowadzono na podstawie symulacji komputerowych. Po wygenerowaniu zestawu danych o rozkładzie normalnym zbudowano model liniowy, w którym zmienna objaśniana jest zależna od wybranych zmiennych wcześniej wygenerowanych. W kolejnym etapie rozszerzono zbiór potencjalnych zmiennych objaśniających i zastosowano badane metody wyboru modelu. Powyższe kroki powielono, a następnie porównano, jak często każda z badanych metod wskazała właściwy zbiór zmiennych i tym samym właściwy model. Porównania metod dokonano także na danych empirycznych.

Słowa kluczowe: kryteria informacyjne Akaike i Schwarza, metoda Hellwiga, entropia krzyżowa.

1. Wstęp

W nauce organizacji produkcji, w celu wspomagania podejmowania decyzji, prognozowania wybranych wartości bądź szukania źródłowych przyczyn pewnych zjawisk, zaleca się stosowanie programów do modelowania i symulacji. Słuszność stosowania modeli symulacyjnych jest niepodważalna, jednak w praktyce przemysłowej bardzo często zdarza się, że firmy nie decydują się na zakup odpowiedniego oprogramowania, ze względu na bardzo wysokie koszty i skomplikowaną obsługę, która wymaga zazwyczaj znajomości programowania. Dlatego też warto zauważyć, że zastosowanie narzędzi modelowania ekonometrycznego w dziedzinie organizacji produkcji może stać się alternatywą dla drogiego i często trudnego oprogramowania

* Zadanie współfinansowane ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego. Projekt systemowy „Grant Plus" Program Operacyjny Kapitał Ludzki, Priorytet VIII Regionalne Kadry Gospodarki, Działanie 8.2 Transfer Wiedzy, Poddziałania 8.2.2. Regionalne Strategie Innowacji.

do symulacji. Oczywiście takie modele nie są tak uniwersalne jak modele symulacyjne zbudowane w odpowiednim środowisku informatycznym, jednak – odpowiednio skonstruowane – mogą wspomagać przedsiębiorców w podejmowaniu prawidłowych decyzji, prognozowaniu czy też wyjaśnianiu wystąpienia pewnych zjawisk.

W celu stosowania najlepszej metody wyboru modelu w nauce organizacji produkcji w niniejszej pracy porównano wybrane metody selekcji zmiennych objaśniających – bardzo popularne kryterium informacyjne Akaike (AIC) i bayesowskie kryterium informacyjne Schwarza (BIC), metodę entropii krzyżowej, zalecaną w polskich podręcznikach akademickich do ekonometrii metodę wskaźników pojemności informacji Hellwiga.

Podstawowym celem pracy jest zbadanie oraz porównanie wymienionych metod selekcji zmiennych objaśniających pod kątem efektywności konstrukcji modeli regresyjnych. Aby zadanie to zrealizować, opracowano specjalny program w środowisku R (*The R Project for Statistical Computing*, *R language*). Program ten umożliwi dokonanie oceny efektywności zastosowania badanych metod do modelowania procesów zachodzących w obszarze produkcji. Na podstawie wyników generowanych z tego programu będzie można stwierdzić, która z metod jest bardziej efektywnym narzędziem wyboru zmiennych objaśniających do modelu i tym samym narzędziem wyboru modelu optymalnego. Dodatkowym celem pracy jest ocena możliwości zastosowania badanych metod w modelowaniu wybranych zależności zachodzących na różnych etapach procesu produkcyjnego. Ocena ta zostanie przeprowadzona w oparciu o analizę problemu kompensacji błędów obróbki pewnego centrum tokarskiego.

2. Przykłady możliwości zastosowania modelowania ekonometrycznego w inżynierii produkcji

W inżynierii produkcji zachodzi wiele zjawisk i procesów, do których modelowania można zastosować metody ekonometryczne. Na każdy proces wpływa liczna grupa czynników. W celu analizy kształtowania się danego procesu należy rozpoznać pewne zmienne, które wnoszą określone informacje o tym procesie. Do modelowania takich zmiennych można wykorzystać analizę regresji. Każda zmienna posiada pewien rozkład prawdopodobieństwa. Za najlepiej zbadany rozkład zmiennej losowej ciągłej można uznać rozkład normalny [Firkowicz 1970, s. 48-49]. W praktyce rozkład w przybliżeniu normalny ma bardzo często cecha mierzalna danego wyrobu lub określona funkcja tej cechy (najczęściej logarytm). Cecha mierzalna ma zazwyczaj rozkład normalny, kiedy rozrzut jej wartości jest wynikiem sumowania się wpływów wielu różnych czynników, z których żaden nie jest dominujący. Taka sytuacja bardzo często występuje w praktyce inżynierskiej. Zatem budowa modelu regresyjnego pewnej cechy mierzalnej danego wyrobu może być bardzo przydatna w zarządzaniu jakością. Model taki pozwoli na zbadanie wpływu różnych czynników na wartość analizowanej cechy. Cechą taką może być pewien wymiar danego wyrobu – na przy-

kład średnica otworu. Dokonując pomiarów takiej średnicy i mierząc jednocześnie inne czynniki towarzyszące procesowi wytwarzania wyrobu, a następnie wykorzystując badane w niniejszej pracy metody selekcji zmiennych, można zbudować model, który pozwoliłby jednoznacznie określić, jakie czynniki i w jakim stopniu wpływają na wymiar średnicy otworu w wytwarzanym wyrobie.

Innym przykładem możliwości zastosowania modelowania ekonometrycznego w inżynierii produkcji jest wykorzystanie analizy regresji do budowy funkcji niezawodności czy zależnej od czasu eksploatacji funkcji intensywności uszkodzeń [Radkowski 2003, s. 58-59]. Takie funkcje posiadają rozkład Weibulla. Ekonometria w inżynierii produkcji może znaleźć zastosowanie m.in. w: modelowaniu procesu tworzenia zapasu w ujęciu wartościowym, statystycznym sterowaniu procesem (SPC), kompensacji błędów obrabiarek, problemie wyboru między modelami dotyczącymi średniego czasu pracy do pierwszego uszkodzenia, modelowaniu funkcji kosztów, spektometrii.

3. Metody dobru zmiennych objaśniających

3.1. Problem selekcji zmiennych

W pewnym okresie w budowie modeli istniała niesłuszna tendencja, żeby uwzględniać w badaniach możliwie największą liczbę zmiennych objaśniających. Liczba tych zmiennych przekraczała nawet czasem liczbę obserwacji w próbie. Powodowało to konieczność wykonywania wielu uciążliwych i obszernych rachunków, natomiast uzyskane wyniki często były pozbawione większego znaczenia praktycznego. Wystąpiła potrzeba, aby ograniczać liczbę zmiennych objaśniających już we wczesnym etapie badań. Nieuwzględnione zmienne były traktowane wtedy jako *quasi*-stałe. Przyjmowanie w pewnych granicach danych zmiennych jako stałych, pozwalało pomijać ich wpływ na zmienną endogeniczną. Gdyby zaszła potrzeba, można było je włączyć do badań w dalszym etapie. W efekcie eliminacja „zbędnych" zmiennych oparta była w głównej mierze na informacjach pozastatystycznych, doświadczeniu oraz intuicji badacza. Postępowanie takie było jednak silnie uzależnione od czynników subiektywnych. Zatem ustalenie zbioru zmiennych, które mają wejść do modelu ekonometrycznego można uznać za podstawowe zagadnienie związane z budową tego modelu. Etap ten zwykle jest najtrudniejszy i decyduje o efekcie dalszej analizy [Grabiński, Wydymus, Zeliaś 1982, s. 13]. Można stwierdzić, że ten pierwszy etap procesu budowy modelu ekonometrycznego ma charakter przygotowawczy [Mercik, Szmigiel 2000, s. 68]. Bardzo istotną kwestią przy wstępnym doborze zmiennych do modelu jest merytoryczna znajomość badanego procesu. Znajomość ta umożliwia wytypowanie zbioru potencjalnych zmiennych objaśniających. Należy wybierać tylko takie determinanty, które istotnie wpływają na kształtowanie się badanego zjawiska [Dziechciarz 2002, s. 30]. Problem optymalne-

go wyboru zmiennych objaśniających spośród wytypowanego wcześniej zbioru potencjalnych zmiennych objaśniających sprowadza się zatem do redukcji wstępnie ustalonego zbioru tychże zmiennych [Kukuła 1999, s.16].

3.2. Kryterium informacyjne Akaike

W 1971 roku Hirotsugu Akaike zaproponował kryterium wyboru modelu, które nazwane zostało kryterium informacyjnym Akaike (*Akaike's information criterion AIC*). Szerzej przedstawił to kryterium w pracach *Information theory and an extension of the maximum likelihood principle* z 1973 roku oraz *A new look at the statistical model identification* z 1974 roku. *AIC* może być interpretowane jako miara odległości pomiędzy modelem dopasowanym do zebranych, niekompletnych, danych statystycznych, a modelem efektywnym, który wygenerował te dane [Cavanaugh, Shumway 1998, s. 1]. Kryterium *AIC* opiera się na estymacji informacji Kullbacka-Leiblera za pomocą metody największej wiarygodności. Akaike wyprowadził kryterium informacyjne *AIC*, postaci:

$$AIC(\hat{\theta}) = -2\ln \hat{L} + 2p\,,$$

gdzie: p – liczba parametrów modelu (liczba zmiennych objaśniających wraz z wyrazem wolnym),

$\hat{L} = L(\hat{\theta}) = \max\{L(\theta_i, D)\}$ – maksimum funkcji wiarygodności dla estymowanego modelu.

Kryterium informacyjne Akaike jest to zatem różnica podwojonej liczby parametrów modelu i podwojonego logarytmu naturalnego z maksimum funkcji wiarygodności. Spośród różnych modeli za optymalny uznany jest ten, dla którego *AIC* osiąga najmniejszą wartość [Peracchi 2001, s. 331].

Kiedy analiza statystyczna oparta jest na metodzie najmniejszych kwadratów, wzór na *AIC* można zapisać również następująco[1]:

$$AIC = n\ln(\sum_{i=1}^{n} e_i^2) + 2K\,,$$

gdzie: $\sum_{i=1}^{n} e_i^2$ – suma kwadratów reszt,

K – liczba parametrów modelu.

n – wielkość próby.

Obie powyższe postaci wzorów na *AIC* są równoważne.

Kryterium Akaike jest obecnie bardzo popularne i często stosowane w różnych dziedzinach nauki. Stanowi właściwie uniwersalne kryterium wyboru optymalnego modelu. Należy jednak pamiętać, aby być ostrożnym przy stosowaniu tej metody

[1] http://uczelniawarszawska.pl/upl/1181568659.pdf?PHPSESSID=43887cbc2ca6003603a28048ff-7ce921 z dnia 20 lipca 2008 r.

i nie przyjmować otrzymanych wyników bezkrytycznie. W artykule T.W. Arnolda *Uninformative parameters and model selection using Akaike's information criterion* można znaleźć ważne spostrzeżenie dotyczące kryterium informacyjnego Akaike [Arnold 2010, s. 1175-1178].

3.3. Kryterium informacyjne Schwarza

Statystycy bardzo często napotykają na problem wyboru właściwego wymiaru modelu pasującego do danego zbioru obserwacji [Schwarz 1978, s. 461-464]. Typowym przykładem takiego problemu jest wybór stopnia wielomianowego modelu regresyjnego. W takim przypadku zasada największej wiarygodności niezmiennie prowadzi do wyboru najwyższego możliwego wymiaru. Dlatego też nie może ona zostać uznana za prawidłowy sformalizowany sposób wyboru „właściwego" wymiaru modelu. Akaike zasugerował rozszerzenie zasady największej wiarygodności na nieco bardziej ogólny problem – problem wyboru modelu z grupy modeli o różnej liczbie zmiennych. Rozwiązanie Akaike polega na maksymalizowaniu funkcji wiarygodności osobno dla każdego modelu j, otrzymując $M_j(X_1, \ldots, X_n)$, a następnie na wyborze takiego modelu, dla którego wartość $\log M_j(X_1, \ldots, X_n) - k_j$ jest największa, gdzie k_j jest wymiarem modelu (czyli informuje o liczbie parametrów modelu). Gideon Schwarz zaproponował inne podejście do omówionego problemu. Otóż przedstawił następujące kryterium wyboru modelu (*Bayesian Information Criterion* – BIC) – należy wybrać model, dla którego wartość

$$M_j(X_1, \ldots, X_n) - \frac{1}{2} k_j \log n$$

jest największa. Szczegółowe wyliczenia Schwarz przedstawił w artykule *Estimating the dimension of a model*. Jak można zauważyć, kryterium BIC różni się od kryterium Akaike (AIC) tylko drugą częścią wzoru, a dokładniej przemnożeniem rozmiaru modelu (mierzonego przez liczbę parametrów modelu k_j) przez $\frac{1}{2} k_j \log n$ [Acquah de-Graft 2010, s. 1-6]. Pod względem jakościowym zarówno procedura Schwarza, jak i Akaike umożliwia „matematyczne sformułowanie zasady skąpstwa (*the principal of parsimony*) w budowaniu modeli". Jednakże pod względem ilościowym kryterium Schwarza skłania się bardziej niż kryterium Akaike ku modelom o mniejszych rozmiarach (*lower-dimensional models*). Z perspektywy bayesowskiej kryterium BIC jest tak opracowane, aby wskazywać najbardziej prawdopodobny model dla określonego zbioru danych.

Kryteria informacyjne pozwalają porównywać modele dla tej samej zmiennej objaśnianej. Za najlepszy model uznaje się ten, dla którego wartość kryterium jest najniższa. Należy zauważyć, że wartość kryteriów rośnie wraz ze wzrostem sumy kwadratów reszt (jakość dopasowania) oraz liczby parametrów. Za dobry model uznaje się model prosty – posiadający możliwie najmniej parametrów i jednocześnie dobrze dopasowany.

3.4. Metoda entropii krzyżowej

Statystyczna interpretacja pojęcia entropii została rozwinięta głównie dzięki pracom Boltzmanna, Maxwella i Smoluchowskiego. W ramach teorii informacji niepewność o nieznanej zmiennej jest określana ilościowo przez wielkość zwaną entropią. Dodatkowa wiedza o innych badanych zmiennych przyczynia się do redukcji entropii i dzięki temu informacja o nieznanej zmiennej wzrasta [Ramos, Gonzalez-Rodriguez 2008].

Dywergencja Kullbacka-Leiblera jest naturalną miarą pseudoodległości rozkładu prawdopodobieństwa P od empirycznego P_e [Detyna 2007, s. 93]. Określa się ją wzorem:

$$D(P_e\|P) = -\sum_{x\epsilon X} P_e(x) log \frac{P(x)}{P_e(x)}.$$

Jeśli $P = P_e$, wówczas: $D(P_e\|P) = 0.$

W każdym innym wypadku, jeżeli:

$$\sum_{x\epsilon X} P_e(x) = \sum_{x\epsilon X} P(x) = 1, \qquad \text{to } D(P_e\|P) > 0.$$

W obliczeniach praktycznych znacznie wygodniej jest posługiwać się entropią krzyżową (*cross-entropy*), która jest miarą odstępstwa rozkładu teoretycznego P od empirycznego P_e. Entropię krzyżową definiuje wzór:

$$H(P_e\|P) = -\sum_{x\epsilon X} P_e(x) log P(x) = H(P_e) + D(P_e\|P) \geq 0.$$

Powyższą miarę stosuje się do identyfikacji rozkładu teoretycznego. Dla ustalonych wartości rozkładu empirycznego P_e minimalizowana jest wartość:

$$H(P_e\|P) \text{ dla } P_e = P.$$

W pakiecie R (*R language*), w którym wykonywane będą obliczenia teoretyczne, entropia krzyżowa jest wyliczana w oparciu o algorytm *k* najbliższych sąsiadów kNN (*k nearest neighbours*) – kNN *Cross Entropy Estimators*. Algorytm obliczeń dostępny w języku R został opracowany na podstawie artykułu *kNN-based high-dimensional Kullback-Leibler distance for tracking* autorstwa S. Boltza, E. Debreuve'a i M. Barlauda [2007]. W artykule zaproponowano nowy estymator funkcji gęstości prawdopodobieństwa (*probability density function*, PDF) na podstawie algorytmu *k* najbliższych sąsiadów, który został wykorzystany do zdefiniowania spójnego estymatora entropii. Autorzy zaprezentowali sposób wyliczenia odległości Kullbacka-Leiblera pomiędzy wysokowymiarowymi funkcjami gęstości prawdopodobieństwa (*high-dimensional* PDF) z wykorzystaniem metody kNN. W tym kontekście jawna estymacja funkcji gęstości prawdopodobieństwa nie jest konieczna ze względu na fakt, że odległość jest obliczona na podstawie danych.

3.5. Metoda wskaźników pojemności informacji Hellwiga

Metodę wskaźników pojemności informacji przedstawił Zdzisław Hellwig w roku 1969 na łamach „Przeglądu Statystycznego" [Hellwig 1969]. W metodzie tej procedura wyboru optymalnego zbioru zmiennych objaśniających polega na tym, że dla każdej zmiennej z kombinacji wyznaczana jest indywidualna pojemność nośników informacji [Dziechciarz 2002, s. 33]:

$$h_{kj} = \frac{r_j^2}{\sum |r_{ij}|},$$

gdzie: k – numer kombinacji ($k = 1, 2, \ldots, 2^m-1$),
j – numer zmiennej w danej kombinacji,
r_j – współczynnik korelacji potencjalnej zmiennej objaśniającej o numerze j ze zmienną objaśnianą,
r_{ij} – współczynnik korelacji między i-tą i j-tą potencjalną zmienną objaśniającą.

Wskaźnik h_{kj} mierzy wielkość informacji, jaką zmienna X_j wnosi o zmiennej objaśnianej Y w k-tej kombinacji. Po wyznaczeniu wartości h_{kj} dla wszystkich zmiennych oblicza się pojemność integralną kombinacji nośników informacji. Wyznacza się ją dla każdej kombinacji według wzoru:

$$H_k = \sum h_{kj} .$$

Jest ona sumą indywidualnych pojemności nośników informacji, które wchodzą w skład danej kombinacji. Kryterium wyboru odpowiedniej kombinacji zmiennych objaśniających stanowi pojemność informacji. Według tego kryterium należy wybrać kombinację zmiennych, która wnosi najwięcej informacji, czyli taką, dla której H_k osiąga najwyższą wartość.

W pracy D. Serwy *Metoda Hellwiga jako kryterium doboru zmiennych do modeli szeregów czasowych* [Serwa 2004, s. 5-17] przedstawiono analizę, której celem było rozstrzygnięcie, na ile metoda Hellwiga jest użyteczna w odniesieniu do konstruowania modeli szeregów czasowych i w jakim zakresie jest ona konkurencyjna wobec innych metod, na przykład opartych na kryterium informacyjnym Akaike czy Schwarza. Przeprowadzone tam badania wykazały, że metoda Hellwiga w pewnych przypadkach, nie prowadzi do wyboru modelu oryginalnego. Z przeprowadzonego w przywołanej pracy badania wynika również, iż przy analizie szeregów czasowych metoda Hellwiga nie zawsze pozwala na automatyczny wybór postulowanego modelu „idealnego" lub też model idealny nie zawsze jest równoważny modelowi oryginalnemu. W pracy D. Serwy dokonano analizy metody Hellwiga jako kryterium doboru zmiennych do modeli autoregresji. Zauważono, że przy spełnieniu pewnych prostych warunków metoda Hellwiga sugeruje wybór modelu autoregresji nieodpo-

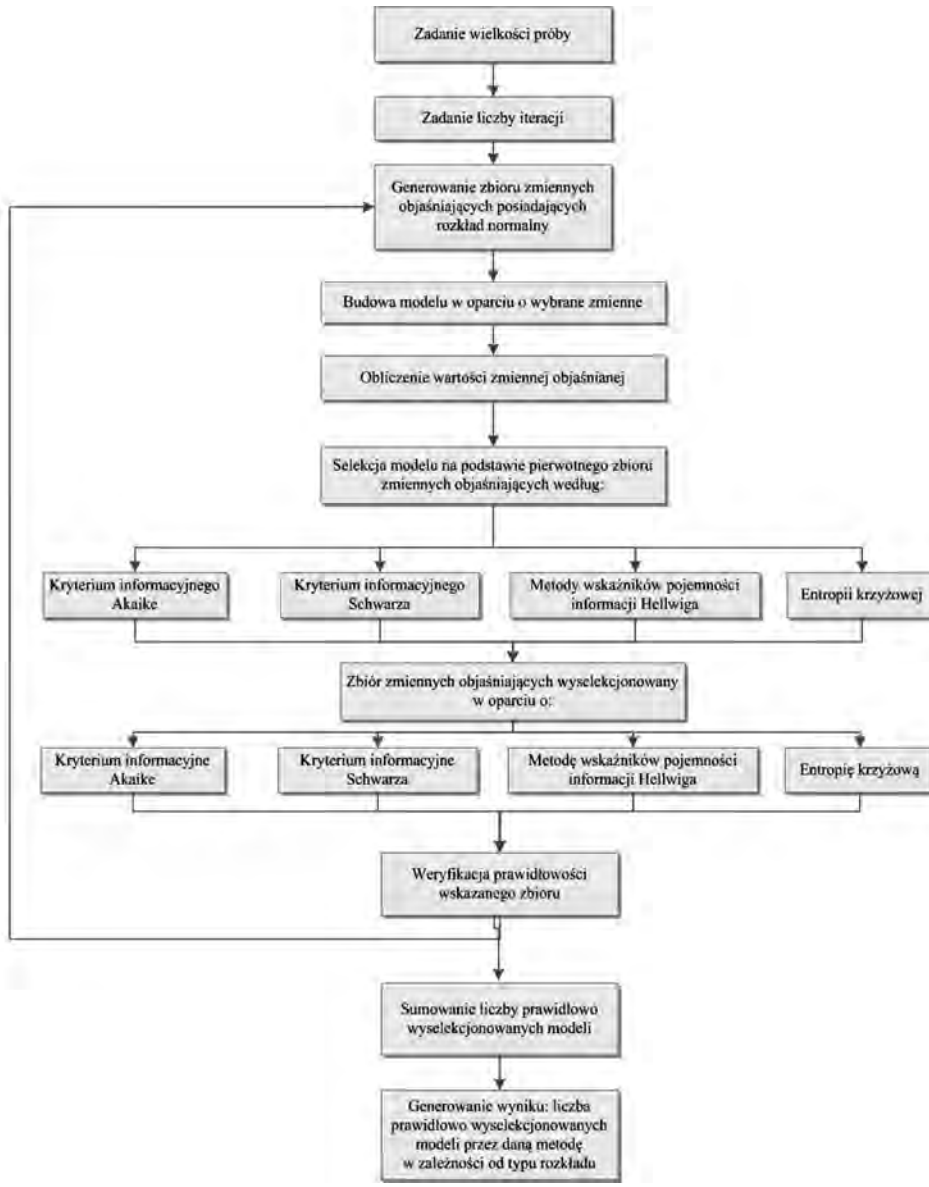
wiedniego rzędu. Wynik ten pozwala na uogólnienie na przypadek innych liniowych modeli szeregów czasowych, zwłaszcza modeli dynamicznych. W modelach tych często występują wzajemnie skorelowane zmienne objaśniające. W wyniku zastosowania metody optymalnego doboru zmiennych objaśniających część z nich może zostać usunięta z grupy kandydatek, pomimo że faktycznie wpływają istotnie na zmienną objaśnianą. Pominięcie ważnych zmiennych objaśniających w modelu może spowodować, że model zostanie odrzucony w procesie weryfikacji z powodu: niestabilności postaci strukturalnej, obciążonych ocen parametrów strukturalnych, autokorelacji składnika losowego bądź występowania dużych błędów prognozy. Tezę tę potwierdzają badania T. Bednarskiego i F. Borowicza, przeprowadzone w pracy *On inconsistency of Hellwig's variable choice method in regression models* [2009, s. 41-51].

4. Badania symulacyjne

4.1. Charakterystyka programu

W celu weryfikacji, która z wybranych metod – kryterium Akaike, kryterium Schwarza, metoda entropii krzyżowej, metoda Hellwiga – jest najefektywniejszym narzędziem doboru zmiennych objaśniających do modelu, opracowano program w środowisku R.

Koncepcję funkcjonowania programu przedstawia algorytm zamieszczony na rys. 1. Zgodnie z algorytmem po uruchomieniu programu w środowisku R należy podać wielkość próby *n* oraz liczbę powtórzeń *i*. Wielkość próby informuje o liczności próby każdej z wygenerowanych zmiennych objaśniających, a w konsekwencji także zmiennej objaśnianej. Z kolei liczba powtórzeń stanowi o tym, ile razy zostanie wykonana pętla obliczeń. Po zadaniu wielkości próby i liczby powtórzeń następuje wygenerowanie zbioru *Z*, zawierającego $m + 1$ potencjalnych zmiennych objaśniających $(x_1, x_2, ..., x_m, e)$, gdzie *e* pełni w modelu rolę składnika losowego), charakteryzujących się rozkładem normalnym. W kolejnym kroku, w oparciu o podzbiór zbioru Z, na który składają się wybrane zmienne objaśniające $(x_1, ..., x_k)$, gdzie $k < m$, zostaje zbudowany model. W następnym etapie dokonywana jest selekcja zmiennych objaśniających według czterech metod: kryterium Akaike, kryterium Schwarza, metody entropii krzyżowej i metody Hellwiga. Następnie zostaje porównane, czy dana metoda wskazała zmienne tożsame z tymi, które wchodzą w skład modelu. Jeśli tak, wówczas zostaje nadana waga 1, jeśli nie – wówczas 0. Ta procedura zostaje powtórzona *i* razy. W ostatnim kroku, po wykonaniu zadanej liczby pętli, następuje sumowanie liczby prawidłowo wyselekcjonowanych modeli (liczby prawidłowo wyselekcjonowanych zbiorów zmiennych objaśniających). Pojawia się wówczas komunikat zawierający informację o liczbie modeli prawidłowo wskazanych przez każdą z metod. Należy dodać, że po niewielkiej modyfikacji program może zostać



Rys. 1. Algorytm prezentujący sposób funkcjonowania programu

Źródło: opracowanie własne.

wykorzystany do badania innych rozkładów prawdopodobieństwa (np. wykładniczego, Weibulla, logarytmiczno-normalnego), a także modeli liniowych i nieliniowych sprowadzalnych do liniowych o dowolnej strukturze. Jedynym ograniczeniem programu jest pojemność pamięci obliczeniowej.

4.2. Analiza otrzymanych wyników

Badania symulacyjne w programie opisanym powyżej wykonano dla czterech modeli, których struktura przedstawiona została w tab. 1:

Tabela 1. Typy modeli stosowanych w programie

Typ modelu	Liczba zmiennych w modelu	Liczba potencjalnych zmiennych	Postać modelu
I	1	3	$\hat{y} = 0{,}8x_1 + 0{,}5e$
II	3	4	$\hat{y} = x_1 + 0{,}5x_2 + 0{,}1x_4 + 0{,}5e$
III	4	7	$\hat{y} = x_1 + 0{,}5x_3 + 0{,}1x_4 + 0{,}7x_6 + 0{,}5e$
IV	6	10	$\hat{y} = x_1 + 0{,}5x_4 + 0{,}1x_5 + 0{,}7x_7 + 0{,}4x_8 + 0{,}9x_9 + 0{,}5e$

Źródło: opracowanie własne.

Plan przeprowadzania badań symulacyjnych został zaprojektowany zgodnie z tab. 2. Dla każdego typu modelu (I-IV) wykonano 10 symulacji. Warunki wstępne (wielkość próby, liczba powtórzeń) każdej z nich zostały przedstawione w tabeli.

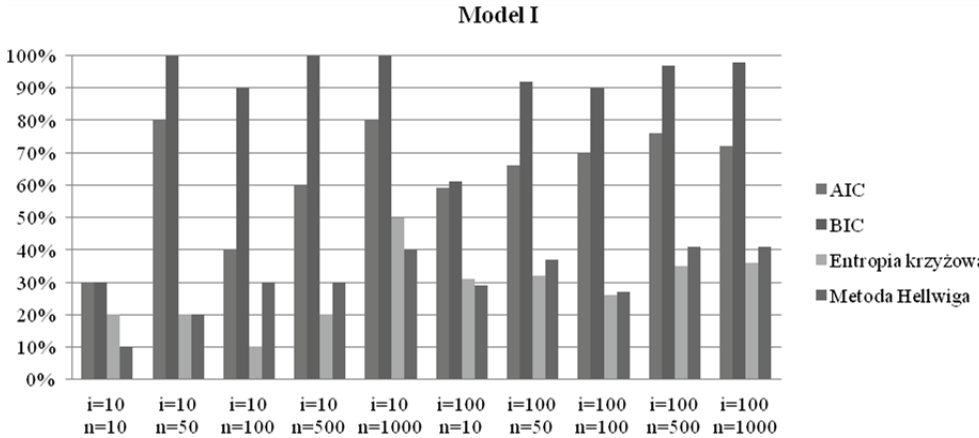
Tabela 2. Struktura przeprowadzania badań symulacyjnych

Numer symulacji	1	2	3	4	5	6	7	8	9	10
Wielkość próby (n)	10	50	100	500	1000	10	50	100	500	1000
Liczba powtórzeń (i)	10	10	10	10	10	100	100	100	100	100

Źródło: opracowanie własne.

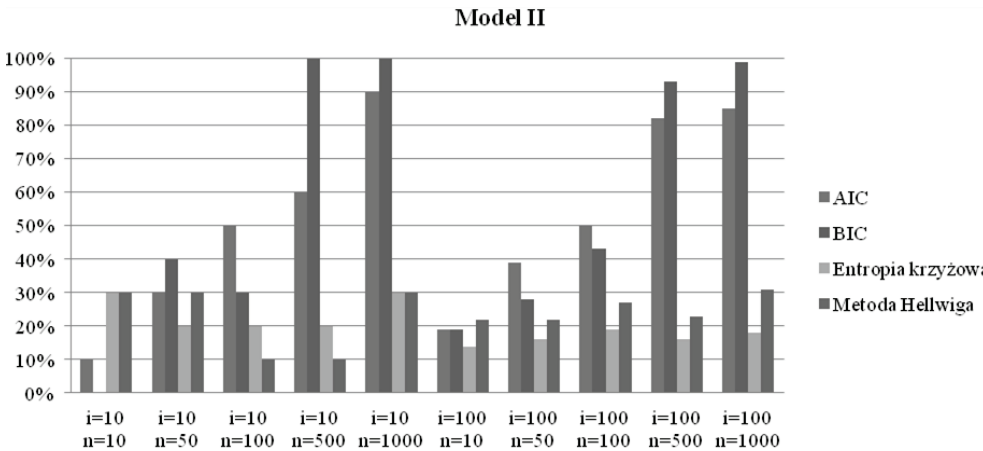
Wyniki otrzymane dla pierwszego modelu (model I z 1 zmienną objaśniającą, 3 potencjalne zmienne objaśniające) zostały przedstawione na rys. 2. Otrzymane wyniki wskazują, że dla modelu z jedną zmienną objaśniającą za najbardziej efektywną metodę selekcji zmiennych można uznać kryterium informacyjne Schwarza. Analizując wyniki dla symulacji, w których wielkość próby wynosiła co najmniej 50, można zauważyć, że za pomocą kryterium informacyjnego Schwarza właściwa postać modelu była wskazywana z częstotliwością wynoszącą co najmniej 90%. Na 10 symulacji 8 razy kryterium informacyjne Akaike wskazywało właściwą postać modelu z prawdopodobieństwem zawierającym się w przedziale 0,6-0,8. Z kolei pozostałe dwie badane metody – entropii krzyżowej i Hellwiga – były efektywne w mniej niż 40% w 9 na 10 symulacji.

Wyniki otrzymane dla drugiego modelu (model II) zostały zamieszczone na rys. 3. Można zauważyć, że w przypadku modelu, w którym zbiór potencjalnych determinant liczy 4 zmienne, a w modelu znajdują się 3 z nich, dla symulacji, w których liczność próby wynosiła $n = 500$ i $n = 1000$, najskuteczniejszą metodą w kontekście



Rys. 2. Wyniki badań dla modelu I

Źródło: opracowanie własne.



Rys. 3. Wyniki badań dla modelu II

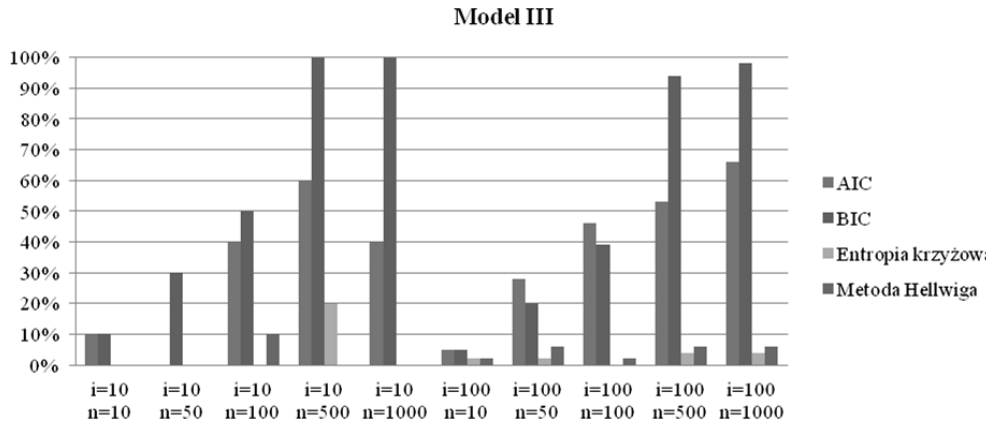
Źródło: opracowanie własne.

wyboru prawidłowego podzbioru zmiennych okazało się kryterium informacyjne Schwarza. Z kolei dla mniej licznych prób bardziej efektywnym narzędziem wyboru modelu okazało się kryterium informacyjne Akaike, należy jednak zaznaczyć, że efektywność ta nie przekroczyła 60%. W przypadku pozostałych metod częstotliwość wyboru prawidłowych zmiennych mających wejść do modelu nie przekroczyła 30%.

Analizując wyniki otrzymane dla trzeciego modelu (rys. 4), składającego się z 4 zmiennych (7-elementowy zbiór potencjalnych zmiennych objaśniających), można

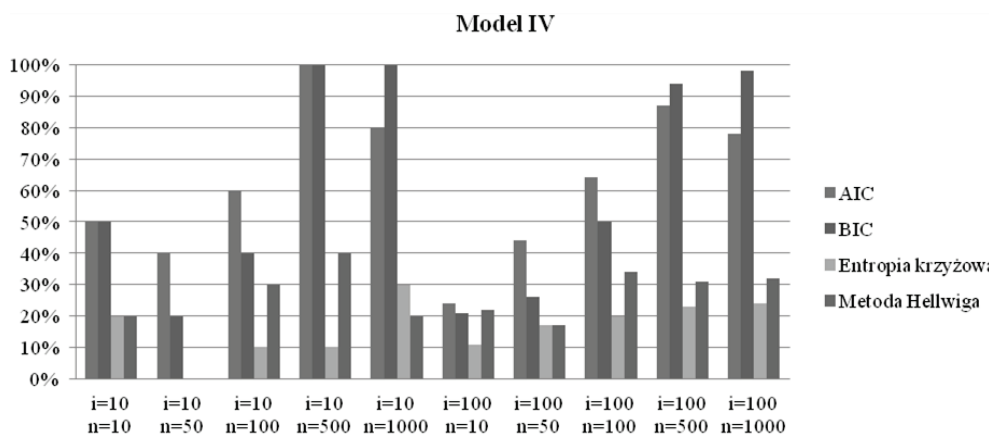
zaobserwować, że ponownie dla symulacji, w których liczność próby wynosiła $n = 500$ i $n = 1000$ za najskuteczniejszą metodę należy uznać kryterium Schwarza. Przy liczbie iteracji równej 100, dla mniejszych prób (50, 100) najbardziej efektywne w porównaniu z pozostałymi metodami okazało się kryterium Akaike. Jednak częstotliwość z jaką wskazywane były prawidłowe zestawy zmiennych była stosunkowo niska (28%-46%). Skuteczność pozostałych metod w 9 na 10 symulacji nie przekroczyła 10%.

Na rysunku 5 przedstawiono wyniki symulacji przeprowadzonych dla czwartego modelu (model IV) z sześcioma zmiennymi. W zbiorze wejściowym znajdowało się



Rys. 4. Wyniki badań dla modelu III

Źródło: opracowanie własne.

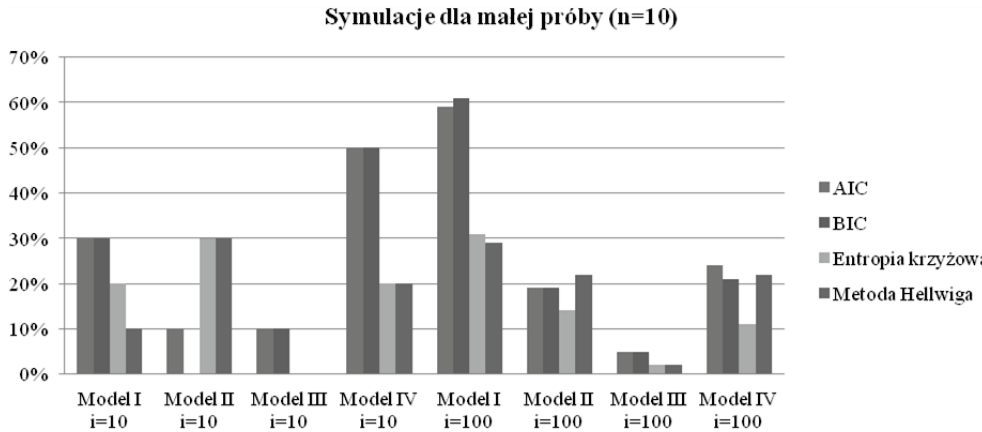


Rys. 5. Wyniki badań dla modelu IV

Źródło: opracowanie własne.

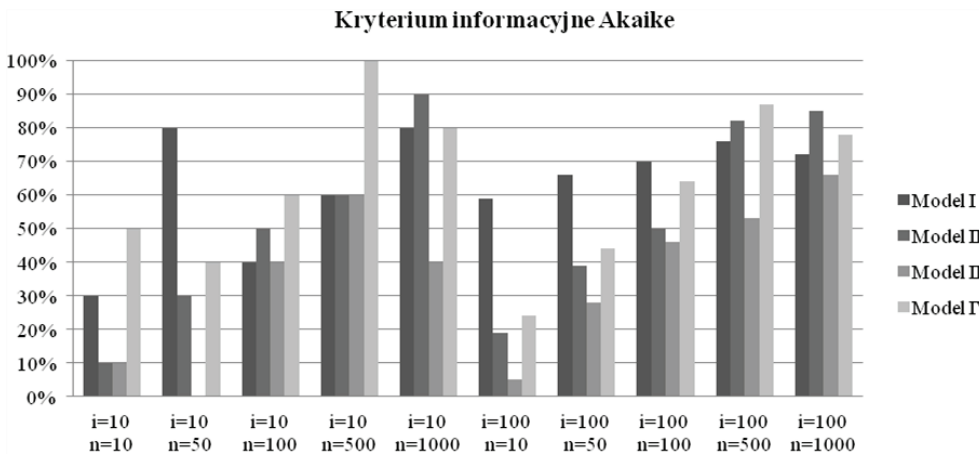
10 potencjalnych zmiennych objaśniających. W przypadku symulacji, w których liczba iteracji wynosiła 10, a liczność próby nie przekroczyła 100, bardziej skuteczną metodą okazało się kryterium AIC. Z kolei przy większej licznie iteracji dla prób o liczności większej lub równej 500 najczęściej prawidłowy podzbiór zmiennych wskazywało kryterium BIC, dla mniejszych prób natomiast bardziej skuteczne było kryterium AIC. Metoda Hellwiga w porównaniu z metodą entropii krzyżowej okazała się bardziej skuteczna, jednak jej efektywność nie przekroczyła 40%.

Na rysunku 6 zestawiono wyniki symulacji, w których badano efektywność metod w przypadku bardzo małej próby ($n = 10$). Niezależnie od typu modelu skutecz-



Rys. 6. Zestawienie wyników symulacji wykonanych dla małej próby

Źródło: opracowanie własne.

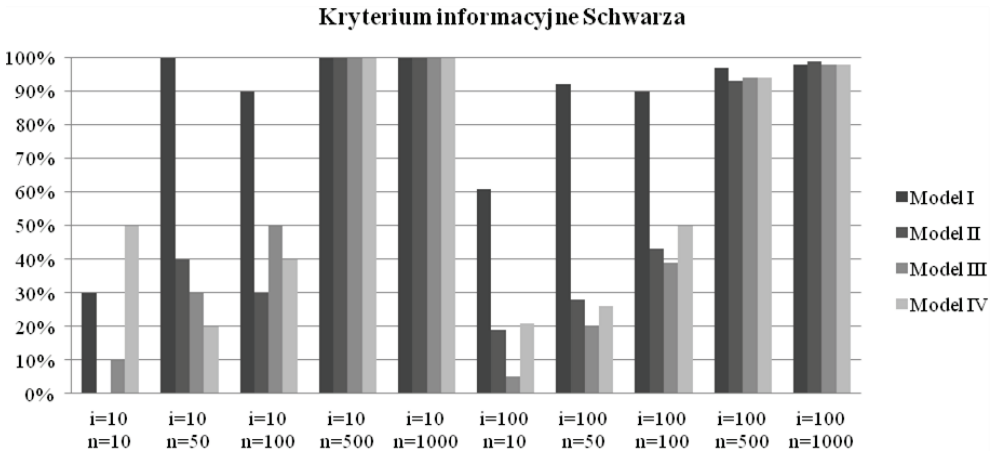


Rys. 7. Kryterium informacyjne Akaike – zestawienie wyników

Źródło: opracowanie własne.

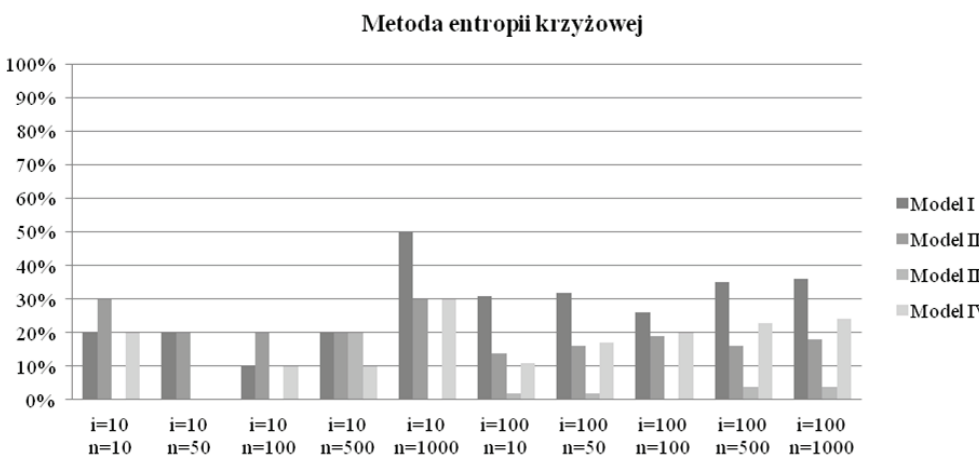
ność każdej z metod jest bardzo niska. Jedynie w przypadku symulacji dla modelu I kryterium AIC i BIC wskazały prawidłowy zbiór zmiennych ze skutecznością przekraczającą 50% (odpowiednio 59% i 61%).

Wykonano łącznie 32 symulacje dla małej próby. W 16 przypadkach na 32 niezależnie od rodzaju metody i typu modelu liczba prawidłowo wskazanych zmiennych nie przewyższyła 20%. Na rysunkach 7-10 zestawiono wyniki dla każdej z badanych metod osobno. Po analizie wyników można jednoznacznie stwierdzić, że kryterium BIC jest najskuteczniejszą metodą selekcji zmiennych wśród czterech ba-



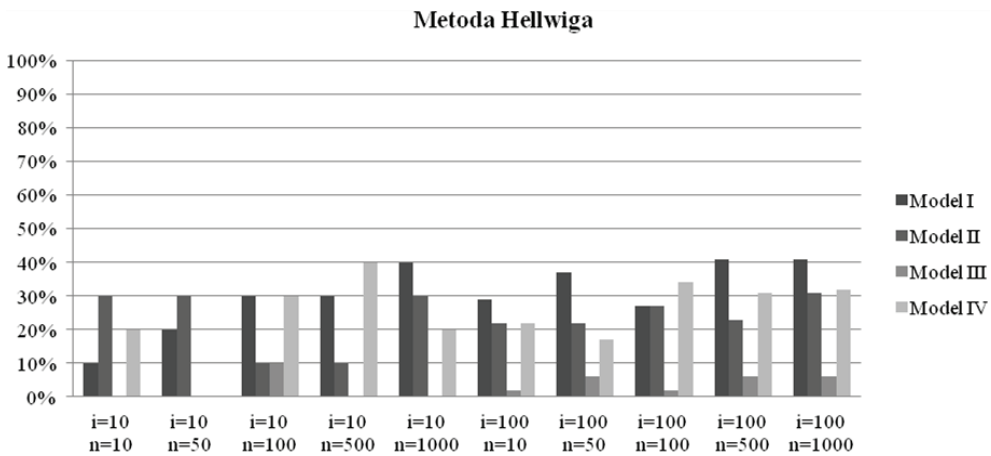
Rys. 8. Kryterium informacyjne Schwarza – zestawienie wyników

Źródło: opracowanie własne.



Rys. 9. Metoda entropii krzyżowej – zestawienie wyników

Źródło: opracowanie własne.



Rys. 10. Metoda Hellwiga – zestawienie wyników

Źródło: opracowanie własne.

danych. Jest wysoce efektywne przede wszystkim w przypadku bardzo licznych prób (n = 500 i n = 1000). Wówczas prawdopodobieństwo wskazania przez to kryterium optymalnego zbioru zmiennych mających wejść do modelu jest bliskie 100%. Kryterium Akaike dla licznych prób sprawdza się z częstotliwością przekraczająco niemal zawsze 60%. Jednak nie jest to metoda tak skuteczna, jak kryterium BIC. Z kolei dla prób mniej licznych skuteczność kryteriów Akaike i Schwarza jest podobna. Natomiast efektywność pozostałych dwóch metod poddanych analizie jest niska (nie przekracza 50%) niezależnie od wielkości próby i struktury modelu.

5. Przykład zastosowania metod selekcji zmiennych w inżynierii produkcji – analiza danych empirycznych

W dalszej części badań porównano efektywność metod na podstawie analizy danych empirycznych. W tym celu wykorzystano wybrane charakterystyki towarzyszące pewnemu procesowi z obszaru inżynierii produkcji. Mianowicie poddano analizie problem kompensacji błędów obróbki na poziomie sterowania. Dane, którymi posłużono się w obliczeniach, pochodzą z pomiarów przeprowadzonych w pewnym centrum tokarskim. Ze względu na fakt, że nowoczesnym centrom obróbkowym stawiane są co raz wyższe wymagania, zachodzi konieczność zwiększania dokładności wykonania przedmiotu obrabianego. W tym celu konieczne jest opracowanie sposobu precyzyjnej kompensacji błędów obróbki. Jest to zagadnienie bardzo obszerne, dlatego w niniejszej pracy ograniczono się do wycinkowego przedstawienia problemu. Wykorzystanie omawianego problemu w niniejszej pracy służy jedynie jako przykład możliwości wykorzystania metod selekcji zmiennych w inżynierii produkcji.

Rozwiązanie problemu polegającego na określeniu wpływu temperatury na całkowite przemieszczenie wrzeciona jest kluczowym aspektem umożliwiającym osiągnięcie odpowiedniej dokładności obróbki. Gdyby udało się precyzyjnie określić, w jakim stopniu temperatura oddziałuje na przemieszczenie wrzeciona, można by to przemieszczenie uwzględnić przy programowaniu danej obrabiarki, dzięki czemu wpływ temperatury powodujący odchylenie wrzeciona zostałby wyeliminowany, a założona dokładność obróbki – osiągnięta. Istnieje wiele czynników powodujących wzrost temperatury obrabiarki. Można do nich zaliczyć między innymi temperaturę panującą na hali produkcyjnej, straty mocy w silniku, straty mocy w napędach posuwu, straty mocy w łożyskach, typ oleju, lepkość oleju, wielkość i typ łożyska wpływające bezpośrednio na temperaturę w łożysku. Wpływ temperatury na elementy wykonane ze stali czy z żeliwa jest jak najbardziej wymierny – wzrost temperatury o 1º powoduje wydłużenie o około 11 μm. W celu określenia wpływu temperatury na przemieszczenie wrzeciona należy dokonać jej pomiaru. Pomiar ten jest wykonywany za pomocą specjalnych czujników, rozmieszczonych w kilku miejscach obrabiarki.

Celem prowadzonej analizy jest zbudowanie modelu, w którym zmienną objaśnianą jest błąd cieplny obróbki, a zbiór potencjalnych zmiennych objaśniających stanowią wartości odczytów temperaturowych mierzonych w różnych miejscach na obrabiarce. Model taki może zostać wykorzystany jako narzędzie do kompensacji błędu obróbki.

Aby porównać i przeanalizować efektywność badanych metod selekcji zmiennych, zostaną one wykorzystane jako kryteria wyboru temperatur pełniących w konstruowanym modelu rolę zmiennych objaśniających. Zbiór potencjalnych zmiennych objaśniających liczy 10 elementów, którymi są pomiary temperatur w 10 stosownie położonych punktach pomiarowych na obrabiarce (x_1, ..., x_{10}). Pomiarów temperatury dla każdego punktu i pomiarów błędu obróbki dokonano 50-krotnie (stąd próba równa $n = 50$). Zastosowanie efektywnej metody selekcji zmiennych pozwoli na wyznaczenie optymalnej liczby miejsc, w których należy dokonywać odczytów temperatury. Pomiary poziomu temperatur z odpowiednich miejsc umożliwią wyznaczenie cieplnego błędu obróbki. Znając prognozowaną wartość tego błędu, można poddać go kompensacji, uzyskując dzięki temu większą dokładność obróbki.

W celu zbudowania możliwie najlepszego modelu przeprowadzone zostaną obliczenia dla czterech wymienionych wyżej metod selekcji zmiennych, a następnie porównane zostaną otrzymane wyniki i na tej podstawie będzie można określić, która metoda jest najefektywniejszym narzędziem ich selekcji. Dla każdego z czterech zbiorów zmiennych wytypowanych przez poszczególne metody wykonano obliczenia metodą MNK w celu oszacowania parametrów modelu oraz wyznaczenia ocen ich precyzji. W kolejnym kroku porównano cztery powstałe w ten sposób modele pod kątem dopasowania do danych empirycznych. Model wyselekcjonowany przez kryterium informacyjne AIC, zbudowany został z 7 zmiennych objaśniających. Wyniki metody MNK przedstawiono w tab. 3:

Tabela 3. Wyniki MNK dla modelu opartego na zmiennych wyselekcjonowanych przez AIC

Błąd standardowy estymacji	0,0055			
Skorygowany R^2	0,8786			
	Wartość oszacowana	**Błąd st.**	***t*(45)**	**poziom *p***
Wyraz wolny	9,3645	10,4934	0,8920	0,377240
Zmienna x_1	–0,005	0,0009	–5,4200	0,000003
Zmienna x_4	–0,200	0,1290	–1,5500	0,128530
Zmienna x_5	0,096	0,0617	1,5610	0,126130
Zmienna x_6	–0,021	0,0091	–2,2580	0,029200
Zmienna x_8	0,017	0,0056	2,9490	0,005190
Zmienna x_9	–0,323	0,1839	–1,7550	0,086490
Zmienna x_{10}	0,403	0,1818	2,2180	0,032000

Źródło: opracowanie własne.

Postać analityczna modelu po oszacowaniu parametrów jest następująca:

$$\hat{y} = 9{,}365 - 0{,}005x_1 - 0{,}200x_4 + 0{,}096x_5 - 0{,}021x_6 \\ +0{,}017x_8 - 0{,}323x_9 + 0{,}403x_{10},$$

gdzie: $\hat{y}$ – oszacowana wartość błędu cieplnego obróbki.

Analizując wartości poziomu istotności *p*, należy sądzić, że jedynie 4 parametry modelu są istotne, a pozostałe należy wykluczyć z modelu (zmienne nieistotne oznaczono w tabeli szarą czcionką). Można zatem stwierdzić, że kryterium AIC nie wyselekcjonowało optymalnego zbioru zmiennych objaśniających. Otrzymany model nie może służyć jako narzędzie, które można by wykorzystać do kompensacji błędów obróbki.

Wyniki uzyskane po zastosowaniu kryterium BIC wskazują, że do modelu powinny wejść 4 zmienne. Wyniki metody MNK przedstawiono w tab. 4:

Tabela 4. Wyniki MNK dla modelu opartego na zmiennych wyselekcjonowanych przez BIC

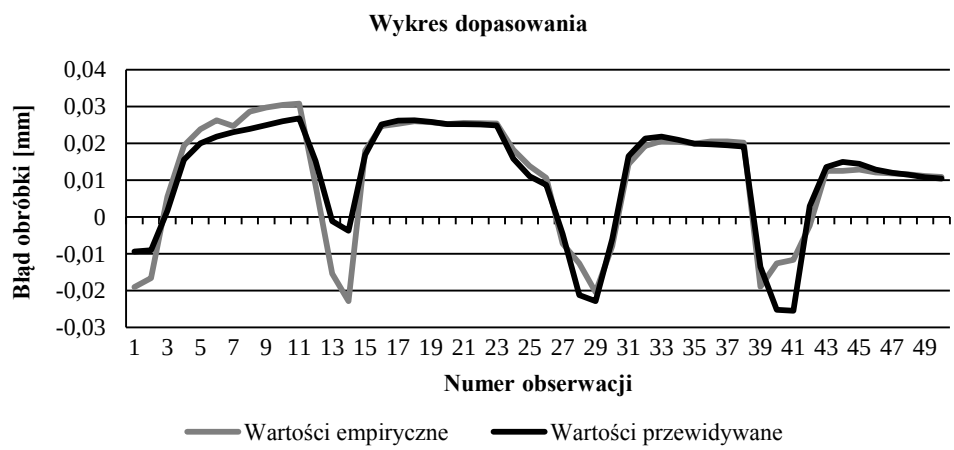
Błąd standardowy estymacji	0,0057			
Skorygowany R^2	0,8722			
	Wartość oszacowana	**Błąd st.**	***t*(45)**	**poziom *p***
Wyraz wolny	–3,838	1,5981	–2,401	0,0205200
Zmienna x_1	–0,004	0,0003	–11,477	0,0000000
Zmienna x_6	–0,010	0,0019	–5,309	0,0000033
Zmienna x_8	0,010	0,0015	6,741	0,0000000
Zmienna x_{10}	0,017	0,0055	3,051	0,0038200

Źródło: opracowanie własne.

Postać analityczna modelu po oszacowaniu parametrów jest następująca:

$$\hat{y} = -3{,}838 - 0{,}004x_1 - 0{,}010x_6 + 0{,}010x_8 + 0{,}017x_{10}.$$

Po przeanalizowanie otrzymanych wyników można stwierdzić, że wszystkie parametry modelu są statystycznie istotne, o czym świadczą wartości poziomu istotności *p*. Standardowe błędy szacunku nie są bardzo wysokie, co oznacza, że oszacowane wartości parametrów nie różnią się znacznie od ich rzeczywistej wartości. Standardowy błąd estymacji świadczy o tym, że oszacowana wartość błędu obróbki różni się średnio o 0,0057 mm od swojej rzeczywistej wartości. Skorygowany współczynnik determinacji równy 0,8722 świadczy o dość dobrym dopasowaniu modelu do danych empirycznych, w ponad 87,2% wyjaśnia on zmienność błędu obróbki. Interpretacja parametrów modelu: jeżeli temperatura na czujniku 1 (x_1) wzrośnie o 1 ºK, to można sądzić, że błąd obróbki zmniejszy się średnio o 0,004 mm, przy założeniu, iż pozostałe zmienne objaśniające będą miały stałą wartość. Analogicznie – jeżeli temperatura na czujniku 6 (x_6), czujniku 8 (x_8), czujniku 10 (x_{10}), wzrośnie o 1 ºK, to można sądzić, że błąd obróbki zmieni się średnio odpowiednio o: –0,010 mm, 0,010 mm, 0,017 mm), przy założeniu, iż pozostałe zmienne objaśniające będą miały stałą wartość. Do oceny dopasowania modelu do danych empirycznych posłużono się również wykresem dopasowania przedstawionym na rys. 11.



Rys. 11. Wykres dopasowania dla modelu opartego na zmiennych wyselekcjonowanych przez BIC

Źródło: opracowanie własne.

Wykres ten świadczy o dość dobrym dopasowaniu modelu błędu obróbki do danych empirycznych. Model wyselekcjonowany przez metodę entropii krzyżowej, zbudowany został z 5 zmiennych objaśniających. Wyniki metody MNK przedstawiono w tab. 5:

Tabela 5. Wyniki MNK dla modelu opartego na zmiennych wyselekcjonowanych przez m. entropii krzyżowej

Błąd standardowy estymacji	0,0070			
Skorygowany R^2	0,8086			
	Wartość oszacowana	**Błąd st.**	***t*(45)**	**poziom *p***
Wyraz wolny	19,646	9,8815	1,9880	0,053040
Zmienna x_2	–0,001	0,0004	–2,3450	0,023622
Zmienna x_3	0,071	0,0187	3,8190	0,000416
Zmienna x_7	–0,001	0,0005	–2,6950	0,009923
Zmienna x_9	–0,259	0,1844	–1,4030	0,167694
Zmienna x_{10}	0,123	0,1349	0,9100	0,367843

Źródło: opracowanie własne.

Postać analityczna modelu po oszacowaniu parametrów:

$$\hat{y} = 19{,}646 - 0{,}001x_2 + 0{,}071x_3 - 0{,}001x_7 - 0{,}259x_9 + 0{,}123x_{10}.$$

Analizując wartości poziomu istotności *p*, należy sądzić, że jedynie 3 parametry modelu są istotne, a pozostałe należy wykluczyć z modelu. Można zatem stwierdzić, że metoda entropii krzyżowej nie wyselekcjonowała optymalnego zbioru zmiennych objaśniających. Otrzymany model nie może służyć jako narzędzie do kompensacji błędów obróbki. Poniżej przedstawiono wyniki obliczeń przeprowadzonych dla ostatniej z badanych metod. Model wyselekcjonowany przez metodę Hellwiga, zbudowany został w oparciu 5 zmiennych objaśniających. Wyniki metody MNK przedstawiono w tab. 6.

Tabela 6. Wyniki MNK dla modelu opartego na zmiennych wyselekcjonowanych przez m. Hellwiga

Błąd standardowy estymacji	0,0057			
Skorygowany R^2	0,8715			
	Wartość oszacowana	**Błąd st.**	***t*(45)**	**poziom *p***
Wyraz wolny	–5,8863	2,2944	–2,5650	0,013800
Zmienna x_1	–0,1170	0,0557	–2,1010	0,041400
Zmienna x_2	0,1119	0,0534	2,0950	0,042000
Zmienna x_6	0,0008	0,0009	0,9140	0,365800
Zmienna x_7	–0,0002	0,0010	–0,1690	0,866800
Zmienna x_{10}	0,0244	0,0095	2,5850	0,013100

Źródło: opracowanie własne.

Postać analityczna modelu po oszacowaniu parametrów jest następująca:

$$\hat{y} = -5{,}8863052 - 0{,}1169418x_1 + 0{,}1119378x_2 + \\ +0{,}0008173x_6 - 0{,}0001652x_7 + 0{,}0244267x_{10}.$$

Po dokonaniu analizy wartości poziomu istotności *p* należy sądzić, że 2 parametry modelu nie są istotne i należy je wykluczyć z modelu. Można zatem stwierdzić, że metoda Hellwiga nie wyselekcjonowała optymalnego zbioru zmiennych objaśniających. Otrzymany model nie może służyć jako narzędzie do kompensacji błędów obróbki.

6. Podsumowanie

Podstawowym celem niniejszej pracy było porównanie kryterium informacyjnego Akaike, kryterium informacyjnego Schwarza, metody entropii krzyżowej i metody wskaźników pojemności informacji Hellwiga pod kątem efektywności konstrukcji modeli regresyjnych. Zadanie to zrealizowano dwojako. W pierwszej kolejności przeprowadzono symulacje w oparciu o program napisany w języku R, a ich celem było wskazanie, która z badanych metod więcej razy wskaże model prawdziwy. Druga analiza porównawcza przeprowadzona została w oparciu o dane empiryczne.

Analiza wyników otrzymanych w wyniku przeprowadzenia symulacji w środowisku R wskazuje, że kryterium informacyjne Schwarza (BIC) jest najskuteczniejszą metodą selekcji zmiennych w porównaniu z pozostałymi trzema badanymi metodami. Kryterium to jest najefektywniejsze w budowie modeli opartych na bardzo licznych próbach ($n = 500$ i $n = 1000$). Prawdopodobieństwo wyselekcjonowania przez to kryterium optymalnego zbioru zmiennych mających wejść do modelu jest wówczas bliskie 100%. Z kolei kryterium Akaike dla licznych prób sprawdza się z częstotliwością przekraczająco niemal zawsze 60%. Nie jest jednak tak efektywną metodą, jak kryterium Schwarza. W przypadku prób mniej licznych skuteczność obu kryteriów jest porównywalna. Przeprowadzone badania wykazały także, że efektywność metody entropii krzyżowej i metody wskaźników pojemności informacji Hellwiga niezależnie od wielkości próby i struktury modelu jest bardzo niska (nie przekracza 50%). Dodatkowym celem pracy była ocena możliwości zastosowania badanych metod w modelowaniu wybranych zależności zachodzących na różnych etapach procesu produkcyjnego. Oceny tej dokonano w oparciu o analizę problemu kompensacji błędów obróbki pewnego centrum tokarskiego. Otrzymane wyniki potwierdziły wnioski pochodzące z badań teoretycznych. Ponownie kryterium informacyjne Schwarza okazało się być najskuteczniejszym narzędziem selekcji zmiennych. Model zbudowany w oparciu o cztery zmienne wytypowane przez kryterium BIC jest dobrze dopasowany do danych empirycznych, a oszacowane wartości parametrów nie różnią się znacznie od ich rzeczywistej wartości. Każda z czterech badanych metod wskazała na inny zestaw zmiennych objaśniających, które powinny wejść do modelu. Jedynie w przypadku kryterium Schwarza parametry wszystkich zmiennych i wyrazu wolnego okazały się istotne. Zarówno kryterium informacyjne Akaike, metoda entropii krzyżowej, jak i metoda wskaźników pojemności informacji Hellwiga wyselekcjonowały zestawy zmiennych i tym samym struktury modeli, które nie mogą być wykorzystane jako narzędzie do kompensacji błędów obróbki.

Podsumowując, należy zatem stwierdzić, że zawarte w pracy analizy, zarówno te przeprowadzone na danych wygenerowanych sztucznie w wyniku symulacji, jak i te przeprowadzone w oparciu o dane empiryczne, dowodzą, że aby skutecznie rozwiązać problem doboru optymalnego zbioru zmiennych objaśniających, który pozwoli skonstruować model efektywny, należy posługiwać się w tym celu kryterium informacyjnym Schwarza, gdyż jest to metoda zdecydowanie bardziej skuteczna niż pozostałe badane w pracy metody. Jak zostało wykazane w badaniach teoretycznych, kryterium AIC również można stosować, ale należy się liczyć z faktem, że nie zawsze jest ono efektywne. Natomiast zdecydowanie niska skuteczność metody entropii krzyżowej i metody Hellwiga sugeruje, że należy zrezygnować ze stosowania tych metod jako efektywnych kryteriów selekcji zmiennych. Wnioski te zdają się potwierdzać wyniki badań D. Serwy [2004, s. 5-17] i badania *stricte* analityczne, które przedstawione zostały w pracy T. Bednarskiego i F. Borowicza [2009, s. 41-51].

Wyniki otrzymane po przeprowadzeniu symulacji oraz wyniki uzyskane dzięki analizie modeli empirycznych wskazały jednoznacznie, że kryteria informacyjne, a w szczególności kryterium Schwarza, są lepszym, skuteczniejszym i bardziej wiarygodnym narzędziem wyboru optymalnego zbioru zmiennych objaśniających i tym samym prawdziwego modelu ekonometrycznego niż metoda entropii krzyżowej czy zalecana w polskich podręcznikach akademickich do ekonometrii metoda Hellwiga.

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
DOLNY ŚLĄSK
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Literatura

Acquah de-Graft H., *Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship*, „Journal of Development and Agricultural Economics" 2010, Vol. 2(1), ss. 001-006.

Arnold T. W., *Uninformative Parameters and Model Selection Using Akaike's Information Criterion*, „Journal of Wildlife Management" 2010, 74(6):1175–1178; ss. 1175-1178.

Bednarski T., Borowicz F., *On inconsistency of Hellwig's variable choice method in regression models*, „Discussiones Mathematicae Probability and Statistics" 2009, 29, ss. 41-51.

Boltz S., Debreuve E., Barlaud M., *kNN-based high-dimensional Kullback-Leibler distance for tracking*, Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'07), IEEE 2007.

Cavanaugh J.E., Shumway R. H., *An Akaike information criterion for model selection in the presence of incomplete data*, „Journal of Statistical Planning and Inference" 1998, Volume 67, ss. 45-65.

Detyna J., *Maksimum entropii jako teoretyczne kryterium statystycznego opisu separacji materii granulowanej*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2007.

Dziechciarz J., *Ekonometria. Metody, przykłady, zadania*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2002.

Firkowicz S., *Statystyczne badanie wyrobów*, Wydawnictwa Naukowo-Techniczne, Warszawa 1970.

Grabiński T., Wydymus S., Zeliaś A., *Metody doboru zmiennych w modelach ekonometrycznych*, Państwowe Wydawnictwo Naukowe, Warszawa 1982.
Hellwig Z., *Problem optymalnego wyboru predykant*, „Przegląd statystyczny" 1969, R.XVI zeszyt 3-4, ss. 225-236.
Kukuła K., *Wprowadzenie do ekonometrii w przykładach i zadaniach*, Wydawnictwo Naukowe PWN, Warszawa 1999.
Mercik J., Szmigiel C., *Ekonometria*, Wyższa Szkoła Zarządzania i Finansów we Wrocławiu, Wrocław 2000.
Peracchi F., *Econometrics*, Jonh Wiley & Sons Ltd, Chichester, West Sussex 2001.
Radkowski S., *Podstawy bezpiecznej techniki*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2003.
Ramos D., Gonzalez-Rodriguez J., *Cross-entropy Analysis of the Information in Forensic Speaker Recognition*, Proceedings of IEEE Oddysey, January 2008.
Schwarz G., *Estimating the dimension of a model*, „The Annals of Statistics" 1978, Vol. 6, No. 2, ss. 461-464.
Serwa D., *Metoda Hellwiga jako kryterium doboru zmiennych do modeli szeregów czasowych,* Szkoła Główna Handlowa, Kolegium Analiz Ekonomicznych, Instytut Ekonometrii, Warszawa 2004, ss. 5-17.

EFFICIENCY ANALYSIS OF CHOSEN METHODS OF EXPLANATORY VARIABLES SELECTION WITHIN THE SCOPE OF REGRESSION MODEL CONSTRUCTION

Summary: The basic aim of this paper is to compare Akaike's information criterion and Schwarz's Bayesian information criterion (BIC), cross entropy and Hellwig's method within the scope of regression model construction efficiency. The study was based on computer simulations. After generating a dataset with normal distribution, a linear model (true model, which in reality is not known) was built. In the model a response variable is dependent on the chosen variables previously generated. Next, a set of potential explanatory variables was extended and the analyzed methods of model selection were applied. These steps were repeated. Subsequently it was compared how often each of the tested methods indicated the right set of variables, and thus the right model. The methods were compared also on the basis of empirical data.

Keywords: Akaike information criterion, Schwarz information criterion, Hellwig's method, cross entropy.