Warsaw University of Life Sciences – SGGW
Institute of Economics and Finance
Department of Econometrics and Statistics

QUANTITATIVE METHODS IN ECONOMICS

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH

Volume XXII, No. 1

Warsaw 2021

ISSN 2082-792X
e-ISSN 2543-8565

Warsaw 2021, Volume XXII, No. 1
The original version is the paper version
Journal homepage: qme.sggw.edu.pl
Published by Warsaw University of Life Sciences Press

QUANTITATIVE METHODS IN ECONOMICS
Volume XXII, No. 1, 2021

CONTENTS

Bieszk-Stolorz Beata, Dmytrów Krzysztof, Majewski Sebastian, Zbaraszewski Wojciech – Drzewa klasyfikacyjne w identyfikacji czynników różnicujących postrzeganie sąsiedztwa parków narodowych w Euroregionie Pomerania 1

Dudziński Marcin, Kaleta Joanna – An Application of the Interval Estimation for the At-Risk-of-Poverty Rate Assessment 14

Karwowski Waldemar – Zastosowanie biblioteki ML.NET do badań ekonomicznych 29

Orzechowski Arkadiusz – Pricing European Options in the Heston and the Double Heston Models 39

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXII/1, 2021, s. 1 – 13

DRZEWA KLASYFIKACYJNE W IDENTYFIKACJI CZYNNIKÓW RÓŻNICUJĄCYCH POSTRZEGANIE SĄSIEDZTWA PARKÓW NARODOWYCH W EUROREGIONIE POMERANIA

Beata Bieszk-Stolorz https://orcid.org/0000-0001-8086-9037
Krzysztof Dmytrów https://orcid.org/0000-0001-7657-6063
Sebastian Majewski https://orcid.org/0000-0003-3072-5718
Instytut Ekonomii i Finansów
Uniwersytet Szczeciński
e-mail: beata.bieszk-stolorz@usz.edu.pl; krzysztof.dmytrow@usz.edu.pl;
sebastian.majewski@usz.edu.pl
Wojciech Zbaraszewski https://orcid.org/0000-0002-1373-1895
Wydział Ekonomiczny
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
e-mail: wojciech.zbaraszewski@zut.edu.pl

Streszczenie: Celem badania jest ocenia różnic w postrzeganiu sąsiedztwa parków narodowych (PN) w Euroregionie Pomerania przez mieszkańców Polski i Niemiec. Do klasyfikacji mieszkańców na podstawie ich odpowiedzi na pytania ankietowe wykorzystano metodę drzew klasyfikacyjnych. Otrzymane wyniki świadczą, że Polacy i Niemcy różnią się pod względem postrzegania sąsiedztwa parków narodowych. Są to: źródła pozyskiwania informacji o PN, ocena znaczenia badanego PN dla turystyki w regionie, ocena liczby turystów w PN oraz ocena pracy zarządu PN. Wykazano również współzależność pomiędzy narodowością respondentów, a odpowiedziami różnicującymi badaną zbiorowość.

Słowa kluczowe: drzewa klasyfikacyjne, analiza korelacji, analiza akceptacji, Euroregion Pomerania

JEL classification: G11, G17

WSTĘP

Powstawanie parków narodowych i parków krajobrazowych związane jest często z prowadzeniem badań dotyczących relacji obszar chroniony – społeczność

https://doi.org/10.22630/MIBE.2021.22.1.1

lokalna. Badania takie prowadzone są na całym świecie, również w krajach tworzących Euroregion Pomerania. W Niemczech noszą one nazwę analizy akceptacji [Pöhlman,. Rall 2011; Stoll-Kleemann (red.) 2015; Job i in. 2019], a w Polsce znane są pod nazwą badań postaw społeczności lokalnych [Hibszer 2013; Zawilińska 2020]. Pomimo różnic w nazewnictwie problemy leżące u podstaw tych badań są w obu krajach bardzo podobne. W polskim i niemieckim społeczeństwie kwestie ochrony przyrody znajdują generalnie szerokie poparcie na szczeblu ogólnonarodowym. Jednak na szczeblu lokalnym konkretne obszary chronione często muszą się zmagać z brakiem akceptacji wśród lokalnej społeczności. Źródłem problemu zazwyczaj jest fakt, że są one tworzone ze względu na walory przyrodnicze. Wiąże się to z wprowadzaniem ograniczeń w dotychczasowym użytkowaniu gruntów. Cały systemem zakazów negatywnie wpływa na poziom akceptacji takiego obszaru wśród lokalnej społeczności. Przyczyn braku akceptacji nie można jednak sprowadzać do jednego czynnika. Jest to wypadkowa czynników indywidulanych, aspektów kulturowych, emocjonalnych, ale także czynników percepcyjnych i komunikacyjnych [Sieberath 2007]. Przejawem niskiego poziomu akceptacji mogą być m.in. demonstracje ludności, czego przykłady można znaleźć zarówno w Niemczech, jak i w Polsce. Brak akceptacji ze strony społeczności lokalnych w Polsce jest na tyle duży, że niweczy nawet powstające na szczeblu rządowym plany utworzenia kolejnych parków narodowych tj. Jurajskiego, Mazurskiego, Turnickiego. Warto jednak podkreślić, że większość parków narodowych na całym świecie napotyka na opór lokalnej społeczności [von Ruschkowski 2009; Schumacher, Job 2013].

Najogólniej, celem prowadzenia badań akceptacji obszaru chronionego jest poznanie opinii lokalnej społeczności o obszarze chronionym i odpowiedź na pytanie: jaki jest stosunek społeczności lokalnej do obszaru chronionego? Dopiero poznanie poziomu akceptacji i przyczyn takiego stanu umożliwia jego administracji lub innym odpowiedzialnym instytucjom i osobom odpowiednią reakcję i podjęcie stosownych działań. Jest to zatem zagadnienie obszerne, które analizuje się biorąc pod uwagę stawiane pytania badawcze, cel badania i zastosowane metody [Job i in. 2019]. W Polsce i w Niemczech zagadnienia akceptacji obszaru chronionego koncentrują się na praktyce rozwiązywania konfliktów [Matuszewska 2003; Królikowska 2007; von Ruschkowski, Mayer 2011].

Celem badania jest ocena różnic w postrzeganiu sąsiedztwa parków narodowych w Euroregionie Pomerania przez mieszkańców Polski i Niemiec. Analizowano dane ankietowe pozyskane w wyniku realizacji polsko-niemieckiego projektu o akronimie REGE[1].

[1] Badanie zostało przeprowadzone w ramach projektu INT107 pt.: „Współpraca transgraniczna między uczelniami i dużymi obszarami chronionymi w Euroregionie Pomerania (akronim REGE)". Projekt jest dofinansowany przez Unię Europejską ze

EUROREGION POMERANIA

Euroregion Pomerania to region utworzony w 1995 roku. Od 2014 roku w jego skład wchodzą w Niemczech: powiaty Vorpommern-Rügen, Vorpommern-Greifswald, Mecklenburgische Seenplatte, Uckermark, Barnim i Märkisch Oderland, a w Polsce województwo zachodniopomorskie. Organizacji Euroregionu przyświecała idea współpracy dla zrównoważonego rozwoju regionu oraz zbliżenia jego mieszkańców zamieszkujących tereny po obu stronach granicy. W ramach wspólnych działań postanowiono, że współpraca między podmiotami po stronie polskiej i niemieckiej będzie dotyczyć bardzo szerokiego wachlarza aktywności rozpoczynając od współpracy gospodarczej, poprzez kulturalną i społeczną a na ekologicznej kończąc. W ramach Euroregionu prowadzone są wspólne projekty dedykowane rozwojowi obszarów a w szczególności jego mieszkańcom.

Projekt o akronimie REGE jest realizowany w ramach Programu Współpracy Interreg VA Meklemburgia-Pomorze Przednie-Brandenburgia-Polska przez cztery uczelnie wyższe: dwie polskie i dwie niemieckie. Głównym celem projektu jest intensyfikacja polsko-niemieckiej współpracy transgranicznej szkół wyższych z Euroregionu a także dziewięciu instytucji odpowiedzialnych za zarządzanie wielkopowierzchniowymi obszarami chronionymi. Zamysły wnioskodawców projektu skierowane były na poprawę współpracy między uczestniczącymi w projekcie uniwersytetami w obszarze badań nad społeczno-ekonomicznymi efektami odziaływania obszarów chronionych na lokalne społeczności. Owa poprawa współpracy miała dotyczyć również wymiany know-how między jednostkami odpowiedzialnymi za obszary chronione po stronie polskiej i niemieckiej. Zespoły badawcze zorganizowane w ramach projektu podjęły się zadania wypracowania uproszczonej metody szacowania efektów ekonomicznych z turystyki realizowanej na obszarach chronionych. Badania naukowe objęły między innymi diagnozę stanu obecnego, którą umożliwiła analiza sąsiedztwa (akceptacji). W dalszych etapach projektu zostaną przeprowadzone kolejne badania, wśród których znajdą się analiza satysfakcji skierowana do turystów oraz badanie przedsiębiorców działających w obszarze oddziaływania obszaru chronionego.

PRZEGLĄD LITERATURY

W literaturze pojawił się nurt opisujący badania związane z analizą akceptacji. Badania takie prowadzone są na wszystkich kontynentach, w krajach o różnym poziomie rozwoju społeczno-gospodarczego. Świadczy to o dużej wadze problemu i o potrzebie współpracy mieszkańców, administracji państwowej, samorządów oraz przedsiębiorców.

środków Europejskiego Funduszu Rozwoju Regionalnego (EFRR) w ramach Programu Współpracy Interreg VA Meklemburgia-Pomorze Przednie/Brandenburgia/Polska.

Na sposób postrzegania obszarów chronionych wpływa poziom wykształcenia, przekonania i sytuacja zawodowa. Obostrzenia w korzystaniu z zasobów naturalnych na obszarze parków narodowych są często sprzeczne z tradycyjną działalnością gospodarczą miejscowej ludności. Rybacy [Seridi, Djebar 2017; Andries i in. 2021], rolnicy [Rao i in. 2003], koczowniczy pasterze [Mashizi, Sharafatmandrad 2020] podkreślają negatywne konsekwencje ekonomiczne zarówno utworzenia obszaru chronionego, jak i nadmiernego rozwoju turystyki. Swoją opinię uzasadniają przede wszystkim wzrostem kosztów utrzymania oraz pogłębieniem się nierówności w dystrybucji korzyści. Pogłębiają się konflikty pomiędzy mieszkańcami reprezentującymi różne grupy zawodowe [Sirakaya i in. 2002; Andereck i in. 2005]. Pojawia się pytanie, kto będzie czerpał największe zyski z turystyki na obszarach chronionych. Mieszkańcy nie widzą siebie, jako rzeczywistych beneficjentów tych korzyści. Uważają, że zyski będą czerpały przedsiębiorstwa zewnętrzne, a mieszkańcy będą musieli radzić sobie z narzuconymi ograniczeniami [Aguirre 2006]. Aby połączyć wszystkie aspekty współpracy istotne jest wdrażanie na obszarach chronionych koncepcji turystyki zrównoważonej [Plummer, Fennell 2009].

Badania wskazują na to, że sukces w tworzeniu obszarów chronionych jest ściśle związany ze stopniem zaangażowania społeczności lokalnej w początkowy proces jego tworzenia. Brak w tym względzie konsensusu społecznego prowadzi do porażki [Pollnac i in. 2001]. Społeczna akceptacja jest czynnikiem, który trzeba brać pod uwagę na każdym etapie tworzenia obszaru chronionego. Szczególnie groźne są istniejące poważne konflikty pomiędzy społecznością lokalną, różnymi użytkownikami środowiska i zarządzającym obszarami chronionymi [Hargreaves-Allen i in. 2011]. Lekceważenie opinii przeciwników utworzenia obszarów chronionych może zaszkodzić przyszłym relacjom z lokalnymi społecznościami i utrudnić powodzenie projektu [Voyer i in. 2012]. Na skłonność do współpracy mieszkańców z zarządami parków wpływa również forma zarządzania. Badania Ayivor i in. [2020] wskazują na to, że preferowane są w tym względzie władze lokalne i samorządowe, gdyż do nich mieszkańcy mają zazwyczaj większe zaufanie. Najwięcej kontrowersji wzbudza centralny, państwowy system zarządzania obszarami chronionymi. Nie zapewnia on lokalnych potrzeb bytowych mieszkańców, nie rozwiązuje problemu ubóstwa, ignoruje podstawowe problemy społeczne, podsyca antagonistyczne relacje między społecznościami a urzędnikami oraz powoduje wzrost zjawisk kryminogennych. Na kształtowanie się akceptacji społecznej obszarów chronionych wpływają: edukacja ekologiczna, prawidłowa identyfikacja potrzeb społecznych mieszkańcow oraz ustalenie wpływu kulturowego na miejscową ludność, a także zadbanie o potrzeby osobiste mieszkańców [Leitinger i in. 2010]. Aby zwiększyć poziom akceptacji obszaru chronionego konieczne są działania wpływające na zwiększenie korzyści czerpanych przez miejscową ludność [Grainger 2003]. Korzyści ekonomiczne odgrywają ważną rolę obok kwestii psychologicznych, socjologicznych i politycznych [Hamin 2002].

METODYKA BADANIA

W celu poznania subiektywnych opinii mieszkańców na temat obszarów chronionych znajdujących się w ich bezpośrednim otoczeniu zostało przeprowadzone badanie ankietowe dotyczące akceptacji funkcjonowania obszarów chronionych. Przebadano mieszkańców okolic sześciu parków narodowych położonych w Euroregionie Pomerania. Badania ankietowe prowadzono w latach 2019-2020 metodą CATI wśród 2357 mieszkańców okolic trzech polskich parków narodowych: Drawieńskiego Parku Narodowego, Parku Narodowego „Ujście Warty", Wolińskiego Parku Narodowego i trzech niemieckich parków narodowych: Nationalpark Jasmund, Nationalpark Vorpommersche Boddenlandschaft i Nationalpark Unteres Odertal. W każdym z parków badaniem objęto od 385 do 403 respondentów, co zagwarantowało poziom istotności 0,05 i względną precyzję szacunku na poziomie 5%. Kwestionariusz składał się z 25 pytań, podzielonych na trzy części: wprowadzenie, pytania dotyczące analizowanego obszaru oraz pytania społeczno-demograficzne. W celu oceny stosunku mieszkańców do badanych parków analizowano drugą część kwestionariusza (część B).

Celem artykułu było zbadanie czy istnieją różnice pomiędzy odpowiedziami mieszkańców okolic polskich i niemieckich parków narodowych. Jeśli takie różnice wystąpiły, to określono, które pytania w największym stopniu determinowały te różnice. W związku z tym dokonano próby klasyfikacji narodowości respondentów na podstawie udzielonych przez nich odpowiedzi.

Ponieważ odpowiedzi na pytania ankietowe mierzone są na słabych skalach (nominalnej lub porządkowej), dlatego wykorzystano nieparametryczną metodę klasyfikacji – drzewo klasyfikacyjne. Budowa każdego drzewa wiąże się z dwoma aspektami [Capelli, Zhang 2007]: z podziałem danych, czyli rozrostem drzewa oraz z przycinaniem drzewa, aby zmniejszyć jego rozmiar i zwiększyć czytelność wyników. Kryterium podziału drzewa może być oparte o „zanieczyszczenie" Giniego. Podział następuje do momentu, aż zostanie osiągnięte kryterium zatrzymania. Konieczność przycinania drzewa wiąże się z tym, że podział danych powoduje, iż drzewo rozrasta się do bardzo dużych rozmiarów i następuje jego przeuczenie. Generalnie, podział przestaje następować, jeżeli zysk informacyjny z kolejnego podziału jest mniejszy, niż określony próg. Innym kryterium jest ustalenie minimalnej liczebności grupy, przy której może nastąpić podział, minimalnej liczebności grupy w węźle końcowym oraz maksymalną głębokość drzewa [Mudunuru 2016].

Obliczenia przeprowadzono w języku **R** przy zastosowaniu pakietów: `caret`, `pROC`, `rattle` oraz `lsr`. Analizę empiryczą przeprowadzono w czterech etapach:

- uczenie modelu drzewa klasyfikacyjnego,
- predykcja na podstawie modelu,

- ocena jakości predykcji,
- ocena współzależności zmiennych wpływających na podział drzewa.

Najpierw podzielono losowo zbiór danych na zbiór uczący (70% obserwacji) i zbiór testowy (pozostałe 30% obserwacji). Uczenia modelu dokonano za pomocą funkcji `train` z pakietu `caret`. Jako miarę jakości klasyfikacji wybrano `ROC`, czyli pole pod krzywą ROC.

Predykcja polega na sklasyfikowaniu ankietowanej osoby do odpowiedniego państwa (Polska lub Niemcy) na podstawie udzielanych przez nią odpowiedzi. Założono, że Polska jest klasą „pozytywną". Zmienną zależną jest narodowość respondentów – zmienna binarna. Zmiennymi niezależnymi są odpowiedzi na pytania. Są one typowe dla analizy sąsiedztwa i dotyczą: oceny zainteresowania PN, stosunku do PN, źródeł wiedzy o PN, stosunku do pracy zarządu, ograniczeń wynikających z zamieszkiwania w sąsiedztwie PN, pytań odnoszących się do turystyki w PN oraz stopnia zgodności wobec stwierdzeń o PN.

Jakość klasyfikacji oceniono za pomocą tablicy pomyłek i krzywej ROC. Tablica pomyłek jest powszechnie stosowanym narzędziem w ocenie jakości zagadnień klasyfikacyjnych. Jest to tablica zawierająca klasę rzeczywistą i predykowaną dla badanych kategorii [Boehmke, Greenwell 2020, pp. 34-35] (tabela 1).

Tabela 1. Tablica pomyłek

Klasa predykowana	Klasa rzeczywista	
	negatywna	pozytywna
negatywna	Prawdziwie negatywna (TN)	Fałszywie negatywna (FN)
pozytywna	Fałszywie pozytywna (FP)	Prawdziwie pozytywna (TP)

Źródło: opracowanie własne na podstawie Boehmke i Greenwell (2020)

Na podstawie tablicy pomyłek obliczono następujące miary oceny klasyfikacji:

- dokładność (ACC) – odsetek poprawnie sklasyfikowanych obiektów:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

- czułość, czyli odsetek prawdziwie pozytywnych (TPR) – odsetek poprawnie sklasyfikowanych obiektów pozytywnych:

$$TPR = \frac{TP}{TP+FN} \tag{2}$$

- specyficzność, czyli odsetek prawdziwie negatywnych (TNR) – odsetek poprawnie sklasyfikowanych obiektów negatywnych:

$$TNR = \frac{TN}{TN+FP} \tag{3}$$

- precyzja, czyli wartość predykcyjna dodatnia (PPV) – odsetek poprawnie sklasyfikowanych obiektów jako pozytywne w całkowitej liczbie obiektów sklasyfikowanych jako pozytywne:

$$PPV = \frac{TP}{TP+FP} \tag{4}$$

Na podstawie krzywej ROC (receiver operating characteristic) oblicza się pole pod krzywą (area under the curve – AUC). Jest to prawdopodobieństwo, że model oceni wyżej losowy element klasy pozytywnej od losowego elementu klasy negatywnej.

W ostatnim etapie badania zbadano istotność i siłę współzależności pomiędzy zmienną objaśnianą (państwem zamieszkania respondentów), a tymi zmiennymi objaśniającymi (odpowiedziami na pytania), które stanowiły kryteria podziału drzewa. Ponieważ zmienna objaśniana jest zmienną nominalną dwustanową, dlatego zastosowano współczynnik V Craméra postaci:

$$V = \sqrt{\frac{\chi^2}{n \min(k-1, r-1)}} \tag{5}$$

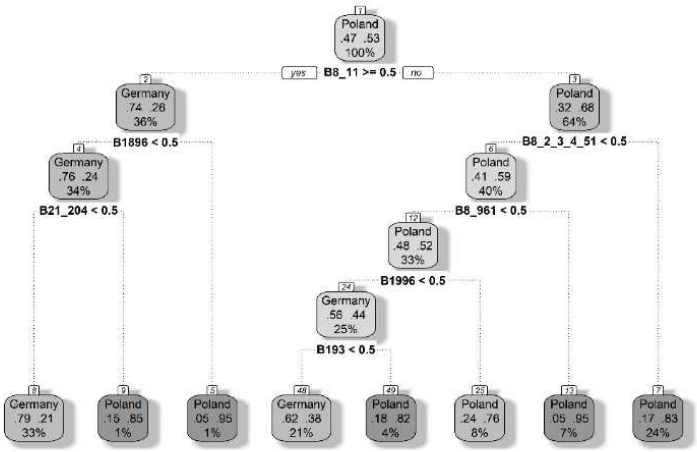
gdzie: χ^2 – wartość statystyki chi-kwadrat, n – liczba obserwacji, k – liczba wierszy w tablicy kontyngencji, r – liczba kolumn w tablicy kontyngencji. Istotność współzależności pomiędzy badanymi zmiennymi zbadano za pomocą testu χ^2 Pearsona.

WYNIKI BADANIA

Z łącznej liczby 2357 ankiet część nie była w pełni wypełniona. Liczba całkowicie wypełnionych ankiet wyniosła 2212, z czego 1548 wybrano losowo do zbioru uczącego, a pozostałe 664 utworzyły zbiór testowy. W wyniku uczenia modelu, wybrano najlepszy (z największą powierzchnią pod krzywą ROC). Drzewo klasyfikacyjne dla zbioru uczącego przedstawia rysunek 1. Cechami, które dzieliły badaną zbiorowość, były odpowiedzi na pytania: B8, B18, B19 oraz B21_04. Pytanie B8 dotyczyło źródeł informacji o PN. Ponieważ było to pytanie wielokrotnego wyboru, dlatego było opisane większą liczbą zmiennych. B8_1 oznacza prasę lokalną, jako źródło informacji. B8_2_3_4 oznacza szeroko rozumiany Internet. Pytanie B18 oznacza ocenę znaczenia badanego PN dla turystyki w regionie, B19 – ocenę liczby turystów w PN, a B21_20 – ocenę stopnia zgodności ze stwierdzeniem, że zarząd PN pracuje dobrze. W pierwszym kroku respondenci zostali podzieleni na Polaków i Niemców według wskazań na prasę lokalną, jako na źródło informacji o PN. Niemcy wskazywali tę odpowiedź znacznie częściej, niż Polacy. Następnie osoby, które wybrały prasę lokalną, zostały podzielone według oceny wpływu PN na turystykę w regionie. Osoby, które nie miały zdania na ten temat, były w większości Polakami i nie zostały dalej podzielone. Z kolei osoby, które miały zdanie na ten temat, były w większości Niemcami. W ostatnim kroku osoby te zostały podzielone według stopnia zgodności ze stwierdzeniem, że zarząd PN pracuje dobrze. W mniejszym stopniu zgadzali się z nim głównie Niemcy, a w większym – Polacy. Wracając do pierwszego kroku podziału drzewa, wybrano osoby, które nie wskazały pracy

lokalnej jako źródłą wiedzy o PN (byli to głównie Polacy). Następnie podzielono je według wskazania, czy Internet był głównym źródłem informacji. Osoby, które odpowiedziały twierdząco, były głównie Polakami i nie zostały dalej podzielone. Osoby, które nie wskazały Internetu, zostały podzielone według tego, czy miały jakiekolwiek zdanie na ten temat, czy nie. Osoby nie mające zdania zostały ostatecznie sklasyfikowane jako Polacy, a pozostałe zostały podzielone według oceny liczby turystów. Osoby nie mające zdania na ten temat zostały ostatecznie sklasyfikowane, jako Polacy. W ostatnim kroku podzielono osoby mające swoją opinię na temat oceny liczby turystów. Osoby odpowiadające, że liczba turystów jest prawidłowa lub zbyt duża, zostały ostatecznie sklasyfikowane jako Niemcy, a osoby twierdzące, że liczba ta jest zbyt mała – jako Polacy.

Rysunek 1. Drzewo klasyfikacyjne dla zbioru uczącego



Źródło: opracowanie własne

Na podstawie drzewa klasyfikacyjnego dokonano klasyfikacji obiektów ze zbioru testowego. Tabela 2 przedstawia tablicę pomyłek dla dokonanej klasyfikacji. Miary jakości klasyfikacji wyznaczone na podstawie tabeli 2 i równań (1)–(4) przedstawia tabela 3.

Tabela 2. Tablica pomyłek dla zbioru testowego

Klasa predykowana	Klasa rzeczywista	
	Niemcy	Polska
Niemcy	229	105
Polska	69	261

Źródło: opracowanie własne

Tabela 3. Miary jakości klasyfikacji

Miary	Wartości
ACC	0,7380
TPR	0,7131
TNR	0,7685
PPV	0,7909

Źródło: opracowanie własne

Jakość klasyfikacji modelu była dobra. Prawie 74% respondentów zostało poprawnie sklasyfikowanych. W większym stopniu (prawie 77%) poprawnie zostali sklasyfikowani mieszkańcy okolic niemieckich parków narodowych. W przypadku Polski było to 71,3%. 79% mieszkańców sklasyfikowanych, jako Polacy, zostało sklasyfikowanych poprawnie. Wartość AUC wyniosła 0,7655, co potwierdza dobre własności predyktywne modelu. Istotność współzależności pomiędzy odpowiedziami na te pytania oraz jej siłę, a narodowością respondentów dla zbioru uczącego i testowego przedstawia tabela 4.

Tabela 4. Ocena współzależności pomiędzy odpowiedziami na pytania, a narodowością respondentów

Pytania	zbiór uczący		zbiór testowy	
	V	wartość p	V	wartość p
B8_1	0,3982	0,0000	0,3173	0,0000
B8_2_3_4	0,2556	0,0000	0,2042	0,0000
B18	0,3826	0,0000	0,3502	0,0000
B19	0,3586	0,0000	0,4055	0,0000
B21_20	0,4229	0,0000	0,4372	0,0000

Źródło: opracowanie własne

Zarówno w zbiorze uczącym, jak i testowym wszystkie współczynniki współzależności były statystycznie istotne. Zależność pomiędzy wskazaniem Internetu, jako źródła informacji o parku narodowym, a narodowością była słaba. W pozostałych przypadkach była to zależność o przeciętnej sile. Powyższe wyniki potwierdzają fakt wpływu narodowości respondentów na udzielane przez nich odpowiedzi.

PODSUMOWANIE

Prezentowane badanie zostało przeprowadzone na polsko-niemieckim pograniczu funkcjonującym w ramach Euroregionu Pomerania. Granica polityczna rozdziela dwa społeczeństwa o różnym poziomie rozwoju gospodarczego, innej kulturze i języku. Przeprowadzona ankieta wskazała na dużo podobieństw w podejściu do ochrony przyrody i konieczności istnienia obszarów chronionych. Bez względu na kraj zamieszkania mieszkańcy pozytywnie oceniają fakt istnienia obszaru i możliwość rozwoju turystyki na tych terenach. Badanie wykazało, że

istnieje kilka zagadnień różnicujących badane społeczeństwa: źródła pozyskiwania informacji o PN, ocena znaczenia badanego PN dla turystyki w regionie, ocena liczby turystów w PN oraz ocena pracy zarządu PN. Respondenci niemieccy częściej wskazywali prasę lokalną, jako źródło informacji, a Polacy – Internet. Respondenci polscy częściej uważali, że w regionie jest zbyt mało turystów, niemieccy zaś – że jest ich albo za dużo, albo ich liczba jest prawidłowa. Polacy częściej oceniali pracę zarządu PN lepiej, niż Niemcy. Inne badania prowadzone przez Mayera et al. [2019] związane z rozwojem turystyki w tym regionie wskazują na to, że polscy i niemieccy respondenci mają nie tylko różny poziom wiedzy na temat obszarów chronionych, ale także odmienne wzorce ich odwiedzin. Badania te również wskazały na większą skłonność korzystania z Internetu przez polskich turystów odwiedzających przygraniczne obszary chronione.

Zaproponowana w artykule metoda badawcza nie wyczerpuje wszystkich możliwości. Można zastosować inne metody klasyfikacji (np. lasy losowe, gradient boosting). Będą one przez Autorów wykorzystywane w dalszych analizach. Można byłoby również rozpatrywać każde pytanie osobno, a uzyskane wyniki porównywać odpowiednim testem statystycznym. Jednak takie podejście uniemożliwia wyodrębnienie tych najbardziej istotnych czynników.

Przeprowadzone badania w związku z realizowanym projektem są unikatowe. Granice państwowe wynikają z prowadzonej polityki. Obszar przyrodniczy w Euroregionie Pomerania stanowi jedną całość. Ochrona przyrody staje się więc wspólnym dobrem. Przeprowadzone badania są bardzo ważne z punktu widzenia prowadzenia wspólnych polsko-niemieckich działań na obszarze pogranicza. Poznanie różnic i podobieństw w postrzeganiu i akceptacji obszarów chronionych przez mieszkańców pozwoli na udoskonalenie wspólnych działań w zakresie ochrony przyrody i rozwoju zrównoważonego.

BIBLIOGRAFIA

Aguirre G. J. (2006) Linking National Parks with its Gateway Communities for Tourism Development in Central America: Nindiri, Nicaragua, Bagazit, Costa Rica and Portobelo, Panama. Pasos. Revista de Turismo y Patrimonio Cultural, 4(3), 351-371, doi:10.25145/j.pasos.2006.04.024.

Andereck K. L., Valentine K. M., Knopf R. C., Vogt C. A. (2005) Residents' Perceptions of Community Tourism Impacts. Annals of Tourism Research, 32(4), 1056-1076, doi:10.1016/j.annals.2005.03.001.

Andries D. M., Arnaiz-Schmitz C., Díaz-Rodríguez P., Herrero-Jáuregui C., Schmitz M. F. (2021) Sustainable Tourism and Natural Protected Areas: Exploring Local Population Perceptions in a Post-Conflict Scenario. Land 10(3), 331, doi:10.3390/land10030331.

Ayivor J. S., Nyametso J. K., Ayivor S. (2020) Protected Area Governance and Its Influence on Local Perceptions, Attitudes and Collaboration. Land 9(9), 310, doi:10.3390/land9090310.

Boehmke B., Greenwell B. M. (2020) Hands-On Machine Learning with R. Chapman and Hall/CRC, Boca Raton.

Cappelli C., Zhang H. (2007) Survival Trees. [w:] Härdle W.,·Mori Y., Vieu P. (red.) Statistical Methods for Biostatistics and Related Fields. Springer-Verlag, Berlin, 167-179.

Grainger J. (2003) 'People Are Living in the Park'. Linking Biodiversity Conservation to Community Development in the Middle East Region: A Case Study from the Saint Katherine Protectorate, Southern Sinai. Journal of Arid Environments, 54(1), 29-38, doi:10.1006/jare.2001.0894.

Hamin M. (2002) Western European Approaches to Landscape Protection: A Review of the Literature. J. Plan. Lit. 16(3), 339-358, doi:10.1177/08854120222093400.

Hargreaves-Allen V., Mourato S., Milner-Gulland E. (2011) A Global Evaluation of Coral Reef Management Performance: Are MPAs Producing Conservation and Socio-Economic Improvements? Environment Management, 47(4), 684-700. doi:10.1007/s00267-011-9616-5.

Hibszer A. (2013) Parki narodowe w świadomości i działaniach społeczności lokalnych. Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Job H., Fließbach-Schendzielorz M., Bittlingmaier S., Herling A., Woltering M. (2019) Akzeptanz der bayerischen Nationalparks: ein Beitrag zum sozioökonomischen Monitoring in den Nationalparks Bayerischer Wald und Berchtesgaden. Würzburg University Press, Würzburg.

Królikowska K. (2007) Konflikty społeczne w polskich parkach narodowych. Oficyna Wydawnicza „Impuls", Kraków.

Leitinger G., Walde J., Bottarin R., Tappeiner G., Tappeiner U. (2010) Identifying Significant Determinants for Acceptance of Nature Reserves: A Case Study in the Stilfserjoch National Park, Italy. Journal on Protected Mountain Areas Research and Management, 2, 15-22, doi:10.1553/eco.mont-2-1s15.

Mashizi A. K., Sharafatmandrad M. (2020) Assessing Ecological Success and Social Acceptance of Protected Areas in Semiarid Ecosystems: A Socio-ecological Case Study of Khabr National Park, Iran. Journal for Nature Conservation, 57, 125898, doi:10.1016/j.jnc.2020.125898.

Matuszewska D (2003) Funkcje turystyczne i konflikty w wybranych parkach narodowych Polski północno-zachodniej. Bogucki Wydawnictwo Naukowe, Poznań.

Mayer M., Zbaraszewski W., Pieńkowski D., Gach G., Gernert J. (2019) Cross-Border Tourism in Protected Areas. Geographies of Tourism and Global Change. Springer, Cham.

Mudunuru V. R. (2016) Modeling and Survival Analysis of Breast Cancer: A Statistical, Artificial Neural Network, and Decision Tree Approach. Graduate Theses and Dissertations. http://scholarcommons.usf.edu/etd/6120, [dostęp 19.01.2019].

Plummer R., Fennell D. A. (2009) Managing Protected Areas for Sustainable Tourism: Prospects for Adaptive Co-management. Journal of Sustainable Tourism, 17(2), 149-168, doi:10.1080/09669580802359301.

Pollnac R. B., Crawford B. R., Gorospe M. L. G. (2001) Discovering Factors Influencing the Success of Community-based Marine Protected Areas in the Visayas, Philippines. Ocean & Coastal Management, 44, 683-710, doi:10.1016/S0964-5691(01)00075-8.

Rao K. S., Nautiyal S., Maikhuri R. K., Saxena K. G. (2003) Local Peoples' Knowledge, Aptitude and Perceptions of Planning and Management Issues in Nanda Devi Biosphere

Reserve, India. Environmental Management, 31(2), 0168–0181, doi:10.1007/s00267-002-2830-4.

Schumacher H., Job H. (2013) Nationalparks in Deutschland – Analyse und Prognose. Natur und Landschaft, 88(7), 309-314.

Seridi A., Djebar A. B. (2017) Proportions of Social Acceptance in the Designation of a Marine Protected Area: Cap de Garde in Annaba, Algeria (SW Mediterranean). Aquaculture, Aquarium, Conservation & Legislation - International Journal of the Bioflux Society, 10(3), 480-498.

Sieberath J. (2007) Die Akzeptanz des Nationalparks Eifel bei der lokalen Bevölkerung Eine empirische Untersuchung zur Verankerung eines Großschutzgebietes in der Region. Bundesamt für Naturschutz, Bonn.

Sirakaya E., Teye V., Sönmez S. (2002) Understanding Residents' Support for Tourism Development in the Central Region of Ghana. Journal of Travel Research, 41(1), 57-67, doi:10.1177/004728750204100109.

Stoll-Kleemann S. (ed.) (2015) Wahrnehmung und Akzeptanz des Bundesländerübergreifenden Naturparks Barnim. Ernst-Moritz-Arndt Universität Greifswald, Greifswald, doi:10.23689/fidgeo-1982.

Von Ruschkowski E. (2009) Ursachen und Lösungsansätze für Akzeptanzprobleme von Großschutzgebieten am Fallbeispiel von zwei Fallstudien im Nationalpark Harz und im Yosemite National Park. Ibidem-Verlag, Stuttgart.

Von Ruschkowski E., Mayer M. (2011) From Conflict to Partnership? Interactions between Protected Areas, Local Communities and Operators of Tourism Enterprises in Two German National Park Regions. Journal of Tourism and Leisure Studies, 17(2), 147-181.

Voyer M., Gladstone W., Goodall H. (2012) Methods of Social Assessment in Marine Protected Area Planning: Is Public Participation Enough? Marine Policy, 36(2), 432-439, doi:10.1016/j.marpol.2011.08.002.

Zawilińska B. (2020) Postawy lokalnych społeczności i turystów wobec parków krajobrazowych w województwie małopolskim. Prace Geograficzne, 160, 95-116, doi:10.4467/20833113PG.20.002.12260.

CLASSIFICATION TREES IN THE IDENTIFICATION OF FACTORS DIFFERENTIATING THE PERCEPTION OF THE NEIGHBOURHOOD OF NATIONAL PARKS IN EUROREGION POMERANIA

Abstract: The aim of the research is the assessment of differences in perception of the neighbourhood of national parks (NP) in the Euroregion Pomerania by inhabitants of Poland and Germany. The classification trees method was used to classify the inhabitants on the basis of their answers to the survey questions. Obtained results prove that there are differences in perception of the neighborhood of national parks by Poles and Germans. These are: sources of information about NP, evaluation of importance of analysed NP for tourism in the region, evaluation of number of tourists in NP and evaluation of work of management of NP. There was also found a significant relationship between the nationality of the respondents and the answers differentiating the analysed group.

Keywords: classification trees, correlation analysis, acceptance analysis, Euroregion Pomerania

JEL classification: G11, G17

QUANTITATIVE METHODS IN ECONOMICS
Volume XXII, No. 1, 2021, pp. 14 – 28

AN APPLICATION OF THE INTERVAL ESTIMATION FOR THE AT-RISK-OF-POVERTY RATE ASSESSMENT

Marcin Dudziński https://orcid.org/0000-0003-4242-8411
Joanna Kaleta https://orcid.org/0000-0001-6628-4251
Institute of Information Technology
Warsaw University of Life Sciences – SGGW, Poland
e-mails: marcin_dudzinski@sggw.edu.pl; joanna_kaleta@sggw.edu.pl

Abstract: In the document [Eurostat (Your Key to European Statistics) 2020], *At-Risk-of-Poverty Rate* (*ARPR* in short) is defined as the percentage of population with an income not exceeding 60% of the general population median income. Extensive and thorough research on the estimation of this measure has been conducted since its introduction. For example, in the paper of [Zieliński 2009a] a non-parametric, distribution-free confidence interval for *ARPR* has been constructed. An example of application of the confidence interval proposed by [Zieliński 2009a] has been given in [Zieliński 2009b]. Some other interesting approach regarding the interval estimation of *ARPR* has been proposed in [Luo and Qin 2017], where the authors introduced new concepts of the interval estimation for the so-called *Low-Income Proportion* (*LIP*) measure, which is a generalization of *ARPR*. The *LIP* measure and thus, the *ARPR* parameter in particular, are important indexes describing the inequality in an income distribution. Based on the construction of the point smoothed kernel estimate for *LIP*, [Luo and Qin 2017] established a smoothed jackknife empirical likelihood approach leading to the introduction of some new non-parametric confidence intervals for the *LIP* measure and consequently, for the *ARPR* index as well. In our work, we aim to apply the most interesting ideas of *LIP* and *ARPR* point and interval estimation for data consisting of 13057 observations concerning an equalised disposable income of households in Poland from 2003. We also discuss the accuracy and adequacy of the empirical results relating to the *ARPR* interval estimation, obtained by the implementation of the constructed confidence intervals.

Keywords: Low-Income Proportion (*LIP)*, At-Risk-of-Poverty Rate (*ARPR*), confidence intervals for *LIP* and *ARPR*, Nonparametric estimation, Kernel estimation

JEL classification: C51, C52

https://doi.org/10.22630/MIBE.2021.22.1.2

INTRODUCTION

At-Risk-of-Poverty Rate (or *ARPR* in short) is a measure that enables to determine the inequality in an income distribution. According to [Eurostat (Your Key to European Statistics) 2020], it is defined as the proportion of general population with an income not exceeding 60% of the median income in the whole population. Using mathematical terms, we may describe this measure in the following pattern. Namely, let $\mathrm{EQ\_INC}_j$ denote an equivalised disposable income of the j-th individual (person or household) and suppose that weight_i stands for the weight of individual i. Firstly, we shall determine the so-called *At-Risk-of-Poverty Threshold* (or *ARPT* in short). It is expressed as (see also [Zieliński 2009a-b])

$$ARPT = \textit{At-Risk-of-Poverty Threshold} = 60\%\mathrm{EQ\_INC}_{\mathrm{MEDIAN}},$$

where

$$\mathrm{EQ}_{\mathrm{INC}_{\mathrm{MEDIAN}}} = \begin{cases} \frac{1}{2}\left(\mathrm{EQ\_INC}_j + \mathrm{EQ\_INC}_{j+1}\right), & \text{if } \sum_{i=1}^{j} \mathrm{weight}_i = \frac{W}{2} \\ \mathrm{EQ\_INC}_{j+1}, & \text{if } \sum_{i=1}^{j} \mathrm{weight}_i < \frac{W}{2} < \sum_{i=1}^{j+1} \mathrm{weight}_i \end{cases},$$

where in turn,

$$W = \sum_{All\ persons} \mathrm{weight}_i.$$

Thus, we can directly come to stating the definition of *ARPR*. Since it is clear now that this measure denotes the percentage of individuals from the whole population with an equivalised disposable income not greater than *ARPT*, then the *ARPR* index is calculated as (see also [Zieliński 2009a-b])

$$ARPR = \frac{\sum_{\substack{All\ persons\ with \\ \mathrm{EQ\_INC} \le ARPT}} \mathrm{weight}_i}{W} \times 100.$$

We are now in a position to discuss the estimation methods for *ARPR*. We will start from the point estimation of this measure. Suppose that $X_1, X_2, \ldots, X_n$ is a sample of the equivalised disposable incomes of randomly drawn n individuals and let Med be the corresponding sample median. A straightforward point estimate for *ARPR* is given by (see, e.g., [Zieliński 2009a-b])

$$\widehat{ARPR} = \frac{1}{n}\#\{X_i\colon X_i \le 0.6 \cdot Med\},$$

with # standing for the cardinality of the considered set. It is obvious that in terms of probability distribution, the *ARPR* index is determined as

$$\theta = ARPR = F\big(0.6 \cdot F^{-1}(0.5)\big),$$

where: F denotes the cumulative distribution function (cdf) of an equivalised disposable income in the investigated general population, F^{-1} is the corresponding quantile function. *ARPR* is a special case of the so-called *Low-Income Proportion* (*LIP*) measure, which is an index defined for two parameters, usually denoted as α and β. Namely, if X denotes an income variable with a cdf F, then *LIP* is given by

$$LIP = \theta_{\alpha,\beta} = P\big(X \le \alpha \cdot \xi_\beta\big) = F\big(\alpha \cdot \xi_\beta\big) = F\big(\alpha \cdot F^{-1}(\beta)\big),$$

where ξ_β denotes the β-th quantile of an income distribution. Thus, for the fixed α and β, *LIP* is the fraction of individuals with an equivalised disposable income not exceeding $\alpha \cdot \xi_\beta = \alpha \cdot F^{-1}(\beta)$ (in other words, it is the proportion of population with an income not greater than the given fraction α of the β-th quantile from an income distribution). It is clear that *LIP* equals *ARPR* for $\alpha = 0.6$ and $\beta = 0.5$. The *Low-Income Proportion*, and consequently the *At-Risk-of-Poverty Rate* as its special case, are the measures that have been extensively used by governing bodies and government experts, as well as by business managers and advisors or academics from different areas of interest, in order to gain a great deal of valuable information and conclusions. It is particularly convenient and useful in the assessment of potential inequalities regarding the socio-economic status. For example, the employees with earnings not exceeding 60% of the population median income are treated as the low-earners by the European Statistical Office 'Eurostat'. Since, as it has already been mentioned, *ARPR* is equivalent to *LIP* with $\alpha = 0.6$ and $\beta = 0.5$, the high values of *ARPR* indicate relatively large social inequalities in the wealth structure, as well as the social instability and uncertainty. All this together should serve as a warning signal for the state decision-makers. Except for the state authorities and business entrepreneurs, *LIP* and *ARPR* have attracted much attention of scholars from various fields of interest. In particular, a large number of inference methods related to both the point and the interval estimation of *LIP* have been proposed or developed. Among numerous research papers devoted to the subject of *LIP* and *ARPR* evaluation, or the risk measures assessment in general, the works of [Gong et al. 2010, Jing et al. 2009, Li et al. 2011, Luo and Qin 2017, Wei et al. 2009, Wei and Zhu 2010] and [Zieliński 2009a-b] - are especially worthwhile to mention. Roughly speaking, there exist two essential concepts concerning the estimation of *LIP*, and *ARPR* in particular. With reference to the issue of *LIP* estimation, these two primary approaches - commonly known as the empirical and kernel methods - may be illustrated as follows. Let $X_1, X_2, \dots, X_n$ denote a simple sample from the income distribution having a cdf F. Then, the empirical estimate for parameter $\theta_{\alpha,\beta} = LIP$ is defined by (see also [Luo and Qin 2017])

$$\hat{\theta}_{\alpha,\beta} = F_n\big(\alpha\hat{\xi}_\beta\big) = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \le \alpha\hat{\xi}_\beta) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,\alpha\hat{\xi}_\beta]}\,(X_i),$$

where: F_n stands for the empirical distribution function of $X_1, \ldots, X_n$, $\hat{\xi}_\beta = F_n^{-1}(\beta)$ is the β-th quantile of the empirical distribution function F_n, while I_A denotes the indicator function of a given set A. Sadly, an application of the empirical point estimate has a relatively serious drawback, which consists in the fact that $\hat{\theta}_{\alpha,\beta} = F_n(\alpha\hat{\xi}_\beta)$ is a non-smoothing estimator of $\theta_{\alpha,\beta}$, as it is a non-smoothing function of the sample quantile $\hat{\xi}_\beta$, whereas $LIP = \theta_{\alpha,\beta} = F(\alpha\xi_\beta)$ is a function related to the smoothing income distribution function F. Therefore, instead of employing a non-smoothing empirical estimator $\hat{\theta}_{\alpha,\beta}$, [Luo and Qin 2017] suggested using the kernel method in order to obtain a smoothed estimator for $\theta_{\alpha,\beta}$. A comprehensive study has shown an advantage of the kernel estimation over an assessment based on the implementation of the empirical estimator (see, e.g., [Falk 1983]-[Falk 1985] in this context). The kernel estimator of the *LIP* index $\theta_{\alpha,\beta}$ is given by the formula

$$\hat{T}_n(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\alpha\hat{\xi}_\beta - X_i}{h}\right),$$

where K is the so-called kernel function and h denotes the chosen bandwidth (h is also interchangeably known under the name of smoothing parameter). It turns out that the kernel estimator $\hat{T}_n(\alpha,\beta)$ has a slightly smaller *Mean Squared Error (MSE)* than the empirical estimator $\hat{\theta}_{\alpha,\beta}$ (see Table 1 in [Luo and Qin 2017]). Apart from the fact that the kernel estimator is more suitable for the *LIP* assessment, this kind of estimator is also used in the definition of a smoothed version of the jacknife empirical likelihood ratio statistic for *LIP*. This smoothed version may by later applied in the constructions of the corresponding confidence intervals. One of significant difficulties arising in calculation of the smoothed estimator $\hat{T}_n(\alpha,\beta)$ is the problem of an appropriate choice of bandwidth h for this kernel estimator. Many methods of bandwidth selection have been proposed so far (see, e.g., [Bowman et al. 1998], among others, for a comprehensive review regarding this matter). In particular, [Luo and Qin 2017] use the twofold cross-validation method for bandwidth selection in order to estimate *LIP* and, after conducting some simulation research, they recommend using bandwidth of the form $h = cn^{-1/3}$, where c is some constant not depending on n. On the other hand, it is worthwhile to mention that intensive simulation studies indicate that the selection of kernel *K* itself is not of so high importance, since the change of kernel does not affect the obtained estimation results too much. In the cited work of [Luo and Qin 2017], the authors use the triweight kernel density function $K(t) = \frac{35}{32}(1-t^2)^3 I(|t| \leq 1)$ in order to evaluate *LIP* for data relating to annual salaries of Professors, Associate Professors and Assistant Professors, employed in the Units of University System or in Military Colleges from a State of Georgia, U.S., during the 2012 fiscal year. Our primary objective is to use some chosen estimation procedures for both the point and the interval estimation of *LIP* and *ARPR*, for evaluation of *ARPR* on the basis of dataset

containing 13057 observations of an equivalised disposable income of Polish households in 2003 from [Statistical Publishing Establishment, Warsaw, 2004]. Above all, we apply and develop the methods proposed by [Luo and Qin 2017] and [Zieliński 2009a-b], as - in our view - these approaches include the most valuable and reliable ideas leading to the assessment of *ARPR* and to the evaluation of other similar poverty (or social inequality) measures. The remainder of our paper is structured as follows. In Section SELECTED CONCEPTS OF THE LOW-INCOME PROPORTION AND THE AT-RISK-OF-POVERTY RATE ESTIMATION, we introduce the essential concepts of *LIP* and *ARPR* assessment, which we later aim to implement in our empirical analyses. In Section EMPIRICAL STUDY, we present the details regarding computational techniques that allow for the corresponding point and interval estimation, as well as we conduct our empirical research concerning the point and interval evaluation of the mentioned risk measures for a dataset containing information on an equivalised disposable income of households in Poland from the year 2003. Finally, Section SUMMARY summarizes and concludes our study. All of our computations have been carried out using the software environment R.

SELECTED CONCEPTS OF THE LOW-INCOME PROPORTION AND THE AT-RISK-OF-POVERTY RATE ESTIMATION

Let $X_1, X_2, \dots, X_n$ be a simple random sample from the income distribution and $X_{1:n} \leq \cdots \leq X_{n:n}$ denote a sequence of the corresponding order statistcs. In view of the definition of the *Low-Income Proportion (LIP)* measure from the previous Section, the estimator of $\theta_{\alpha,\beta} = LIP$ may be defined as

$$\hat{\theta}_{\alpha,\beta} = \frac{1}{n}\#\{X_i\colon X_i \leq \alpha \cdot X_{M:n}\},$$

where $M = \lfloor \beta n \rfloor + 1$ (with $\lfloor x \rfloor$ denoting the largest integer not exceeding x) and # stands for the cardinality of a given set; obviously, $X_{M:n}$ is an estimator for the β-th quantile ξ_β of an income distribution. Therefore, an estimator of the *At-Risk-of-Poverty Rate (ARPR)* may be given by putting $\alpha = 0.6$ and $M = \lfloor 0.5n \rfloor + 1$ into the formula above, i.e. it can be determined as

$$\widehat{ARPR} = \hat{\theta}_{0.6,0.5} = \hat{\theta} = \frac{1}{n}\#\{X_i\colon X_i \leq 0.6 \cdot X_{(\lfloor 0.5n \rfloor+1):n}\}.$$

(Clearly, $X_{(\lfloor 0.5n \rfloor+1):n}$ is an estimator for a median of an income distribution)

As the first example of a confidence interval for $\theta_{\alpha,\beta} = LIP$ (and consequently for $\theta_{0.6,0.5} = ARPR$), we wish to examine an interval constructed in [Zieliński 2009b]. Namely, let ξ be the number of those among $X_1, X_2, \dots, X_n$, which are not greater than $\alpha \cdot X_{(\lfloor \beta n \rfloor+1):n}$, i.e.

$$\xi = \#\{X_i\colon X_i \leq \alpha \cdot X_{(\lfloor \beta n \rfloor+1):n}\}.$$

Then, assuming the fixed confidence level $\gamma \in (0,1)$, the following confidence interval for $\theta_{\alpha,\beta} = LIP$ has been introduced in [Zieliński 2009a]

$$(\beta \cdot B^{-1}(\xi, M-\xi+1; \frac{1-\gamma}{2});\ \beta \cdot B^{-1}(\xi+1, M-\xi; \frac{1+\gamma}{2})),$$

where: $M = \lfloor \beta n \rfloor + 1$, and $B^{-1}(a,b;q)$ denotes a quantile of order q for the beta distribution with parameters a, b. Thus, as a straightforward conclusion, we can write that an interval below is the corresponding confidence interval for $\theta_{0.6,0.5} = \theta = ARPR$

$$(l_0; u_0) =$$
$$\left(0.5 \cdot B^{-1}\left(\tilde{\xi}, \widetilde{M} - \tilde{\xi} + 1; \frac{1-\gamma}{2}\right);\ 0.5 \cdot B^{-1}\left(\tilde{\xi}+1, \widetilde{M} - \tilde{\xi}; \frac{1+\gamma}{2}\right)\right),$$

where:

$$\widetilde{M} = \lfloor 0.5n \rfloor + 1,\ \tilde{\xi} = \#\{X_i\colon X_i \le 0.6 \cdot X_{\widetilde{M}:n}\}.$$

As it has already been mentioned in our preliminary Section, the empirical estimate of $\theta_{\alpha,\beta} = LIP$ may be given by

$$\hat{\theta}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \le \alpha\hat{\xi}_\beta).$$

Furthermore, in view of [Preston 1995], this estimate satisfies the following property

$$\sqrt{n}\left(\hat{\theta}_{\alpha,\beta} - \theta_{\alpha,\beta}\right) \to N\left(0, \sigma^2_{\alpha,\beta}\right),$$

where

$$\sigma^2_{\alpha,\beta} = \theta_{\alpha,\beta}\left(1-\theta_{\alpha,\beta}\right) - 2\alpha(1-\beta)\theta_{\alpha,\beta}\frac{f\left(\alpha\xi_\beta\right)}{f\left(\xi_\beta\right)} + \alpha^2\beta(1-\beta)[\frac{f\left(\alpha\xi_\beta\right)}{f\left(\xi_\beta\right)}]^2,$$

with f standing for density of the corresponding income distribution.

Obviously, it means that $\theta_{\alpha,\beta}$ is asymptotically normal and hence, the following $(1-\delta)$-level normal approximation-based confidence interval for $\theta_{\alpha,\beta}$ may be established

$$(l_1; u_1) = (\hat{\theta}_{\alpha,\beta} - \frac{z_{1-\delta/2} \cdot \hat{\sigma}_{\alpha,\beta}}{\sqrt{n}};\ \hat{\theta}_{\alpha,\beta} + \frac{z_{1-\delta/2} \cdot \hat{\sigma}_{\alpha,\beta}}{\sqrt{n}}),$$

where $z_{1-\delta/2}$ stands for the $(1-\delta/2)$-th quantile of the standard normal distribution and $\hat{\sigma}_{\alpha,\beta}$ denotes a consistent estimator of the standard deviation $\sigma_{\alpha,\beta}$. However, since - as it has already been noted in our introductory part - $\hat{\theta}_{\alpha,\beta}$ is a non-smoothing estimate of $\theta_{\alpha,\beta}$, another approach, adopting the concept of the kernel estimation, has been proposed in order to obtain the smoothed estimator of *LIP*. Based on a simple random sample $X_1, X_2, \ldots, X_n$, the corresponding kernel estimate for $\theta_{\alpha,\beta}$ is determined as follows

$$\hat{T}_n(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\alpha\hat{\xi}_\beta - X_i}{h}\right),$$

where: K, h are the selected kernel and the stated bandwidth, respectively, and $\hat{\xi}_\beta$ is the β-th empirical quantile of the considered distribution.

Due to Theorem 2.1 in [Luo and Qin 2017], we directly get that $\sqrt{n}\big(\hat{T}_n(\alpha,\beta) - \theta_{\alpha,\beta}\big) \to N\big(0,\sigma^2_{\alpha,\beta}\big)$, where $\sigma^2_{\alpha,\beta}$ is the same as earlier. Thus, the following $(1-\delta)$-level normal approximation-based confidence interval for $\theta_{\alpha,\beta}$ may be obtained

$$(l_2;u_2) = (\hat{T}_n(\alpha,\beta) - \frac{z_{1-\delta/2}\cdot\hat{\sigma}_{\alpha,\beta}}{\sqrt{n}}; \hat{T}_n(\alpha,\beta) + \frac{z_{1-\delta/2}\cdot\hat{\sigma}_{\alpha,\beta}}{\sqrt{n}}).$$

The definition of smoothed estimator $\hat{T}_n(\alpha,\beta)$ is applied in establishing the so-called smoothed log jackknife empirical likelihood ratio statistic for *LIP* and later, for creating the corresponding confidence interval. In order to introduce the estimation concepts leading to both of the mentioned constructions, it is needed to define the so-called jackknife pseudo-values for *LIP*. By [Tukey 1958], the jackknife pseudo-values for *LIP* are defined as follows

$$\hat{V}_k(\alpha,\beta) = n\hat{T}_n(\alpha,\beta) - (n-1)\hat{T}_{n-1,k}(\alpha,\beta), \quad k = 1,2,\dots,n,$$

where: $\hat{T}_{n-1,k}(\alpha,\beta) = \frac{1}{n-1}\sum_{j\neq k}^{n} K\left(\frac{\alpha\hat{\xi}_{\beta,-k} - X_j}{h}\right)$ refers to the determined smoothed estimator $\hat{T}_n(\alpha,\beta)$, but it is computed on $n-1$ observations $X_1, X_2, \dots, X_{k-1}, X_{k+1}, \dots, X_n$, and $\hat{\xi}_{\beta,-k} = F^{-1}_{n,-k}(\beta)$ is the β-th quantile from an empirical distribution $F_{n,-k}(x) = \frac{1}{n-1}\sum_{j\neq k}^{n} I\left(X_j \le x\right)$, based on $n-1$ observations (i.e., on all observations except for the *k*-th one).

Using the jackknife pseudo-values $\hat{V}_k(\alpha,\beta)$, $k = 1, \dots n$, we may define the log jackknife empirical likelihood ratio statistic in the form as below

$$l_n\big(\theta_{\alpha,\beta}\big) = -2\log L_n\left(\theta_{\alpha,\beta}\right) = 2\sum_{k=1}^{n} \log\{1 + \lambda\big(\hat{V}_k(\alpha,\beta) - \theta_{\alpha,\beta}\big)\},$$

where $L_n\big(\theta_{\alpha,\beta}\big)$ denotes the jackknife empirical ratio statistic for $\theta_{\alpha,\beta}$ and $\lambda = \lambda\big(\alpha,\beta,\theta_{\alpha,\beta}\big)$ is the solution to

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\hat{V}_k(\alpha,\beta) - \theta_{\alpha,\beta}}{1 + \lambda\big(\hat{V}_k(\alpha,\beta) - \theta_{\alpha,\beta}\big)} = 0.$$

It is known that, under certain conditions, $l_n\big(\theta_{\alpha,\beta}\big) \to \chi^2(1)$, where $\chi^2(1)$ stands for the chi-squared distribution with one degree of freedom. Thus, the $(1-\delta)$-level confidence interval for $\theta_{\alpha,\beta} = LIP$ may be given in the form

$$(l_3;u_3) = \{\theta: l_n(\theta) \le \chi^2_{1,1-\delta}\},$$

where $\chi^2_{1,1-\delta}$ denotes the $(1-\delta)$-th quantile of the $\chi^2(1)$ distribution. It is worthwhile to mention here that the variance $var(\sqrt{n}\hat{T}_n(\alpha,\beta))$ can be estimated by the sample variance of jackknife pseudo-values $\{\hat{V}_1(\alpha,\beta),\dots\hat{V}_n(\alpha,\beta)\}$ and that the jackknife variance estimator of $T_n(\alpha,\beta)$ is determined as

$$v_{JACK}(\alpha,\beta) = var\left(\sqrt{n}\hat{T}_n(\alpha,\beta)\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{V}_i(\alpha,\beta) - \frac{1}{n}\sum_{j=1}^{n}\hat{V}_j(\alpha,\beta))^2.$$

In view of Theorem 3.1 in [Luo and Qin 2017], we get $v_{JACK}(\alpha,\beta) \to \sigma^2_{\alpha,\beta}$, where $\sigma^2_{\alpha,\beta}$ is the same as in the earlier considerations. Thus, the following $(1-\delta)$-level normal approximation-based confidence interval for $\theta_{\alpha,\beta}$ may be introduced

$$(l_4;u_4) = (\hat{T}_n(\alpha,\beta) - \frac{z_{1-\delta/2}\cdot\sqrt{v_{JACK}(\alpha,\beta)}}{\sqrt{n}}; \hat{T}_n(\alpha,\beta) + \frac{z_{1-\delta/2}\cdot\sqrt{v_{JACK}(\alpha,\beta)}}{\sqrt{n}}).$$

The confident intervals $(l_1;u_1)$, $(l_2;u_2)$ and $(l_4;u_4)$ are established on the basis of normal approximation theorems. Such the approximation-based confidence intervals may perform poorly for the estimation of income-related ratios, since the income datasets tend to be skewed or have outliers. In order to overcome this drawback, another technique that enables to create the confidence intervals for the rates like *LIP*, and *ARPR* in particular, has been proposed in the case when asymptotic variance of the corresponding point estimator is unknown. This idea - called the bootstrap method - is due to [Efron 1979] and has become a celebrated estimation approach in recent decades. With reference to Efron's design, [Luo and Qin 2017] combined the bootstrap approach with the kernel estimation in order to obtain appropriate confidence intervals for $\theta_{\alpha,\beta}$. The concept introduced in the cited work of [Luo and Qin 2017] may be depicted as follows. Namely, assume that $(X_1^*, X_2^*, \dots, X_n^*)$ is a bootstrap sample from the original sequence $(X_1, X_2, \dots, X_n)$, i.e. $(X_1^*, X_2^*, \dots, X_n^*)$ is repeatedly drawn, with replacement, from $(X_1, X_2, \dots, X_n)$. Then, the bootstrap equivalent of the kernel estimate $\hat{T}_n(\alpha,\beta)$ is given by

$$\hat{T}_n^*(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\alpha\hat{\xi}_\beta^* - X_i^*}{h}\right).$$

After repeating the bootstrap procedure $B \geq 500$ times, i.e. after drawing $B \geq 500$ bootstrap samples $(X_{1b}^*, X_{2b}^*, \dots, X_{nb}^*)$, where $b = 1, \dots, B$, B bootstrap copies $\{\hat{T}_{nb}^*(\alpha,\beta)\}_{b=1,\dots,B} = \{\hat{T}_b^*\}_{1,\dots,B}$, of the estimate $\hat{T}_n(\alpha,\beta)$, are computed. Finally, based on the obtained bootstrap replicates $\{\hat{T}_1^*, \hat{T}_2^* \dots, \hat{T}_B^*\}$, the following $(1-\delta)$-level bootstrap kernel-based confidence intervals for $\theta_{\alpha,\beta}$ are established:

$$(l_5;u_5) = (\hat{T}_n(\alpha,\beta) - z_{1-\delta/2}\cdot\sqrt{V_T^*}; \hat{T}_n(\alpha,\beta) + z_{1-\delta/2}\cdot\sqrt{V_T^*}),$$

$$(l_6;u_6) = (\bar{T}^* - z_{1-\delta/2}\cdot\sqrt{V_T^*}; \bar{T}^* + z_{1-\delta/2}\cdot\sqrt{V_T^*}),$$

where: $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{T}_b^*$, $V_T^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{T}_b^* - \bar{T}^*\right)^2$.

In the subsequent Section, we aim to use the above discussed point and interval estimation procedures in order to evaluate *ARPR* for data containing the equivalised disposable incomes of households from Poland, gained for the year 2003. We wish to pay our special attention to the issue of *ARPR* estimation methods which apply the non-smoothing kernel-based designs, in particular to those methods which combine the kernel estimation with the jackknife resampling technique.

EMPIRICAL STUDY

In our computations, we consider a sample of data $(x_1, x_2, \dots, x_n)$, comprising of n = 13507 observations referring to an equivalised disposable income of the Polish households in the year 2003, collected from [Statistical Publishing Establishment, Warsaw, 2004]. Before we compute the realizations of confidence intervals for *ARPR*, we need to check whether the given observations come from a random simple sample. For this reason, we apply the so-called *Runs Test*. Based on our dataset and assuming the most common confidence level 0.95, we obtain the value of test statistic $U = 0.4025805$ and the critical (rejection) region $(-\infty; -1.96 > \cup < 1.96; \infty)$. Consequently, we do not reject the null hypothesis that our observations come from a random simple sample. Thus, we may proceed to computation of the empirical confidence intervals for *ARPR* (i.e., to calculation of these confidence intervals realizations). As we have already mentioned, the point estimate of *ARPR* may be expressed in the form $\widehat{ARPR} = F_n\left(0.6 \cdot \hat{\xi}_{0.5}\right) = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \leq 0.6 \cdot \hat{\xi}_{0.5})$, or by $\widehat{ARPR} = \frac{1}{n}\#\{X_i\colon X_i \leq 0.6 \cdot X_{(\lfloor 0.5n \rfloor + 1):n}\}$.

For our dataset, we have:

$$\widetilde{M} = (\lfloor 0.5n \rfloor + 1) = 6529,\ X_{\widetilde{M}:n} = \hat{\xi}_{0.5} = F_n^{-1}(0.5) = 5570.073,$$
$$\tilde{\xi} = \#\{X_i\colon X_i \leq 0.6 \cdot X_{\widetilde{M}:n}\} = 2083,$$

and hence, $\widehat{ARPR} = \hat{\theta}_{0.6,0.5} = 0.1595$.

Due to the earlier given formula for the confidence interval $(l_0; u_0)$, introduced by [Zieliński 2009a], we immediately obtain the following confidence interval for *ARPR*, at the confidence level $\gamma = 0.95$,

$$(l_0; u_0) = (0.5 \cdot B^{-1}(\tilde{\xi}, \widetilde{M} - \tilde{\xi} + 1; \frac{1-\gamma}{2});\ 0.5 \cdot B^{-1}(\tilde{\xi} + 1, \widetilde{M} - \tilde{\xi}; \frac{1+\gamma}{2})) = (0.1539; 0.1652).$$

Furthermore, it is easy to compute the realization of 95% (0.95-level) normal approximation-based confidence interval for $ARPR = \theta_{0.6,0.5}$. Namely, for $\alpha = 0.6$ and $\beta = 0.5$, we obtain that

$$(l_1; u_1) = (\hat{\theta}_{0.6,0.5} - \frac{z_{1-0.05/2} \cdot \hat{\sigma}_{0.6,0.5}}{\sqrt{13057}};\ \hat{\theta}_{0.6,0.5} + \frac{z_{1-0.05/2} \cdot \hat{\sigma}_{0.6,0.5}}{\sqrt{13057}}) = (0.1537; 0.1654).$$

Among the realizations of the kernel-based confidence intervals $(l_2; u_2)$-$(l_5; u_5)$, we shall consider the realization of $(l_3; u_3)$ first, as it is recommended by [Luo and Qin 2017], since the empirical study conducted there shows that, among the presented intervals, $(l_3; u_3)$ displays the best statistical performance in terms of coverage probabilities. Out of all the introduced confidence intervals for *ARPR*, $(l_3; u_3)$ seems to be relatively the most difficult one to evaluate, as a technique leading to the construction of $(l_3; u_3)$ combines the kernel estimation with the jackknife resampling concept. In particular, computing the realization of $(l_3; u_3)$ requires the selection of an appropriate kernel function K together with its bandwidth h. Although, it has been checked that the choice of K does not affect the accuracy of a calculated estimate too much, various research studies exhibit that it is not so in case of the bandwidth selection, since it has been shown that the change of h may have a significant impact on the estimator value. Thus, following the suggestions from [Luo and Qin 2017], we apply the triweight kernel $K(t) = \frac{35}{32}(1 - t^2)^3 I(|t| \leq 1)$ and implement a bandwidth $c \cdot n^{-1/3}$, where a constant c is selected on the grounds of the two-fold cross-validation method. The procedure leading to the calculation of c involves employing several steps, which may be described as follows.

Step 1°. We randomly split the given sample into two parts of possibly equal size, the first of which is the training sample, while the second is treated as the test sample;

Step 2°. Based on the training sample, we compute the kernel estimate $\hat{T}_{n,c}^{(1)}(\alpha = 0.6, \beta = 0.5) = \hat{T}_{n,c}^{(1)}(0.6,0.5)$ for $ARPR$ and based on the test sample, we compute its empirical estimate $\hat{\theta}^{(2)}_{\alpha=0.6,\beta=0.5} = \hat{\theta}^{(2)}_{0.6,0.5}$;

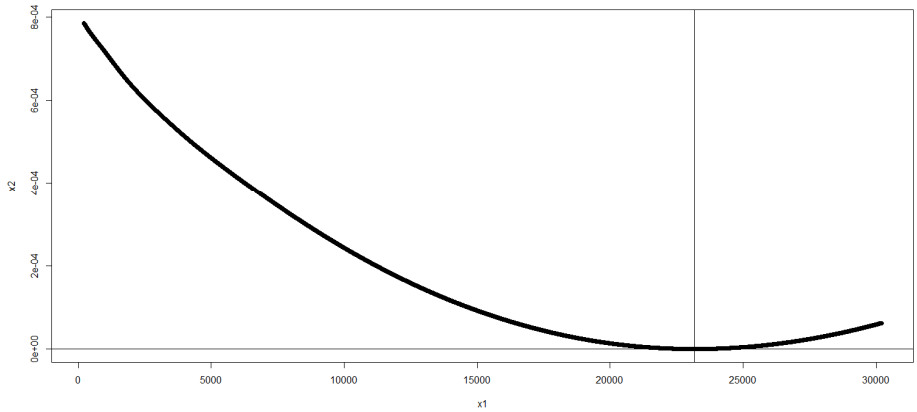
Step 3°. We repeat the random split and the computation, described in the previous steps, $L \geq 30$ times and obtain a set of L pairs consisting of the kernel estimates $\hat{T}_{n,c}^{(1,l)}(0.6,0.5)$ and the empirical estimates $\hat{\theta}^{(2,l)}_{0.6,0.5}$, where $l = 1, \dots, L$ (i.e., we get a set $\{\left(\hat{T}_{n,c}^{(1,l)}(0.6,0.5), \hat{\theta}^{(2,l)}_{0.6,0.5}\right) : l = 1, \dots, L\}$);

Step 4°. We choose a constant c by minimizing the following cross-validation estimate of *MSE*

$$CV_c = \frac{1}{n}\sum_{l=1}^{L}\left[\hat{T}_{n,c}^{(1,l)}(0.6,0.5) - \hat{\theta}^{(2,l)}_{0.6,0.5}\right]^2.$$

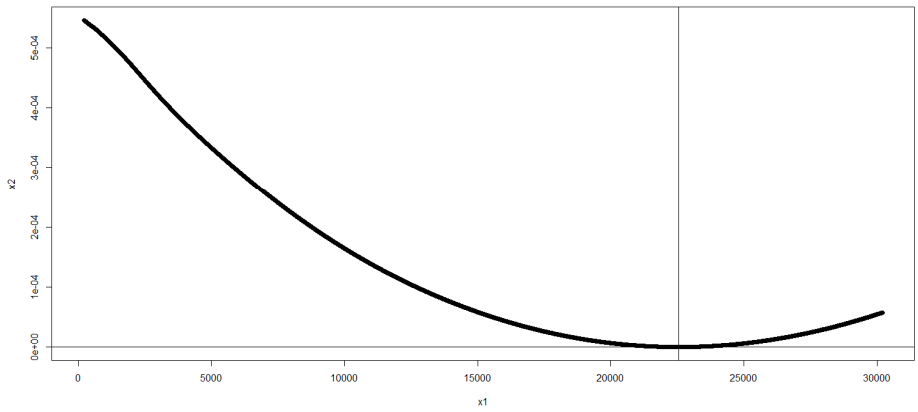
We conducted the steps 1°– 4° above for our data - composed of the equivalised disposable incomes of 13057 Polish households from the year 2003 - for the cases when: $L = 30,40,50,60$. As a result, CV_c reached its minimum for: $c = 23174.6$ - if $L = 30$, $c = 22553.6$ - if $L = 40$, $c = 23240.6$ - if $L = 50$, $c = 24634.1$ - if $L = 60$. That can be illustrated in the figures below.

Figure 1. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 30; c_{min} = 23174.6



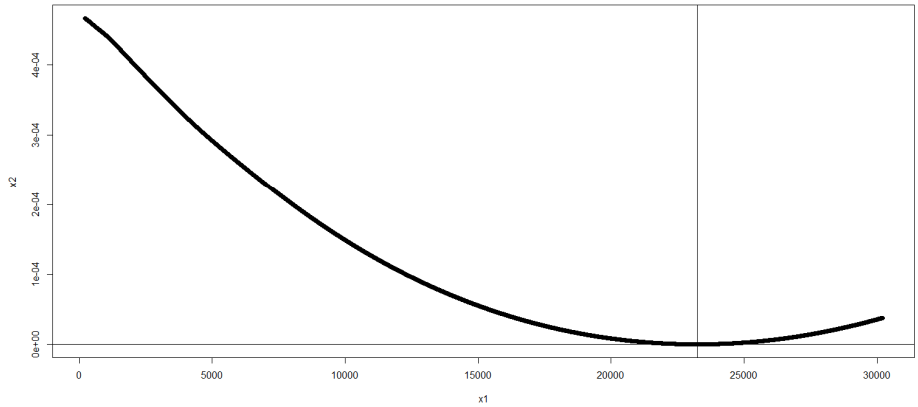
Source: own elaboration

Figure 2. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 40; c_{min} = 22553.6



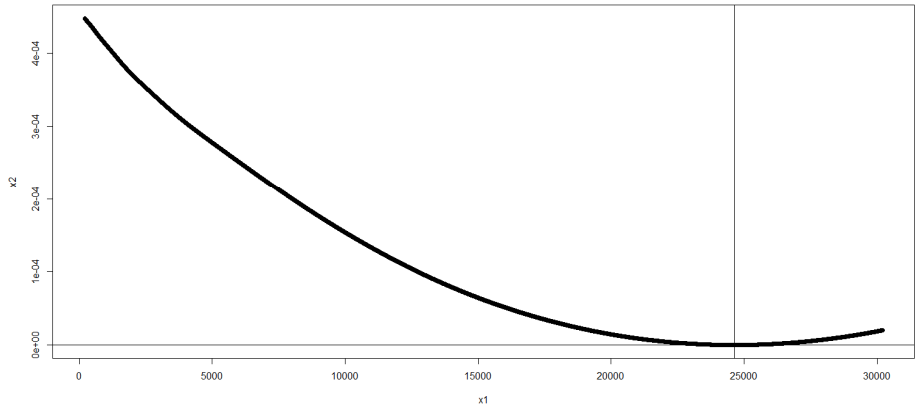
Source: own elaboration

Figure 3. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 50; c_{min} = 23240.6



Source: own elaboration

Figure 4. The choice of c in bandwidth selection, obtained by minimizing MSE for L = 60; c_{min} = 24634.1



Source: own elaboration

Consequently, since the values for c_{min} were computed based on the subsamples of size $\lceil n/2 \rceil = 6529$, then - according to the recommendation in [Luo and Qin 2017] - we applied the following values for bandwidths: $h = 23174.6 \cdot (6529)^{-1/3} = 1239.929$, or $h = 22553.6 \cdot (6529)^{-1/3} = 1206.703$, or $h = 23240.6 \cdot (6529)^{-1/3} = 1243.46$, or $h = 24634.1 \cdot (6529)^{-1/3} = 1318.017$, for: $L = 30,40,50,60$, respectively. In Table 1 below, we collected the 95% (0.95-level) realizations of the kernel-based confidence intervals $(l_2; u_2)$-$(l_6; u_6)$ for $\theta_{\alpha,\beta}$. We

limited ourselves to the case when $\alpha = 0.6$ and $\beta = 0.5$, i.e. to evaluation of the realizations of confidence intervalas for *ARPR*. The values of B from this table denote the sizes of bootstrap copies used in computations of the bootstrap kernel-based confidence intervals $(l_5; u_5)$-$(l_6; u_6)$.

Table 1. The 95% realizations of the selected kernel-based confidence intervals for ARPR (the bootstrap kernel-based realizations $(l_5; u_5)$-$(l_6; u_6)$ were obtained for: B = 500 - if: L = 30, 40, 50, or B = 1000 - if L = 60)

I \ *L*	30	40	50	60
$(l_2; u_2)$	0.1481-0.1599	0.1443-0.1560	0.1486-0.1603	0.1572-0.1689
$(l_3; u_3)$	≤ 0.1548	≤ 0.1520	≤ 0.1544	≤ 0.1604
$(l_4; u_4)$	0.1540-0.1541	0.1501-0.1502	0.1544-0.1545	0.1630-0.1631
$(l_5; u_5)$	0.1433-0.1647	0.1398-0.1605	0.1439-0.1650	0.1540-0.1721
$(l_6; u_6)$	0.1428-0.1642	0.1393-0.1601	0.1435-0.1646	0.1539-0.1720

Source: own elaboration

SUMMARY

The main goal of our study was to apply some selected estimation procedures in evaluation of the *Low-Income Proportion* (*LIP*) and *At-Risk-of-Poverty Rate* (*ARPR*) measures. We primarily focused on interval estimation of the *ARPR* index and computed the realizations of the corresponding confidence intervals for dataset consisting of 13057 observations referring to an equivalised disposable income of households in Poland from the year 2003, gained from [Statistical Publishing Establishment, Warsaw, 2004]. As recommended in [Luo and Qin 2017], we mainly considered the kernel-based confidence intervals. It may be easily seen that, depending on the number L (which is the number of random splits into the training and test samples), the ranges of lower/upper limits of the selected confidence intervals are (with an exception of the realizations of $(l_3; u_3)$) as follows: (i) if $L = 30$, then the lower limits of the obtained realizations range from 0.1428 to 0.1540 and its upper limits range from 0.1541 to 0.1647, (ii) if $L = 40$, then the lower limits of the obtained realizations range from 0.1393 to 0.1501 and its upper limits range from 0.1502 to 0.1605, (iii) if $L = 50$, then the lower limits of the obtained realizations range from 0.1435 to 0.1544 and its upper limits range from 0.1545 to 0.1650, (iv) if $L = 60$, then the lower limits of the obtained realizations range from 0.1539 to 0.1630 and its upper limits from 0.1631 to 0.1721. Furthermore, comparing the obtained realizations of the kernel-based confidence intervals $(l_2; u_2)$ and $(l_4; u_4)$-$(l_6; u_6)$ for *ARPR* with its empirical estimate $\widehat{ARPR} = \hat{\theta}_{0.6,0.5} = 0.1595$, we observe that: three out of four computed realizations of $(l_2; u_2)$ contain $\widehat{ARPR}$, none of four computed realizations of $(l_4; u_4)$ contains $\widehat{ARPR}$, all of four computed realizations of $(l_5; u_5)$ contain $\widehat{ARPR}$, and also all of four computed realizations of $(l_6; u_6)$ contain $\widehat{ARPR}$. Thus, it seems reasonable to limit our

attention to interpretation of the obtained realizations of confidence intervals $(l_2; u_2)$ and $(l_5; u_5)$-$(l_6; u_6)$. If we do so, then - taking into account all of the considered numbers of iterations L - we observe that the lower limits of these realizations range from 0.1393 to 0.1572, whereas the upper ones are between 0.1560 and 0.1721. Obviously, if we average the minimum and maximum of the considered lower limits range, we get an average 0.148 and by averaging the minimum and maximum of the considered upper limits range, we have an average 0.164. Thus, roughly speaking, we may state that, based on the obtained 95% realizations of selected confidence intervals for $ARPR$, the $ARPR$ measure ranges, on average, between 0.148 and 0.164. In other words, we may claim that with high probability, the percentage of Polish households with the equivalised disposable incomes not exceeding 60% of the whole population median amounted between 15% and 16% in the year 2003. That was the last year before Poland's entry to the European Union and a natural question arises, whether $ARPR$ has changed throughout the years since Poland has become a member of the EU. An answer to this question has been - at least partially - delivered in the report from [Eurostat Statistics Explained 2020] (which is an electronic publishing platform containing Eurostat's statistical information). It shows that in 2019, the $ARPR$ measure in Poland was between 15% and 16% - approximately the same in range that we obtained for the year 2003 using the chosen procedures of interval estimation. Thus, we may conclude that the percentages of low earners in Poland in the years 2003 and 2019 were roughly similar. It may seem slightly unusual that a reliable poverty measure was the same both in the year directly preceding Poland's accession to the EU and 15 years after that, especially since the results presented in [Eurostat Statistics Explained 2020] were computed for data including social transfers. It would be vital to study this issue in our further research. Also, it would be worthwhile to estimate $ARPR$ for dataset covering a period when various Coronavirus lockdown rules have been introduced. Directly before this period, the estimated value of this index for Poland, amounting to between 15% and 16%, has ranked Poland, in the group of EU countries with the $ARPR$ measure below the EU average of 21.1% and we think it would be desirable to check whether it is also the case after almost two turbulent years of SARS-CoV-2 era.

REFERENCES

Bowman A. W., Hall P., Prvan T. (1998) Cross-Validation for the Smoothing of Distribution Functions. Biometrika, 85, 799-808.

Efron B. (1979) Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7, 1-26.

Eurostat Statistics Explained (2020) File: At-Risk-of-Poverty Rate and At-Risk-of-Poverty Threshold 2019 LCIE20.png, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:At-risk-of-poverty_rate_and_at-risk-of-poverty_threshold,_2019_LCIE20.png.

Eurostat (Your Key to European Statistics) (2020) https://ec.europa.eu/eurostat/web/products-datasets/-/tespm010.

Falk M. (1983) Relative Efficiency and Deficiency of Kernel Type Estimators of Smooth Distribution Functions. Statistica Neerlandica, 37, 73-83.

Falk M. (1985) Asymptotic Normality of the Kernel Quantile Estimator. The Annals of Statistics, 13, 428-433.

Gong Y., Peng Y., Qi Y. C. (2010) Smoothed Jackknife Empirical Likelihood Method for ROC Curve. Journal of Multivariate Analysis, 101, 1520-1531.

Jing B., Yuan J., Zhou W. (2009) Jackknife Empirical Likelihood. Journal of American Statistical Association, 104, 1224-1232.

Li Z., Gong Y., Peng L. (2011) Empirical Likelihood Intervals for Conditional Value-at-Risk in Heteroscedastic Regression Models. Scandinavian Journal of Statistics, 38, 781-787.

Luo S., Qin G. (2017) New Non-Parametric Inferences for Low-Income Proportions. Annals of the Institute of Statistical Mathematics, 69(3), 599-626.

Preston I. (1995) Sampling Distributions of Relative Poverty Statistics. Applied Statistics, 44, 91-99 [Correction, 45, 399 (1996)].

Statistical Publishing Establishment, Warsaw (2004) Household Budget Surveys in 2003.

Tukey J. W. (1958) Bias and Confidence is Not-Quite Large Sample. The Annals of Mathematical Statistics, 29, 614.

Wei Z., Wen S., Zhu L. (2009) Empirical Likelihood-Based Evaluations of Value at Risk Models. Science in China Series A, Mathematics, 52, 1995-2006.

Wei Z., Zhu L. (2010) Evaluation of Value at Risk: An Empirical Likelihood Approach. Statistica Sinica, 20, 455-468.

Zieliński W. (2009a) A Nonparametric Confidence Interval for At-Risk-of-Poverty-Rate. Statistics in Transition, new series, 10 (3), 437-444.

Zieliński W. (2009b) A Nonparametric Confidence Interval for At-Risk-of-Poverty-Rate: Example of Application. Polish Journal of Environmental Studies, 18 (5B), 217-219.

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXII/1, 2021, s. 29 – 38

ZASTOSOWANIE BIBLIOTEKI ML.NET DO BADAŃ EKONOMICZNYCH

Waldemar Karwowski https://orcid.org/0000-0002-9988-0209
Instytut Informatyki Technicznej
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: waldemar_karwowski@sggw.edu.pl

Streszczenie: W artykule podjęto próbę oceny przydatności biblioteki ML.NET do badań w dziedzinie ekonomii. Przedstawiono możliwości wykorzystania uczenia maszynowego w aplikacjach i krótko omówiono dostępne bezpłatnie biblioteki programistyczne związane z uczeniem maszynowym. Omówiono funkcjonalność biblioteki ML.NET ze szczególnym uwzględnieniem przydatności do badań i zastosowań w dziedzinie ekonomii. W celu weryfikacji możliwości ML.NET do generacji modelu jak i wykorzystania go do prognozy, utworzono prostą aplikację w języku C# prognozującą poziom inflacji na podstawie zgromadzonych danych.

Słowa kluczowe: uczenie maszynowe, ML.NET, zastosowanie informatyki w ekonomii

JEL classification: C80

WSTĘP

W ostatnich latach na całym świecie bardzo popularne stało się wykorzystanie technik uczenia maszynowego (ang. machine learning). Techniki te są integralną częścią sztucznej inteligencji najogólniej rozumianej jako zdolność systemów informatycznych do rozwiązywania zadań, które gdy wykonywane są przez człowieka wymagają inteligencji. W szczególności dotyczy to zadań, w których interpretujemy nowe dane na podstawie wcześniej zgromadzonej wiedzy, na ogół wiedza ta zawarta jest w przygotowanym modelu. Uczenie maszynowe bazuje głównie na metodach matematycznych, przede wszystkim metodach statystycznych oraz nowoczesnych narzędziach przetwarzania informacji

https://doi.org/10.22630/MIBE.2021.22.1.3

takich jak sieci neuronowe. Bardzo często przygotowanie danych w celu utworzenia modelu wymaga współpracy człowieka, na przykład oznaczenia przypadków jako pozytywne lub negatywne. Rozwój i szerokie zastosowanie uczenia maszynowego w dużym stopniu stały się możliwe dzięki wzrastającej sile obliczeniowej komputerów. Do realizacji zadań uczenia maszynowego konieczne są nie tylko wydajne komputery ale przede wszystkim specjalistyczne oprogramowanie. Powstało wiele narzędzi oraz bibliotek pomagających zarówno w tworzeniu modeli jak i korzystaniu z nich. Gotowe narzędzia przeznaczone są zasadniczo dla wszystkich użytkowników, oparte są one przede wszystkim na interfejsie graficznym. Biblioteki przeznaczone są głównie dla programistów, ułatwiają tworzenie zintegrowanych rozwiązań, które oprócz innej funkcjonalności umożliwiają tworzenie, trenowanie i wykorzystanie modeli. Obecnie, ze względu na dużą popularność uczenia maszynowego, zwiększa się liczba dostępnych rozwiązań przeznaczonych dla różnych systemów operacyjnych i różnych języków programowania. W dużym stopniu są to biblioteki możliwe do bezpłatnego wykorzystania, często z otwartym kodem źródłowym. Najbardziej znane aplikacje komercyjne to IBM Watson Studio, IBM SPSS Modeler, Mathematica, MATLAB, SAS Enterprise Miner czy Oracle Data Mining; oferują one możliwość pracy z interfejsem graficznym oraz udostępniają interfejs programistyczny. Obecnie wiele aplikacji jest udostępniane jako usługi w chmurze, najbardziej znane to Amazon Machine Learning, Azure Machine Learning czy Oracle AI Platform Cloud Service. Wśród aplikacji bezpłatnych są zarówno aplikacje z interfejsem graficznym takie jak Deep Learning Studio, które nie wymagają nic więcej niż wklejenie tabeli z danymi i użycia myszy komputerowej. Przede wszystkim są to jednak biblioteki programistyczne. Obecnie najpopularniejszym językiem programowania w dziedzinie sztucznej inteligencji jest Python i najczęściej używane biblioteki z implementacją algorytmów uczenia maszynowego są napisane w Pythonie, który umożliwia pracę interaktywną i jednocześnie dostarcza dodatkowe biblioteki np. do obliczeń matematycznych takie jak NumPy, SciPy, oraz do obróbki i wizualizacji danych takie jak Pandas i Matplotlib. Należy jednak zaznaczyć, że krytyczne fragmenty oprogramowania są w rzeczywistości napisane w języku C/C++ a tylko interfejs użytkownika jest napisany w Pythonie. Najpopularniejsze biblioteki do uczenia maszynowego to: TensorFlow i PyTorch, ponadto Keras (wykorzystuje TensorFlow) oraz SciKit Learn (bazuje na NumPy i SciPy) [Gevorkyan i in. 2019]. Ponieważ Python jest językiem skryptowym, to bardzo dobrze nadaje się do pracy interaktywnej natomiast nieco gorzej wymienione biblioteki integrują się ze środowiskami programistycznymi takimi jak Java czy .NET. Z jednej strony tworzone są interfejsy programistyczne w Javie czy C# do wymienionych wyżej bibliotek z drugiej strony tworzone są dedykowane biblioteki dla danego języka programowania. W Javie przykładem takiego rozwiązania jest Weka czy Java-ML, natomiast w środowisku .NET jest to ML.NET. ML.NET jest biblioteką dostępną bezpłatnie, rozwijaną bezpośrednio przez Microsoft od roku 2018 [Zeeshan i in. 2019; Capellman 2020]. Pierwsza

pełna wersja 1.0 została wydana w połowie roku 2019 i jest systematycznie rozwijana, obecnie dostępna jest wersja 1.7.

Celem artykułu jest zbadanie przydatności biblioteki ML.NET do zagadnień w dziedzinie ekonomii. Platforma .NET i język C# należą do jednych z najpopularniejszych środowisk programistycznych dla aplikacji pracujących w systemie Windows. Aplikacje te często dotyczą zagadnień związanych z ekonomią. Możliwość łatwego rozszerzenia takich aplikacji o możliwości uczenia maszynowego stwarza nową jakość i pozwala spełnić wymagania współczesnych użytkowników.

MOŻLIWOŚCI ML.NET

ML.NET powstał z potrzeby umożliwienia programistom .NET integracji aplikacji z algorytmami uczenia maszynowego. ML.NET daje możliwość dodawania uczenia maszynowego do aplikacji .NET w scenariuszach online lub offline. Microsoft już wcześniej udostępniał usługi związane z uczeniem maszynowym na komercyjnej platformie Microsoft Azure, natomiast ML.NET jest narzędziem bezpłatnym o otwartym kodzie źródłowym. Jednym z jego założeń jest unifikacja procesów rozwojowych modeli oraz aplikacji, w których maja one być wykorzystane. Biblioteka uporządkowana jest w formie przestrzeni nazw: główna przestrzeń ML zawiera podstawowe klasy i metody rozszerzające, ML.Data zawiera komponenty do ładowania i zapisu danych, definicji schematów danych oraz metryk dla modeli; ML.Transforms zawiera komponenty do transformacji danych; ML.Trainers zawiera trenery, parametry modeli i narzędzia. Dodatkowo w ML.Calibrators zawarte są metody, które obliczają prawdopodobieństwa z wyników dla klasyfikatorów binarnych. Schemat korzystania z biblioteki jest następujący: zbieranie, ładowanie i przekształcanie danych treningowych, określanie potoku operacji przede wszystkim algorytmu uczenia maszynowego, trenowania modelu, ocena modelu, zapisanie modelu do wykorzystania w aplikacji do tworzenia prognoz. Po etapie oceny modelu jest możliwa korekta danych i ponowne trenowanie modelu. Najważniejszym elementem w ML.NET jest model uczenia maszynowego. Model określa kroki potrzebne do przekształcenia danych wejściowych w prognozę. Dzięki ML.NET można trenować model niestandardowy, określając algorytm lub importować wstępnie wytrenowane modele TensorFlow i ONNX.

Centralnym elementem w programie korzystającym z biblioteki ML.NET jest obiekt klasy MLContext, jest to singleton koordynujący wszystkie pozostałe elementy, który zapewnia sposoby tworzenia komponentów do przygotowania danych, cech, trenowanie, proces predykcji i ewaluacji modelu.

W przestrzeni ML klasa DataOperationsCatalog jest używana do tworzenia składników, które działają na danych, ale nie są częścią potoku uczenia modelu. Zawiera metody do ładowania, zapisywania, buforowania, filtrowania, mieszania i dzielenia danych. Możliwe jest wczytywanie danych z plików tekstowych

w różnych formatach a także bezpośrednio z większości popularnych systemów bazodanowych. Oprócz tego dostępne są filtry, pozwalają one na wybranie ze zbioru jedynie rekordów spełniających dany warunek. Mamy trzy zasadnicze metody filtracji: FilterRowsByMissingValue - odrzuca wiersze, w których wybrana kolumna nie ma wartości; FilterRowsByColumn - wybiera tylko rekordy, w których wartość wybranej kolumny mieści się w ustalonym przedziale oraz FilterRowsByKeyColumnFraction – wybiera rekordy operując na kluczach. Przygotowanie i wybieranie wierszy danych możliwe jest także dzięki metodom: SkipRaws TakeRows, ShuffleRows. Z kolei CrossValidationSplit dzieli zestaw danych na zestaw treningowy i zestaw testowy, podobnie jak TrainTestSplit.

W klasie TransformsCatalog są metody, które pozwalają na różnego rodzaju transformacje i zmiany wykonywane na danych. Surowe dane muszą być przygotowane lub wstępnie przetworzone, zanim będą mogły zostać użyte do znalezienia parametrów modelu. Mapowanie i grupowanie kolumn jest możliwe dzięki metodom: Concatenate - łączenia jednej lub więcej kolumn wejściowych w nową kolumnę wyjściową, CopyColumns, DropColumns, SelectColumns. Metody do oznaczania i uzupełniania brakujących wartości to: IndicateMissingValues - tworzy nową kolumnę wyjściową logiczną, której wartość jest true, gdy brakuje wartości w kolumnie wejściowej; ReplaceMissingValues - tworzy nową kolumnę wyjściową, której wartość jest ustawiona na wartość domyślną, jeśli wartości nie ma w kolumnie wejściowej. Inne metody umożliwiają przekształcenia danych, na przykład: NormalizeMeanVariance oblicza średnią i wariancję; inne operacje do normalizacji to: NormalizeMeanVariance, NormalizeLogMeanVariance. NormalizeLpNorm - skaluje wektory wejściowe według ich Lp-norm, NormalizeMinMax - skaluje dane w zakresie między minimalną i maksymalną wartością w danych treningowych.

Metody z grupy TransformsCatalog.ConversionTransforms pozwalają konwertować dane np. z wartości tekstowych na reprezentację liczbową, są to: ConvertType - konwertuje typ kolumny wejściowej na nowy typ; Hash mieszanie (haszowanie) wartości w kolumnach wejściowych oraz grupa metod do mapowania na klucze i odwrotnie np. MapValue mapuje wartości do kluczy (kategorii) na podstawie dostarczonego słownika mapowań; MapKeyToVector konwertuje klucze z powrotem na wektory oryginalnych wartości. W grupie TransformsCatalog.CategoricalTransforms mamy OneHotEncoding kodowanie 1 z n jednej lub więcej kolumn tekstu na wektory oraz OneHotHashEncoding kodowanie w oparciu o mieszanie (haszowanie).

W przestrzeni nazw ML.Transforms.TimeSeries umieszczono narzędzia do analizy szeregów czasowych. Dodatkowo dostępne są klasy do transformacji tekstu - TextCatalog oraz przekształcenia obrazu - ImageEstimatorsCatalog oraz klasy z przestrzeni ML.Vision i narzędzia z przestrzeni ML.Transforms.Image.

Cała rodzina klas jest dedykowana podstawowym funkcjonalnościom: AnomalyDetectionCatalog i TimeSeriesCatalog umożliwiają analizę szeregów czasowych, BinaryClassificationCatalog - klasyfikację binarną, ClusteringCatalog -

klasteryzację, MulticlassClassificationCatalog - klasyfikację wieloklasową, RegressionCatalog – regresję. Do każdej z tych klas mamy odpowiednią klasę trenerów. Do trenowania wykrywania anomalii szeregów czasowych mamy metodę RandomizedPca - analizę głównych składowych. Do trenowania klasteryzacji mamy tylko metodę KMeans opartą na algorytmie k-średnich. Do trenowania klasyfikacji binarnej mamy do dyspozycji wiele metod: LbfgsLogisticRegression model liniowej regresji logistycznej trenowany przy pomocy zmodyfikowanego algorytmu Broydena-Fletchera-Goldfarba-Shanno; dwie kolejne metody korzystają z maszyny wektorów podpierających, LdSvm bazuje na modelu Local Deep SVM natomiast LinearSvm, oparta jest na liniowym modelu klasyfikacji binarnej wytrenowanym na danych etykietowanych wartościami logicznymi; SdcaLogisticRegressionBinary oraz SdcaNonCalibratedBinary używają liniowego modelu klasyfikacji opartego o algorytm Stochastic Dual Coordinate Ascent; SgdCalibrated oraz SgdNonCalibrated używają liniowego modelu klasyfikacji opartego o algorytm stochastic gradient descent; FastForestBinary wykorzystuje model klasyfikacji binarnej oparty na drzewie decyzyjnym; GamBinary wykorzystuje uogólnione modele addytywne; FieldAwareFactorizationMachine, wykorzystuje maszynę faktoryzacji wytrenowaną na danych etykietowanych wartościami logicznymi; LightGbmBinary używa wzmocnienia gradientowego dla drzew decyzyjnych; AveragedPerceptron wykorzystuje perceptron dla danych etykietowanych wartościami logicznymi; Prior, bazuje na wcześniejszym modelu klasyfikacji binarnej.

Do trenowania klasyfikacji wieloklasowej również dostępnych jest wiele metod. LightGbm, LbfgsMaximumEntropyMulticlass, NaiveBayes, OneVersusAll, PairwiseCoupling, SdcaMaximumEntropy, SdcaNonCalibrated, oraz ImageClassification, która wykorzystuje głęboką sieć neuronową (DNN) do klasyfikowania obrazów.

Do trenowania regresji można wykorzystać: LightGbm, Ols opartą na metodzie najmniejszych kwadratów (Ordinary least squares), LbfgsPoissonRegression, OnlineGradientDescent, Sdca, FastForest, FastTree oraz Gam.

ML.NET dla każdego modelu oferuje różne zestawy metryk do oceny modelu. Wymienimy głównie nazwy, bez szczegółowych wyjaśnień, są one znaczące a termin angielski nie nastręcza problemu. Dla wykrywania anomalii mamy dwie metryki: AreaUnderRocCurve, DetectionRateAtFalsePositiveCount - stosunek poprawnie zidentyfikowanych anomalii przy określonej liczbie fałszywych trafień. Dla klasteryzacji są trzy metryki: AverageDistance, DaviesBouldinIndex oraz NormalizedMutualInformation. Dla klasyfikacji binarnej są to: Accuracy, AreaUnderPrecisionRecallCurve, AreaUnderRocCurve, ConfusionMatrix, F1Score, NegativePrecision, NegativeRecall, PositivePrecision oraz PositiveRecall. Dla klasyfikacji wielokryterialnej są to ConfusionMatrix, LogLoss, LogLossReduction, MacroAccuracy, MicroAccuracy, PerClassLogLoss,

TopKAccuracy, TopKAccuracyForAllK oraz TopKPredictionCount. Dla modelu regresji są to: LossFunction, MeanAbsoluteError, MeanSquaredError, RootMeanSquaredError. Dodatkowo możliwa jest walidacja krzyżowa, metoda CrossValidate dzieli najpierw zbiór na określoną liczbę podzdbiorów, model jest uczony i ewaluowany dla każdego z nich. Metoda zwraca kolekcję obiektów, z których każdy zawiera model oraz wyznaczone metryki.

WYNIKI EKSPERYMENTU

W celu weryfikacji możliwości ML.NET do budowy modelu jak i wykorzystania go do prognozy, utworzono prostą aplikację w języku C# prognozującą poziom inflacji na podstawie zgromadzonych danych. Celem badania było określenie jak trendy inflacyjne innych państw wpływają na inflację w Polsce przy wykorzystaniu modelu regresji liniowej. Aplikacja trenuje model regresji, na podstawie danych, a następnie ocenia jakość powstałego modelu. Poziom inflacji w Polsce został potraktowany jako zmienna zależna. Utworzony model jest weryfikowany na wybranych danych dla losowo wybranego miesiąca. W badaniu porównano dwa algorytmy regresji: Poissona (LbfgsPoissonRegression) oraz oparty na metodzie optymalizacyjnej Stochastic Dual Coordinate Ascent (SDCA). Ocena jakości modelu polega na wyznaczeniu współczynnika determinacji R^2 dla danych testowych oraz błędu średniokwadratowego. W celu ustalenia zależności zbudowano kolejno modele dla wybranych grup państw.

Algorytm regresji Poissona w klasie LbfgsPoissonRegressionTrainer został zaimplementowany przy pomocy techniki optymalizacji opartej na metodzie ograniczonej pamięci Broydena-Fletchera-Goldfarba-Shanno (L-BFGS) [Liu i in. 1989]. L-BFGS jest metodą quasi-newtonowską, która zastępuje kosztowne obliczenia Hesjanu aproksymacją, ale nadal charakteryzuje się szybkim tempem zbieżności, podobnie jak metoda, w której obliczana jest pełna macierz Hessego [Nash i in. 1991]. Metoda ta nadaje się do zagadnień z wektorami cech o dużych wymiarach. Pierwsze implementacje metody były napisane w Fortranie [Zhu i in. 1997]. Efektywność metody może być poprawiana dzięki regularyzacji [Hardik i in. 2021], implementacja w bibliotece ML.NET pozwala na regularyzację przy pomocy kombinacji norm L1 i L2.

Stochastic Dual Coordinate Ascent to algorytm optymalizacji, w którym maksymalizujemy funkcję dla problemu dualnego. Algorytm sukcesywnie maksymalizuje wzdłuż kierunków współrzędnych. W każdej iteracji algorytm określa współrzędną za pomocą reguły losowego wyboru współrzędnych, a następnie maksymalizuje odpowiednią ustalając wszystkie inne współrzędne [Shalev-Shwartz i in. 2013]. Jest to nowoczesna technika optymalizacji wypukłych funkcji celu, algorytm ten można skalować, ponieważ jest to algorytm uczenia strumieniowego [Tran i in. 2015]. Metoda ma wiele modyfikacji, które poprawiają wydajność [Shalev-Shwartz i in. 2016].

Do analizy wybrano dane z 29 państw łącznie z Polską. Wybrano następujące państwa Wielka Brytania (GB), Irlandia (IE), Niemcy (DE), Francja (FR), Hiszpania (ES), Portugalia (PT), Austria (AT), Włochy (IT), Czechy (CZ), Słowacja (SK), Grecja (GR), Białoruś (BY), Ukraina (UA), Norwegia (NO), Szwecja (SE), Węgry (HU), Finlandia (FI), Rosja (RU), Stany Zjednoczone (US), Chiny (CN), Japonia (JP), Indie (IN), Tajwan (TW), Korea Południowa (KR), ZEA (AE), Katar (QA), Kuwejt (KW), Turcja (TR) oraz Polska (PL). Wszystkie państwa mają kontakty handlowe z Polską i mają wpływ na polską gospodarkę a więc także poziom inflacji. Dla każdego z państw ustalono poziom inflacji dla kolejnych miesięcy 2021 roku (grudzień - prognoza). Dane w trzech grupach przedstawiono w tabeli 1.

Tabela 1. Wartości wskaźnika inflacji w roku 2021 w wybranych państwach

Mies.	GB	IE	DE	FR	ES	PT	AT	IT	CZ	SK
1	0,7%	-0,2%	1,0%	0,6%	0,5%	0,3%	0,8%	0,4%	2,2%	0,7%
2	0,4%	-0,4%	1,3%	0,6%	0,0%	0,5%	1,2%	0,6%	2,1%	0,9%
3	0,7%	0,0%	1,7%	1,1%	1,3%	0,5%	2,0%	0,8%	2,3%	1,4%
4	1,5%	1,1%	2,0%	1,2%	2,2%	0,6%	1,9%	1,1%	3,1%	1,6%
5	2,1%	1,7%	2,5%	1,4%	2,7%	1,2%	2,8%	1,3%	2,9%	2,2%
6	2,5%	1,6%	2,3%	1,5%	2,7%	0,5%	2,8%	1,3%	2,8%	2,9%
7	2,0%	2,2%	3,8%	1,2%	2,9%	1,5%	2,9%	1,9%	3,4%	3,3%
8	3,2%	2,8%	3,9%	1,9%	3,3%	1,5%	3,2%	2,0%	4,1%	3,8%
9	3,1%	3,7%	4,1%	2,2%	4,0%	1,5%	3,3%	2,5%	4,9%	4,6%
10	4,2%	5,1%	4,5%	2,6%	5,4%	1,8%	3,7%	3,0%	5,8%	5,1%
11	5,1%	5,3%	5,2%	2,8%	5,5%	2,6%	4,3%	3,7%	6,0%	5,6%
12	5,4%	5,3%	5,4%	3,0%	5,7%	2,6%	4,3%	3,9%	6,7%	5,6%

Mies.	GR	BY	UA	NO	SE	HU	FI	RU	US	CN
1	-2,0%	7,7%	5,0%	2,5%	1,6%	2,7%	0,9%	5,2%	1,4%	-0,3%
2	-1,3%	8,7%	6,1%	3,3%	1,4%	3,1%	0,9%	5,7%	1,7%	-0,2%
3	-1,6%	8,5%	7,5%	3,1%	1,7%	3,7%	1,3%	5,8%	2,6%	0,4%
4	-0,3%	8,7%	8,5%	3,0%	2,2%	5,1%	2,1%	5,5%	4,2%	0,9%
5	0,1%	9,4%	8,4%	2,7%	1,8%	5,1%	2,2%	6,0%	5,0%	1,3%
6	1,0%	9,9%	9,5%	2,9%	1,3%	5,3%	2,0%	6,5%	5,4%	1,1%
7	1,4%	9,8%	9,5%	3,0%	1,4%	4,6%	1,9%	6,5%	5,4%	1,0%
8	1,9%	9,8%	10,2%	3,4%	2,1%	4,9%	2,2%	6,7%	5,3%	0,8%
9	2,2%	10,2%	10,2%	4,1%	2,5%	5,5%	2,5%	7,4%	5,4%	0,7%
10	3,4%	10,5%	11,0%	3,5%	2,8%	6,5%	3,2%	8,1%	6,2%	1,5%
11	4,8%	10,3%	10,9%	5,1%	3,3%	7,4%	3,7%	8,4%	6,8%	2,3%
12	5,0%	10,8%	10,3%	4,8%	3,3%	7,2%	3,7%	8,4%	6,9%	2,6%

Mies.	JP	IN	TW	KR.	AE	QA	KW	TR	PL
1	-1,2%	4,1%	-0,2%	0,6%	-2,4%	-1,3%	2,2%	15,0%	2,6%
2	-0,7%	5,0%	1,4%	1,1%	-2,1%	-1,4%	2,0%	15,6%	2,4%
3	-0,5%	5,5%	1,2%	1,5%	-1,9%	-0,3%	2,5%	16,2%	3,2%
4	-0,4%	4,2%	2,1%	2,3%	-2,0%	1,0%	2,8%	17,1%	4,3%
5	-1,1%	6,3%	2,4%	2,6%	-1,1%	2,5%	3,0%	16,6%	4,7%
6	-0,8%	6,3%	1,8%	2,4%	-0,5%	2,0%	3,0%	17,5%	4,4%
7	-0,5%	5,6%	1,9%	2,6%	-0,4%	3,1%	3,0%	19,0%	5,0%
8	-0,3%	5,3%	2,3%	2,6%	-0,5%	3,0%	3,2%	19,3%	5,5%
9	-0,4%	4,4%	2,6%	2,5%	0,0%	2,7%	3,1%	19,6%	5,9%
10	0,2%	4,5%	2,5%	3,2%	0,6%	4,3%	3,2%	19,9%	6,8%
11	0,1%	4,9%	2,8%	3,7%	1,2%	6,1%	3,5%	21,3%	7,4%
12	0,4%	4,9%	2,9%	3,7%	1,3%	6,1%	3,5%	23%%	8,1%

Źródło: opracowanie własne na podstawie https://tradingeconomics.com

W tabeli 2 przedstawiono wyniki eksperymentu.

Tabela 2. Wynik dla wszystkich wybranych państw

	Metoda	Współczynnik determinacji	Pierwiastek błędu średniokwadratowego	Estymowana inflacja w Polsce	Rzeczywista inflacja w Polsce
28 państw	Regresja Poissona	0,98	0,22	7,6073%	7,4%
	SDCA	1,0	0,01	7,4135%	7,4%
Strefa Euro	Regresja Poissona	0,97	0,3	7,7047%	7,4%
	SDCA	1,0	0,11	7,5996%	7,4%
Strefa Euro + GB	Regresja Poissona	0,97	0,28	7,6734%	7,4%
	SDCA	1,0	0,1	7,6003%	7,4%
UE	Regresja Poissona	0,97	0,28	7,6806%	7,4%
	SDCA	1,0	0,09	7,5258%	7,4%
Silne gospodarki	Regresja Poissona	0,97	0,28	7,4634%	7,4%
	SDCA	1,0	0,08	7,5928%	7,4%
Silne militarnie	Regresja Poissona	0,97	0,32	7,522%	7,4%
	SDCA	0.98	0,22	7,621%	7,4%
Sąsiedzi z UE	Regresja Poissona	0,95	0,39	7,6054%	7,4%
	SDCA	0.82	0,75	8,1375%	7,4%
Zatoka Perska	Regresja Poissona	0,94	0,42	8,0765%	7,4%
	SDCA	0.96	0,33	7,6698%	7,4%

Źródło: opracowanie własne

Kolejno uwzględniono wpływ wszystkich 28 państw na inflację w Polsce. Następnie wybrano podgrupę państw ze strefy Euro (Irlandia, Niemcy, Francja,

Hiszpania, Portugalia, Austria, Włochy, Słowacja, Grecja, Finlandia) oraz strefę Euro i Zjednoczone Królestwo. Kolejną czwartą, badaną podgrupą były kraje Unii Europejskiej (Irlandia, Niemcy, Francja, Hiszpania, Portugalia, Austria, Włochy, Czechy, Słowacja, Grecja, Szwecja, Węgry, Finlandia). Piątą grupą były państwa o dużym potencjale gospodarczym (Niemcy, Francja, Rosja, USA, Chiny, Japonia, Indie, Tajwan, Korea Południowa). Szóstą podgrupę stanowiły państwa o dużym potencjale militarnym (Rosja, USA, Chiny, Indie). Siódmą grupę stanowili sąsiedzi z Unii Europejskiej (Niemcy, Czechy, Słowacja). Na zakończenie sprawdzono grupę państw z rejonu Zatoki Perskiej (ZEA, Katar, Kuwejt).

PODSUMOWANIE

ML.NET umożliwia integrację algorytmów uczenia maszynowego z praktycznymi aplikacjami. Przygotowanie aplikacji w języku C# nie jest trudne, w prosty sposób można wygenerować model i zastosować go do predykcji. Wspomagające narzędzie Model Builder pozwalające na automatyczną generację modelu okazało się z kolei niezbyt przydatne. Model Builder tylko w szczególnych przypadkach generuje model, proces często kończy się błędem. W przypadku rozważanej regresji, podczas generacji sprawdzana jest kolejno każda dostępna metoda, i finalnie wybierana ta która ma najwyższy współczynnik determinacji. Niestety błąd którejkolwiek metody, spowodowany specyfiką danych (np. zbyt mało danych) powoduje przerwanie autogeneracji, nie ma możliwości predefiniowania tylko wybranych metod. Innym problemem jest kwestia wydajności. W przypadku niewielkiego zbioru danych wybranego w eksperymencie czas trenowania modelu jest krótki, jednakże w pracy [Kędziora i in. 2021] wyniki dla biblioteki ML.NET były znacznie gorsze niż dla TensorFlow.

BIBLIOGRAFIA

Ahmed Z. et al. (2019) Machine Learning at Microsoft with ML.NET. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, n. pag.

Capellman J. (2020) Hands-On Machine Learning with ML.NET: Getting Started with Microsoft ML.NET to Implement Popular Machine Learning Algorithms in C#. Packt Publishing, Birmingham.

Gevorkyan M. N., Demidova A. V., Demidova T. S., Sobolev A. A. (2019) Review and Comparative Analysis of Machine Learning Libraries for Machine Learning. Discrete and Continuous Models and Applied Computational Science, 27(4), 305-315, doi: 10.22363/2658-4670-2019-27-4-305-315.966-979.

Kędziora E. J. Maksim G. K. (2021) Analiza wydajności bibliotek uczenia maszynowego. Journal of Computer Sciences Institute, 20, 230-236.

Liu D. C., Nocedal J. (1989) On the Limited Memory BFGS Method for Large Scale Optimization. Mathematical Programming, 45, 503-528.

Nash S. G., Nocedal J. (1991) A Numerical Study of the Limited Memory BFGS Method and the Truncated-Newton Method for Large Scale Optimization. SIAM Journal of Optimization, 1, 358-372.

Shalev-Shwartz S., Zhang T. (2013) Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. Journal of Machine Learning Research, 14, 567-599.

Shalev-Shwartz S., Zhang T. (2016) Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization. Mathematical Programming, 155, 105-145.

Tankaria H., Sugimoto S., Yamashita N. (2021) A Regularized Limited Memory BFGS Method for Large-Scale Unconstrained Optimization and its Efficient Implementations. arXiv: Optimization and Control, n. pag.

Tran K., Hosseini S., Xiao L., Finley T., Bilenko M. (2015) Scaling up Stochastic Dual Coordinate Ascent. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, n. pag.

Zhu C. Byrd R. H., Lu P., Nocedal J. (1997) Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software, 23(4), 550-560, doi:10.1145/279232.279236.

https://dotnet.microsoft.com/en-us/apps/machinelearning-ai/ml-dotnet [dostęp: 20.11.2021].

APPLICATION OF THE ML.NET LIBRARY TO ECONOMIC ISSUES

Abstract: The aim of the paper was to attempt to validate the suitability of the ML.NET library for research in the field of economics. The possibilities of using machine learning in applications are presented, and the free programming libraries related to machine learning are briefly discussed. The functionality of the ML.NET library, with particular emphasis on its suitability for research and applications in the field of economics, was discussed. In order to verify the possibility of ML.NET to the model generation and its use for forecasting, a simple application in C# language to forecast the level of inflation based on the collected data was created.

Keywords: machine learning, ML.NET, application of computer science in economics

JEL classification: C80

QUANTITATIVE METHODS IN ECONOMICS
Volume XXII, No. 1, 2021, pp. 39 – 50

PRICING EUROPEAN OPTIONS IN THE HESTON AND THE DOUBLE HESTON MODELS

Arkadiusz Orzechowski[1] https://orcid.org/0000-0003-2872-189X
Institute of Risk and Financial Markets
Warsaw School of Economics – SGH, Poland
e-mail: aorzec@sgh.waw.pl

Abstract: Two models of pricing European options are presented and compared in this paper, i.e. the Heston model and the double Heston model. As the models belong to the class of stochastic volatility models, particular attention is paid to the way the characteristic functions and their inverse Fourier transforms are determined. The aim of the study is to investigate computational efficiency of pricing European calls. The method applied is based on the assumption that the prices of the derivatives are evaluated by means of Gauss-Kronrod quadrature.

Keywords: option pricing, the Heston model, the double Heston model, characteristic functions

JEL classification: C02, G13

INTRODUCTION

2021 has seen a significantly increased interest in the options market. According to Options Clearing Corporation (OCC), over 850 million option contracts were traded in December, 2021, a 12,4% growth as compared to December, 2020. Full year average daily cleared contract volume for 2021 was over 39 million, a 32,5% growth as compared to 2020[2]. In the period analyzed, equity options were the fastest to gain the market share (a 33,7% growth as compared to 2020), followed by index and ETF options (8,8% and 5,1% increase, respectively, compared to 2020).

[1] The views expressed in the article are the personal views of the author and do not express the official position of the institution in which he is employed.

[2] https://www.theocc.com/Newsroom/Press-Releases/2022/01-04-OCC-Clears-Record-Setting-9-93-Billion-Total [access: 04.01.2022].

https://doi.org/10.22630/MIBE.2021.22.1.4

In 2021 significant increase in trading options activity was reported among retail investors who were responsible for more than 25% of the total options trading volume. Easy access to commission-free online brokers was one of the main factors influencing the market structure. Such trading environment allowed retail investors to implement strategies based on execution of high-speed transactions.

As vast majority of retail option traders in 2021 were involved in basic call or put contracts, numerous different models of pricing options could be applied. Among the possible approaches to the valuation of options the most popular is the Black-Scholes model [Black & Scholes 1973]. The model has been widely used for years by both theoreticians and practitioners, because it enables fast obtainment of option prices. Unfortunately, this affects the accuracy of pricing, because the model is based on many unrealistic assumptions, e.g. constant variance of underlying asset's returns. As a result, many alternative approaches to the valuation of options have been proposed. They all can be divided into three categories, i.e. pure jump models, jump-diffusion models and stochastic volatility models (with all their variations).

In the pure jump models it is assumed that the price of the underlying asset changes in a discrete manner. As the discontinuities in the price movements of the underlying asset can be modelled in a number of ways many different methods of pricing options have been proposed, e.g. Madan et al. [1998], Madan et al. [1991], Carr et al. [2002], Eberlein et al. [1998].

An extension of the Black-Scholes model by possible discontinuities in the prices of the underlying asset allows for the valuation of options in the jump-diffusion models, e.g. Merton [1976], Kou [2002]. The construction of the models is based on the assumption stating that the continuous changes in the prices of the underlying asset can be occasionally disrupted by jumps. In these models both the frequency and the amplitude of the jumps can be modelled using different processes.

The stochastic volatility models assume the volatility of the underlying asset prices to be inconstant. The process responsible for the dynamics of the volatility can take different forms, depending on the model applied to the valuation of options [Heston 1993, Christoffersen et al. 2009]. It is worth noting that the stochastic volatility models are extended not only by changing the dynamics of the volatility, but also by introducing assumptions concerning the price dynamics of the underlying asset, e.g. Bates [2006].

The aim of the article is to compare the Heston model [Heston 1993] with the Christoffersen et al. model (further referred to as the double Heston model) [Christoffersen et al. 2009] in terms of computational speed, based on the example of pricing European calls. The article consists of several sections. In the first section two models of pricing European calls are formulated. In the second section the characteristic functions are applied to the process of option valuation. The third section includes the determination of the inverse Fourier transforms for the previously introduced characteristic functions. The speed of pricing European calls is also analyzed. Finally, the article has been summarized and major conclusions have been drawn.

THE HESTON AND THE DOUBLE HESTON MODELS

In this section, the Heston and the double Heston models are formulated and then applied to the valuation of the European calls. For this purpose the originally derived characteristic functions and their inverse Fourier transforms are applied.

The Heston model

The derivation procedure of the Heston model [Heston 1993] starts from two equations:

$$dS_t = \mu S_t dt + \sqrt{\sigma_t^2} S_t dW_{1,t}. \quad (1)$$

$$d\sigma_t^2 = \kappa(\theta - \sigma_t^2)dt + \upsilon\sqrt{\sigma_t^2} dW_{2,t}. \quad (2)$$

where: S_t denotes the spot price of the underlying asset at time t, σ_t^2 is the instantaneous variance, $\mu, \theta, \kappa, \upsilon$ are the constants associated with the drift, the long-term variance, the mean-reversion rate, and the volatility of the variance process, respectively. In the Heston model the Brownian motions W_1 and W_2 are correlated with a constant ρ.

Valuation of a European call is based on the following formula:

$$C^H(s_t, \sigma_t^2, t) = S_t P_1^H(s_t, \sigma_t^2, \tau) - e^{-r\tau} K P_2^H(s_t, \sigma_t^2, \tau). \quad (3)$$

where: $\tau = T - t$, r is the risk-free rate, K is the exercise price, $P_1^H(s_t, \sigma_t^2, \tau)$ and $P_2^H(s_t, \sigma_t^2, \tau)$ are unknown probabilities of expiring a European call in-the-money calculated as the inverse Fourier transform of characteristic function (for $j = 1,2$), i.e.:

$$P_j^H(s_t, \sigma_t^2, \tau) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Re\left(\frac{e^{-\mathbb{I}\xi lnK}\phi^{j,H}(\xi, s_t, \sigma_t^2)}{\mathbb{I}\xi}\right) d\xi. \quad (4)$$

where: $\Re(.)$ is the real part of the subintegral function, $\mathbb{I}$ is the imaginary unit of the complex number, $\phi^{j,H}(\xi, s_t, \sigma_t^2)$ is the characteristic function of $s_t = lnS_t$ (corresponding to $P_j^H(s_t, \sigma_t^2, \tau)$). The remaining notation is the same as previously introduced.

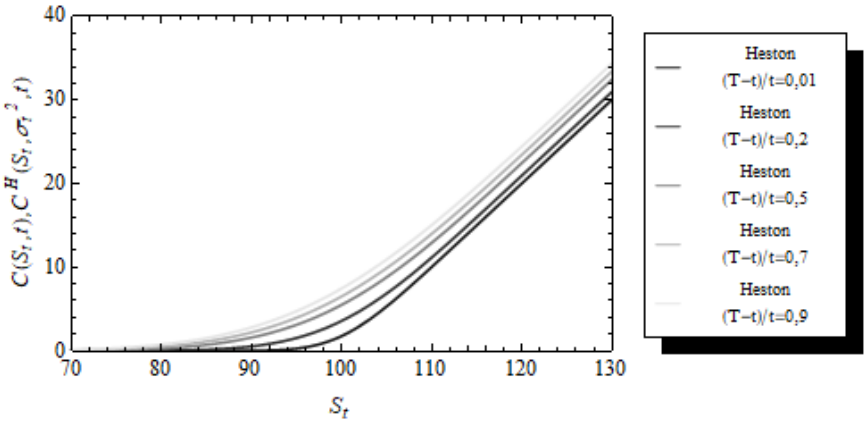
In the Heston model the general form of the characteristic function of s_t (corresponding to P_j^H, for $j = 1,2$) is expressed in the following form:

$$\phi^{j,H}(\xi, s_t, \sigma_t^2) = e^{C_j(\xi,\tau) + D_j(\xi,\tau)\sigma_t^2 + \mathbb{I}\xi s_t}. \quad (5)$$

where: $C_j(\xi,\tau) = r\mathbb{I}\xi\tau + \frac{a}{\upsilon^2}\left[(b_j - \upsilon\rho\mathbb{I}\xi + d_j)\tau - 2ln\left(\frac{1-g_j e^{d_j\tau}}{1-g_j}\right)\right]$, $D_j(\xi,\tau) = \frac{b_j - \upsilon\rho\mathbb{I}\xi + d_j}{\upsilon^2}\left(\frac{1-e^{d_j\tau}}{1-g_j e^{d_j\tau}}\right)$, $u_1 = \frac{1}{2}$, $u_2 = -\frac{1}{2}$, $a = \kappa\theta$, $b_1 = \kappa + \lambda - \upsilon\rho$, $b_2 = \kappa + \lambda$, $g_j = \frac{b_j - \upsilon\rho\mathbb{I}\xi + d_j}{b_j - \upsilon\rho\mathbb{I}\xi - d_j}$, $d_j = \sqrt{(\upsilon\rho\mathbb{I}\xi - b_j)^2 - \upsilon^2(2u_j\mathbb{I}\xi - \xi^2)}$.

The figure below presents the payoff functions of a European call in the Heston model ($C^H(S_t, \sigma_t^2, t)$) assuming that: $S_t \in [70, 130]$, $K = 100$, $\sigma_t = 0.2$, $r = 5\%, \upsilon = 0.3, \kappa = 1.5, \lambda = 3, \theta = 0.04, \rho = 0.8$ for different periods remaining to expiration, i.e. $\frac{T-t}{T} \in \{0.01; 0.2; 0.5; 0.7; 0.9\}$.

Figure 1. Payoff functions of a European call in the Heston model assuming that:
$S_t \in [70, 130]$, $K = 100$, $\sigma_t = 0.2$, $r = 5\%$, $\frac{T-t}{T} \in \{0.01; 0.2; 0.5; 0.7; 0.9\}$,
$\upsilon = 0.3$, $\kappa = 1.5$, $\lambda = 3$, $\theta = 0.04$, $\rho = 0.8$



Source: developed by the author

One of the Heston model features is its computational inefficiency. This is the result of the fact that in its original form two characteristic functions are used in the formula for the price of a European call. It makes the pricing process computationally more costly comparing to other approaches where only one characteristic function is implemented. This issue will be analyzed in more detail in the next section of this article.

The double Heston model

In the double Heston model [Christoffersen et al. 2009] three equations are used to describe the price process of the underlying asset, i.e.:

$$dS_t = \mu S_t dt + \sqrt{\sigma_{1,t}^2} S_t dW_{1,t} + \sqrt{\sigma_{2,t}^2} S_t dW_{2,t}. \tag{6}$$

$$d\sigma_{1,t}^2 = \kappa_1(\theta_1 - \sigma_{1,t}^2)dt + \upsilon_1 \sqrt{\sigma_{1,t}^2} dW_{3,t}. \tag{7}$$

$$d\sigma_{2,t}^2 = \kappa_2(\theta_2 - \sigma_{2,t}^2)dt + \upsilon_2 \sqrt{\sigma_{2,t}^2} dW_{4,t}. \tag{8}$$

where: S_t denotes the spot price of the underlying asset at time t, $\sigma_{1,t}^2$, $\sigma_{2,t}^2$ are two variance factors of the price process S_t, μ, θ_1, θ_2, κ_1, κ_2, υ_1, υ_2 are the constants

associated with the drift, the long-term variance factors, the mean-reversion rates, and the volatilities of the variance factors processes, respectively. In the double Heston model the Brownian motions W_1 and W_2 are correlated with Brownian motions W_3 and W_4 with a constants ρ_1 and ρ_2.

The price of a European call in the double Heston model can be determined using the same formula as in the case of the Heston model except of the fact that the probabilities of expiring a European call in-the-money are calculated as follows:

$$P_1^{dH}\left(s_t, \sigma_{1,t}^2, \sigma_{2,t}^2, \tau\right) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Re\left(\frac{e^{-\mathbb{I}\xi lnK}\phi^{dH}\left(\xi-\mathbb{I}, s_t, \sigma_{1,t}^2, \sigma_{2,t}^2\right)}{\mathbb{I}\xi s_t e^{r\tau}}\right) d\xi. \tag{9}$$

$$P_2^{dH}\left(s_t, \sigma_{1,t}^2, \sigma_{2,t}^2, \tau\right) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \Re\left(\frac{e^{-\mathbb{I}\xi lnK}\phi^{dH}\left(\xi, s_t, \sigma_{1,t}^2, \sigma_{2,t}^2\right)}{\mathbb{I}\xi}\right) d\xi. \tag{10}$$

where: $\phi^{dH}\left(\xi, s_t, \sigma_{1,t}^2, \sigma_{2,t}^2\right)$ is the characteristic function of $s_t = lnS_t$. The remaining notation is the same as previously introduced.

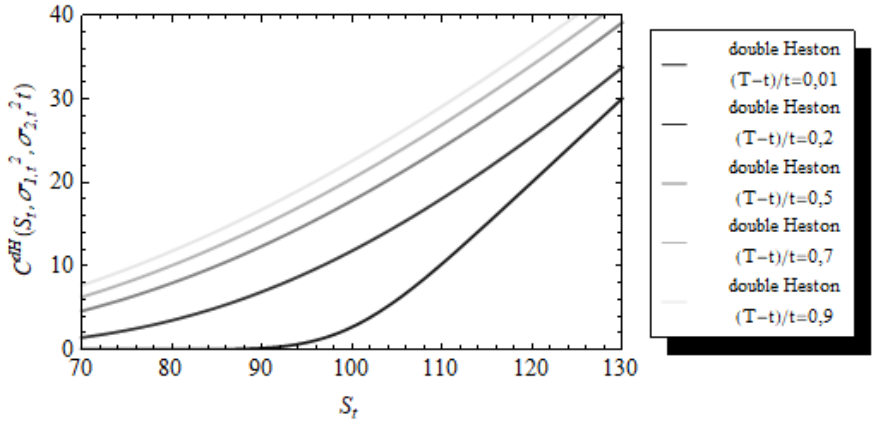
In the double Heston model the general form of the characteristic function of s_t differs from the one appearing in the Heston model and takes the following form:

$$\phi^{dH}\left(\xi, s_t, \sigma_{1,t}^2, \sigma_{2,t}^2\right) = e^{A(\xi,\tau)+B_1(\xi,\tau)\sigma_{1,t}^2+B_2(\xi,\tau)\sigma_{2,t}^2+\mathbb{I}\xi s_t}. \tag{11}$$

where: $A(\xi,\tau) = r\mathbb{I}\xi\tau + \sum_{j=1}^{2}\frac{\kappa_j\theta_j}{v_j^2}\left[\left(\kappa_j - \rho_j v_j \mathbb{I}\xi + d_j\right)\tau - 2ln\left(\frac{1-g_j e^{d_j\tau}}{1-g_j}\right)\right]$, $B_j(\xi,\tau) = \frac{\kappa_j - v_j\rho_j\mathbb{I}\xi + d_j}{v_j^2}\left(\frac{1-e^{d_j\tau}}{1-g_j e^{d_j\tau}}\right)$, $g_j = \frac{\kappa_j - v_j\rho_j\mathbb{I}\xi + d_j}{b_j - v_j\rho_j\mathbb{I}\xi - d_j}$, and $d_j = \sqrt{\left(\kappa_j - v_j\rho_j\mathbb{I}\xi\right)^2 + v_j^2\xi(\xi+\mathbb{I})}$. The remaining notation is the same as previously introduced.

The figure below presents the payoff functions of a European call in the double Heston model ($C^{dH}\left(S_t, \sigma_{1,t}^2, \sigma_{2,t}^2, t\right)$) assuming that: $S_t \in [70, 130]$, $K = 100$, $\sigma_{1,t} = 0.2$, $\sigma_{2,t} = 0.25$, $r = 5\%$, $v_1 = 0.3$, $v_2 = 0.35$, $\kappa_1 = 1.5$, $\kappa_2 = 1.1$, $\lambda = 3$, $\theta_1 = 0.04$, $\theta_2 = 0.06$, $\rho_1 = 0.8$, $\rho_2 = 0.2$ for different periods remaining to expiration, i.e. $\frac{T-t}{T} \in \{0.01; 0.2; 0.5; 0.7; 0.9\}$.

Figure 2. Payoff functions of a European call in the double Heston model assuming that: $S_t \in [70, 130]$, $K = 100$, $\sigma_{1,t} = 0.2$, $\sigma_{2,t} = 0.25$, $r = 5\%$, $\frac{T-t}{T} \in \{0.01; 0.2; 0.5; 0.7; 0.9\}$, $\upsilon_1 = 0.3$, $\upsilon_2 = 0.35$, $\kappa_1 = 1.5$, $\kappa_2 = 1.1$, $\lambda = 3$, $\theta_1 = 0.04$, $\theta_2 = 0.06$, $\rho_1 = 0.8$, $\rho_2 = 0.2$



Source: developed by the author

It is worth noting that the double Heston model shares the same drawbacks as the Heston model. It means that in its original form it is inefficient. Luckily there are some other methods of calculating inverse Fourier transforms of characteristic function $\phi^{dH}\left(\xi, s_t, \sigma_{1,t}^2, \sigma_{2,t}^2\right)$ which allow to lessen computational effort related to pricing European options.

CHARACTERISTIC FUNCTIONS

There are many approaches to determining characteristic function of s_t and calculating its inverse Fourier transform [Carr & Madan 1999, Attari 2004, Bates 2006 and Orzechowski 2018]. As some of the approaches has already been presented [Orzechowski 2020] in the later part of the article only formulas concerning the double Heston model are of interest. As before, for the purpose of the article, it is assumed that $t = 0$. The remaining notation remains consistent with the previously introduced.

The double Heston model

1. The Carr & Madan approach [Carr, Madan 1999] for α = 1:

$$C^{dH}\left(S_0, \sigma_{1,0}^2, \sigma_{2,0}^2, 0\right) = \frac{e^{-\alpha k}}{\pi} \int_0^{\infty} \Re\left(e^{-\mathbb{I}\xi k} \frac{e^{-rT}\phi^{dH}\left(\xi-(\alpha+1)\mathbb{I}, s_0, \sigma_{1,0}^2, \sigma_{2,0}^2\right)}{\alpha^2+\alpha-\xi^2+\mathbb{I}(2\alpha+1)\xi}\right) d\xi. \qquad (12)$$

2. The Attari approach [Attari 2004]:

$$C^{dH}\left(S_0,\sigma_{1,0}^2,\sigma_{2,0}^2,0\right)=S_0\left(1+\frac{e^l}{\pi}\int_0^\infty \Re\left(\frac{e^{-\mathbb{I}\xi l}}{\mathbb{I}(\xi+\mathbb{I})}\psi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)\right)d\xi\right)+$$
$$-e^{-rT}K\left(\frac{1}{2}+\frac{1}{\pi}\int_0^\infty \Re\left(\frac{e^{-\mathbb{I}\xi l}}{\mathbb{I}\xi}\psi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)\right)d\xi\right). \quad (13)$$

where: $l=ln\left(\frac{K}{S_0e^{rT}}\right)$.

3. The Bates approach [Bates 2006]:

$$C^{dH}\left(S_0,\sigma_{1,0}^2,\sigma_{2,0}^2,0\right)=$$
$$=S_0-e^{-rT}K\left(\frac{1}{2}+\frac{1}{\pi}\int_0^\infty \Re\left(\frac{e^{-\mathbb{I}\xi ln\left(\frac{K}{S_0}\right)}}{\mathbb{I}\xi(1-\mathbb{I}\xi)}\right)\varphi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)d\xi\right). \quad (14)$$

4. The Orzechowski approach [Orzechowski 2018]:

$$C^{dH}\left(S_0,\sigma_{1,0}^2,\sigma_{2,0}^2,0\right)=\frac{1}{2}S_o+e^{-rT}\frac{1}{\pi}\int_0^\infty \Re\left(e^{-\mathbb{I}\xi k}\frac{\phi^{dH}\left(\xi-\mathbb{I},s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)}{\mathbb{I}\xi(\mathbb{I}\xi+1)}\right)d\xi. \quad (15)$$

It is worth noting that: $\phi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)$, $\psi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)$ as well as $\varphi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)$ are characteristic functions determined by the following equations:

$$\phi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)=e^{A(\xi,\tau)+B_1(\xi,\tau)\sigma_{1,t}^2+B_2(\xi,\tau)\sigma_{2,t}^2+\mathbb{I}\xi s_t}. \quad (16)$$

$$\psi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)=\phi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)e^{-\mathbb{I}\xi s_0-\mathbb{I}\xi rT}. \quad (17)$$

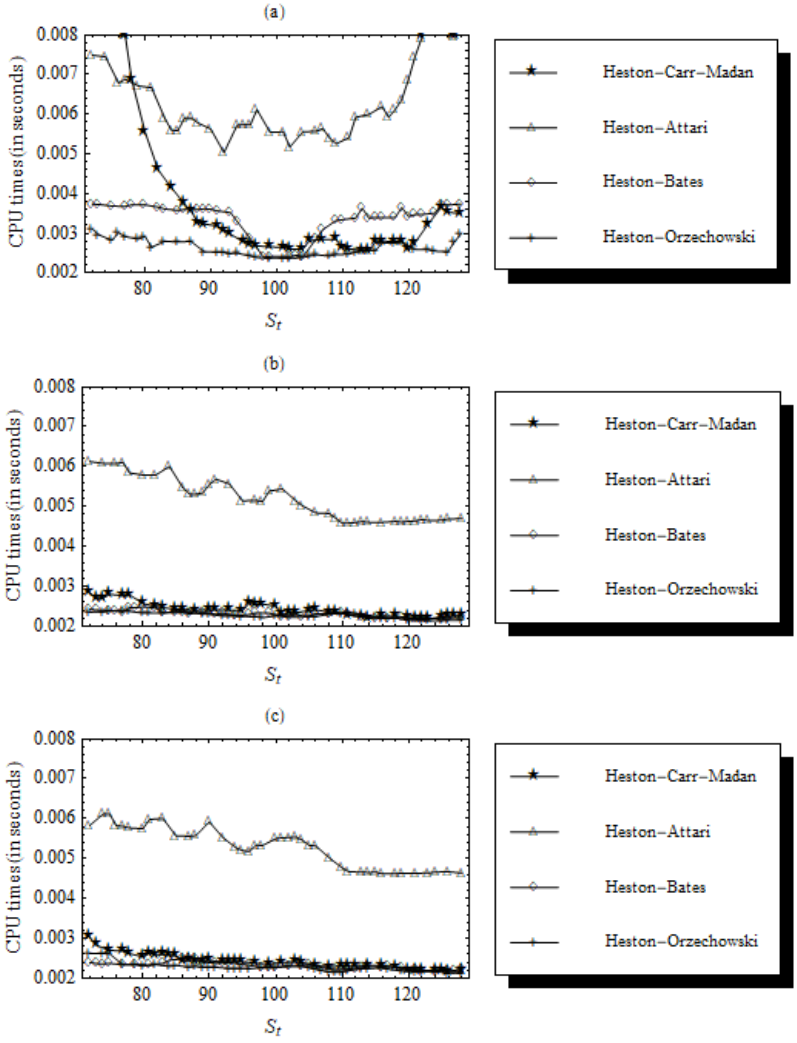
$$\varphi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)=\phi^{dH}\left(\xi,s_0,\sigma_{1,0}^2,\sigma_{2,0}^2\right)e^{-\mathbb{I}\xi s_0}. \quad (18)$$

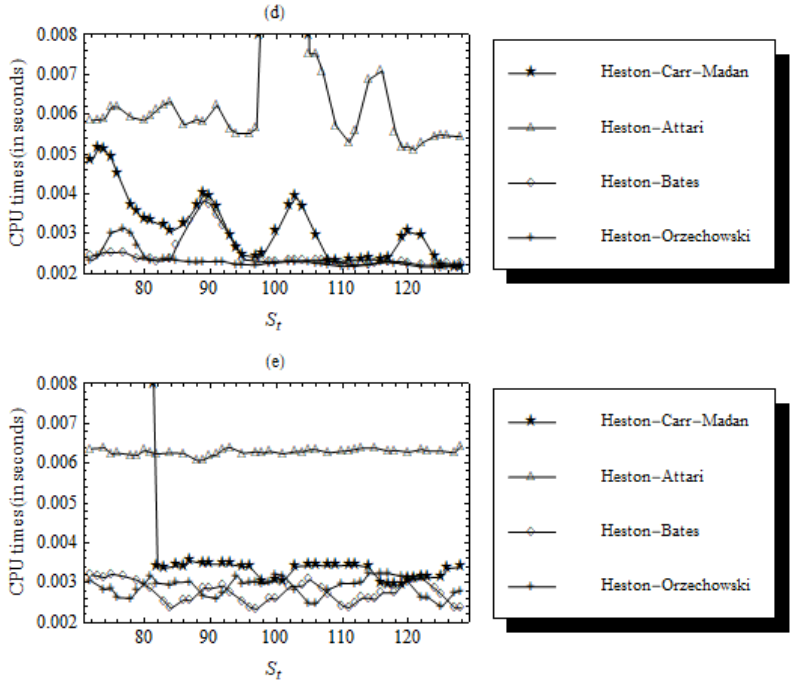
RESULTS

The determination of the most efficient approach both in the Heston and the double Heston models, is based on the results generated with the use of codes developed in Mathematica 10.2. The methodology proposed in the research is compatible with the approach applied previously [Orzechowski 2020]. It means that the theoretical prices of European calls have been evaluated numerically by means of the Gauss-Kronrod quadrature. Additionally the graphs have been smoothed by averaging runs of five elements. It is also worth noting that hardware with the same characteristics has been used for the computation purposes. Cache memory has been cleared before starting codes allowing for the valuation of options.

The results of the research carried out are shown in the graphs below - see Figures 3 and 4.

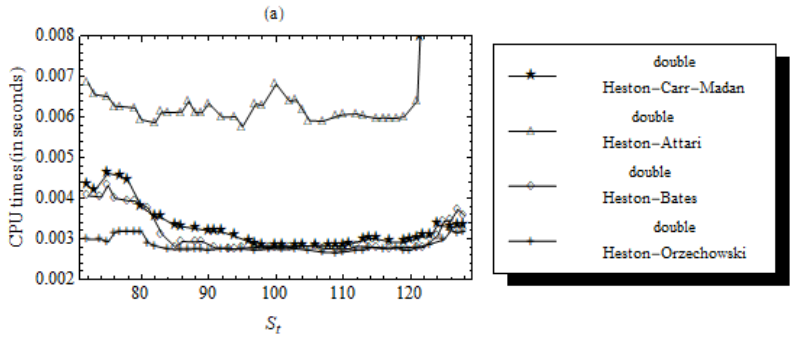
Figure 3. Computational speed in the Heston model assuming that: $S_t \in [70, 130]$, $K = 100$, $\sigma_t = 0.2$, $r = 5\%$, $\upsilon = 0.3$, $\kappa = 1.5$, $\lambda = 3$, $\theta = 0.04$, $\rho = 0.8$ for (a) $\frac{T-t}{T} = 0.01$, (b) $\frac{T-t}{T} = 0.2$, (c) $\frac{T-t}{T} = 0.5$, (d) $\frac{T-t}{T} = 0.7$ and (e) $\frac{T-t}{T} = 0.9$

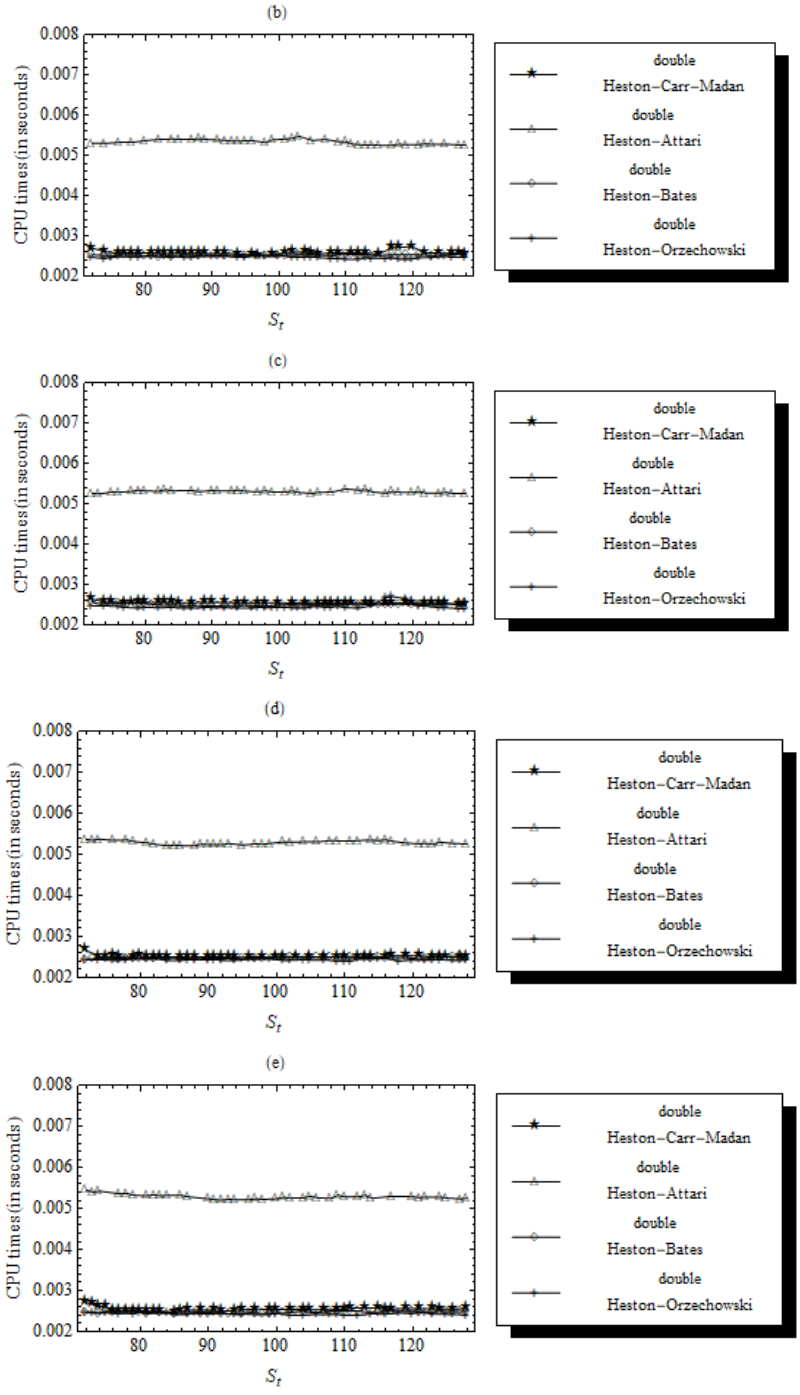




Source: developed by the author

Figure 4. Computational speed in the double Heston model assuming that: $S_t \in [70, 130]$, $K = 100$, $\sigma_{1,t} = 0.2$, $\sigma_{2,t} = 0.25$, $r = 5\%$, $\upsilon_1 = 0.3$, $\upsilon_2 = 0.35$, $\kappa_1 = 1.5$, $\kappa_2 = 1.1$, $\lambda = 3$, $\theta_1 = 0.04$, $\theta_2 = 0.06$, $\rho_1 = 0.8$, $\rho_2 = 0.2$ for (a) $\frac{T-t}{T} = 0.01$, (b) $\frac{T-t}{T} = 0.2$, (c) $\frac{T-t}{T} = 0.5$, (d) $\frac{T-t}{T} = 0.7$ and (e) $\frac{T-t}{T} = 0.9$





Source: developed by the author

The results obtained allow to state that the conclusions drawn previously [Orzechowski 2020] can be extended onto the double Heston model. It means that the computational efficiency of pricing European options depends on the way the characteristic functions and their inverse Fourier transforms are calculated. Such statement is correct not only for the Heston model, but also for the double Heston model used to pricing European options that are close to expiration. On the basis of Figures 3 and 4 it can be also easily concluded that the closer the time to expiration of the European options, the more computationally efficient is the method based on eq. 15. This is, however, not true for the European options close to the moment of their writing. In this case the results are more ambiguous.

SUMMARY

Two models of pricing European options, i.e. the Heston model and the double Heston model were analyzed in this article and then compared in terms of computational speed. Special attention in this regard was paid to the way the characteristic functions and their inverse Fourier transforms are calculated.

On the basis of the results obtained it can be concluded that the closer the time to expiration the greater is the advantage of the method based on eq. 15, regardless the model being considered. At the same time the closer the time to writing European options the more blurred become the differences in efficiency between the models. It is also important to note that these properties hold under the assumptions that the prices of the European calls are evaluated numerically by means of the Gauss-Kronrod quadrature.

REFERENCES

Attari M. (2004) Option Pricing Using Fourier Transform: A Numerically Efficient Simplification. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=520042, [access: 20 12.2021].

Bates D. S. (2006) Maximum Likelihood Estimation of Latent Affine Processes. Review of Financial Studies, 19(3), 909-965.

Black F., Scholes M. (1973) The Pricing of Options and Corporate Liabilities. Journal of Political Economy, 81(3), 637-654.

Carr P., Geman H., Madan D. B., Yor M. (2002) The Fine Structure of Asset Returns: An Empirical Investigation. Journal of Business, 75(2), 305-332.

Carr P., Madan D. B. (1999) Option Valuation Using the Fast Fourier Transform. Journal of Computational Finance, 2(4), 61-73.

Christoffersen P., Heston S. L., Jacob K. (2009) The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work So Well. Management Science, 55(12), 1914-1932.

Eberlein E., Keller U., Prause K. (1998) New Insights into Smile, Mispricing and Value at Risk: The Hyperbolic Model. Journal of Business, 71(3), 371-405.

Heston S. (1993) A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. The Review of Financial Studies, 6(2), 327-343.

Kou S. G. (2002) Jump - Diffusion Model for Option Pricing. Management Science, 48(8), 1086-1101.

Madan D. B., Carr P., Chang E. (1998) The Variance Gamma Process and Option Pricing Model. European Finance Review, 2(1), 79-105.

Madan D. B., Milne F. (1991) Option Pricing with VG Martingale Components, Mathematical Finance, 1(4), 39-55.

Merton R. C. (1976) Option Pricing When Underlying Stock Returns Are Discontinuous. Journal of Financial Economics, 3(1-2), 125-144.

Orzechowski A. (2018) Pricing Correlation Options: from the P. Carr and D. Madan Approach to the New Method Based on the Fourier Transform. Economics and Business Review, 4(1), 16-28.

Orzechowski A. (2020) Pricing European Options in Selected Stochastic Volatility Models. Quantitative Methods in Economics, 21(3), 145-156.