DEBIUTY STUDENCKIE

2023

ZASTOSOWANIE METOD ILOŚCIOWYCH W EKONOMII I FINANSACH

pod redakcją
Alicji Grześkowiak
i Piotra Peterneka

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2023

Recenzja
Katarzyna Ostasiewicz

Redakcja wydawnicza
Elżbieta Żurawska-Łuczyńska

Korekta
Katarzyna Gwizda

Skład i łamanie
Beata Mazur

Projekt okładki
Beata Dębska

Na okładce wykorzystano zdjęcia z zasobów 123 Royalty Free

ISBN 978-83-67899-08-6 (wersja papierowa)
ISBN 978-83-67899-09-3 (wersja elektroniczna)

DOI: 10.15611/2023.09.3

Druk i oprawa: TOTEM

Maciej Marszałek

e-mail: 174964@student.ue.wroc.pl

Uniwersytet Ekonomiczny we Wrocławiu

Nowoczesne metody imputacji braków danych – porównanie wybranych metod

DOI: 10.15611/2023.09.3.04
JEL Classification: C55

Streszczenie: Artykuł opisuje wyniki przeprowadzonych imputacji metodami imputacji wielokrotnej procedurą MICE oraz przy użyciu sieci neuronowych na niewielkim zbiorze zawierającym 690 obserwacji i 6 zmiennych. Uzyskane wyniki porównano i wybrano imputację MICE jako lepszą metodę uzupełniania braków danych w badanym zbiorze.

Słowa kluczowe: braki danych, nowoczesne metody imputacji, uczenie maszynowe

1. Wstęp

W artykule podjęto tematykę związaną z brakami danych w zbiorach danych. Zjawisko to jest bardzo często spotykane w przypadku pracy na rzeczywistych zbiorach danych. Przyczyny powstawania braków mogą być różne, lecz najczęściej wiążą się z błędem ludzkim. Problem ten jest bagatelizowany, a brakujące wartości są często zastępowane wartościami średnimi albo usuwane z analizy. Oba te podejścia mogą prowadzić do znacznego zniekształcenia analizowanej bazy danych.

Celem badania jest przeprowadzenie 2 imputacji, czyli zastąpienia braków danych innymi wartościami w niewielkim zbiorze danych, zawierającym 690 obserwacji oraz 6 zmiennych numerycznych. Zbiór ten posiada losowo wygenerowane braki danych w zmiennych. Zbiór zawiera również zmienną wynikową – target, ponieważ oryginalnie zbiór służy do uczenia klasyfikatorów algorytmami uczenia maszynowego. W zmiennej wynikowej nie wystąpiły jednak braki, dlatego nie wzięła ona udziału w badaniu. Imputacje zostały przeprowadzone metodami imputacji wielokrotnej i imputacji za pomocą sieci neuronowych. Po ich wykonaniu podjęto próbę porównania wyników osiągniętych przy użyciu obu metod oraz wybór lepszej metody dla badanego zbioru danych.

W trakcie badania podjęto próbę odpowiedzi na następujące pytania:

- Jak wygląda badany zbiór danych, ile ma braków i gdzie występują?
- Jakie są możliwe metody postępowania z brakami danych?
- Jak poradziły sobie poszczególne metody imputacji?
- Jak porównać poszczególne metody imputacji?
- Która metoda zwróciła najlepsze wyniki imputacji?
- Czy istnieje jedna najlepsza metoda imputacji?

2. Braki danych – problematyka i metody postępowania

Braki w zbiorach danych są zjawiskiem powszechnym. S. Van Buuren (2012) stwierdza, że braki danych otaczają nas na co dzień, a problem przez nie generowany zbyt długo był ignorowany. Braki danych występują zarówno w ankietach społecznych, czego przyczyną może być zwyczajnie odmowa odpowiedzi respondenta, jak i w innych bazach danych, gdzie przyczyną może być błąd w zapisie, przypadkowe pominięcie pytania czy też nieprecyzyjny pomiar. W książce *Advising on Research Methods*, D. J. Hand, wraz z H. J. Adèrem stwierdzają, że braki w informacji mogą poważnie zagrozić jakości zbioru danych, zniekształcić wyniki analiz bądź uniemożliwić ich przeprowadzenie. Jest to powód, dla którego badacze próbują wypełniać puste miejsca zanim przejdą do etapu analiz statystycznych. Według autorów jednak braki danych w zbiorach są często nieuniknione (Hand, Adèr i Mellenbergh, 2007). Wprawdzie problem braków danych poruszany jest w literaturze przy niemal każdym opracowaniu dotyczącym analizy danych i powstało także wiele opracowań poświęconych postępowaniu z brakami danych, jednak nie istnieją sposoby idealne, które wyeliminowałyby ten problem. Sytuację z brakami danych najlepiej opisują G. M. Cox i G. M. Cochran (1957) stwierdzając, że najlepszym rozwiązaniem problemu braków danych jest ich w ogóle nie mieć.

Najłatwiejszym sposobem poradzenia sobie z brakami danych jest odrzucenie obiektów z wartościami brakującymi. Jest to metoda najszybsza, ale też niosąca ze sobą największe szkody dla zbioru danych. Wyrzucenie dużej części obiektów spowoduje znaczne uszczuplenie informacji, jaką te dane niosą. W efekcie może się okazać, że stworzony model nie będzie w pełni odzwierciedlał rzeczywistych zdarzeń. W niektórych przypadkach usunięcie obserwacji z brakami danych będzie wręcz pozbawione sensu. W przypadku wystąpienia braków danych w szeregu czasowym usunięcie ich z analizy w celu stworzenia prostej trendu nie tylko znacząco zniekształci otrzymany wynik, ale również będzie sprzeczne z założeniem ciągłości obserwacji w kolejnych jednostkach czasu.

Drugi sposób polega na imputacji brakujących wartości, czyli na wyeliminowaniu braków przez zastąpienie ich wybraną wartością. Istnieje wiele podziałów imputacji ze względu na sposób postępowania. Jednego z podstawowych podziałów dokonali D. Kasprzyk i G. Kalton (1982), dzieląc je na imputacje dedukcyjne oraz statystyczne. Imputacje dedukcyjne oparte są na zależności między zmiennymi. Mają one charakter deterministyczny oraz pozwalają na jednoznaczne wyznaczenie wartości imputacyjnej. Natomiast imputacje statystyczne opierają się na pozostałej części zbioru danych imputowanej zmiennej (Grochowina, 2014). Imputacje statystyczne możemy podzielić na metody klasyczne oraz nowoczesne. Do metod klasycznych zaliczamy (Pokropek, 2018):

1. Imputację za pomocą średniej, która polega na zastąpieniu brakujących danych wartością średnią wyznaczoną na pozostałej grupie obserwacji.
2. Imputację za pomocą mediany. Podobnie jak w przypadku imputacji z wykorzystaniem średniej metoda sprowadza się do zastąpienia brakującej wartości me-

dianą wyznaczoną na kompletnej grupie obserwacji. Możliwe jest zastosowanie mediany wyznaczonej osobno w klasach (grupach) wyłonionych na podstawie przyjętego kryterium. Wówczas mamy do czynienia z imputacją medianą warunkową.

3. Imputację regresyjną opartą na równaniu regresji. Wówczas za zmienną objaśnianą w regresji przyjmowana jest zmienna, której dotyczy brak danych, a za zmienne objaśniające przyjmowane są pozostałe zmienne (kompletne). Imputacja sprowadza się do odpowiedniego doboru modelu regresji, oszacowania modelu i ostatecznie zastąpienia brakującej wartości, wartością teoretyczną wynikającą z modelu.
4. Imputację nieparametryczną, która polega na znajdowaniu tzw. dawców wśród par zmiennych, z których jedna obarczona jest brakiem danych. Procedura konceptualnie jest bardzo prosta. Zakładamy, że mamy zestaw par zmiennych, ale w jednej z tych par brakuje danych. Teraz chcemy znaleźć inną parę zmiennych, gdzie obie strony (y_1 i y_2) mają dane, które są do siebie zbliżone lub nawet identyczne. Taką parę nazywamy dawcą. Gdy znajdziemy takiego dawcę, to biorąc pod uwagę, że y_1(a) (czyli y_1 dla pary z brakującymi danymi) jest podobne lub równe y_1(b) (czyli y_1 dla pary z danymi), to możemy użyć y_2(a) (czyli y_2 dla pary z danymi) jako zastępstwo dla brakującej wartości y_2(b). Jeśli mamy więcej zmiennych, to chcemy znaleźć dawcę, który jest jak najbardziej podobny pod względem wszystkich tych zmiennych, a nie tylko jednej.

Niestety każda z wymienionych klasycznych metod imputacji niesie ze sobą ryzyko znaczącego zmniejszenia odchylenia standardowego i zaniżania wariancji, przez co zdecydowana większość parametrów estymowanych modeli statystycznych będzie błędnie szacowana.

Oddzielną grupą metod są metody nowoczesne, do których zaliczamy między innymi wielokrotne imputacje. Imputacje wielokrotne są metodą polegającą na wypełnieniu braków danych przez wielokrotne losowe przypisanie wartości z dostępnej puli danych. Jest to jedna z najczęściej stosowanych metod radzenia sobie z brakami danych, ponieważ jest prosta w implementacji i pozwala na zachowanie zmienności w danych (Little, 2002).

Imputacja wielokrotna składa się z kilku kroków. Pierwszy krok polega na określeniu braków danych oraz ich rozmieszczenia w zbiorze danych. Następnie dla każdej brakującej wartości wybierana jest losowo jedna z dostępnych wartości z tej samej zmiennej. Proces ten jest powtarzany wielokrotnie, najczęściej kilkadziesiąt razy. W ostatnim kroku, dla każdej zmiennej z brakami danymi, tworzona jest nowa zmienna, która jest średnią z wielokrotnie przypisanych wartości (Rubin, 1976).

Imputacja wielokrotna ma kilka zalet. Po pierwsze, pozwala na zachowanie zmienności w danych, co jest szczególnie ważne przy analizie statystycznej. Po drugie, jest prosta w implementacji i nie wymaga specjalistycznej wiedzy statystycznej.

Jednak imputacja wielokrotna ma również pewne wady. Po pierwsze, może prowadzić do błędów w wynikach analizy statystycznej, jeśli braki danych są nieodpowiednio rozmieszczone (Little, 2002). Po drugie, nie uwzględnia relacji między

zmiennymi, co może prowadzić do błędnych wniosków (Allison, 2001). Imputacje wielokrotne pozwalają na zachowanie większej liczby danych w badaniu, co ma pozytywny wpływ na jakość wyników. Jednak należy pamiętać, że ta metoda może prowadzić do błędów w wynikach, jeśli braki danych są związane z jakimś zjawiskiem (np. skorelowanym z innymi zmiennymi).

Pomimo tych ograniczeń imputacja wielokrotna jest nadal często stosowana jako metoda radzenia sobie z brakami danych w badaniach statystycznych. Należy pamiętać, że imputacje wielokrotne nie są jednak panaceum na problemy braków danych i nie powinny być stosowane bez wcześniejszej analizy danych oraz rozważenia innych metod imputacji.

Do nowoczesnych metod imputacji braków danych można również zaliczyć imputację za pomocą sieci neuronowych. Metoda ta rzadko jest wykorzystywana do imputacji ze względu na trudności przy poszukiwaniu optymalnych parametrów. Sieci neuronowe to rodzaj modeli uczenia maszynowego, które naśladują działanie ludzkiego mózgu składającego się z miliardów neuronów. Sieci te składają się z połączonych ze sobą neuronów sztucznych, które przetwarzają wejściowe dane, wykonują na nich obliczenia i zwracają wynik na wyjście.

Sztuczne neurony w sieci neuronowej mają wagę, która jest modyfikowana podczas procesu uczenia, a także funkcję aktywacji, która określa, czy neuron zostanie pobudzony i prześle sygnał dalej. Proces uczenia polega na dostosowaniu wag połączeń między neuronami w celu minimalizacji błędu predykcji.

W tym artykule do uzupełnienia braków danych wykorzystano dwie różne metody. Pierwsza z nich to metoda imputacji wielokrotnych według procedury MICE (ang. *Multiple Imputation by Chained Equations*). Jest to metoda imputacji brakujących danych, która pozwala na uzyskanie kilku zestawów zastępczych wartości brakujących. Procedura MICE jest realizowana przez tworzenie wielu zestawów (ang. *imputations*) danych, w których wartości brakujące są zastępowane przez wartości przewidywane na podstawie innych zmiennych. W każdym zestawie brakujące wartości są zastępowane innymi, losowo wygenerowanymi wartościami. Następnie na każdym z tych zestawów przeprowadzana zostaje analiza statystyczna, a otrzymane wartości sumuje się w celu uzyskania końcowych wyników. Procedura MICE pozwala na wykorzystanie różnych metod imputacji. Do badania wybrano i porównano 6 z nich:

- PMM (ang. *Predictive mean Matching*) – metoda ta polega na znalezieniu rekordu z wartością brakującą, a następnie dopasowaniu (*matching*) go do rekordu nieposiadającego braku, który charakteryzuje się zbliżonymi wartościami innych zmiennych objaśniających. Wartość brakująca jest następnie zastępowana wartością zmiennych objaśniających z rekordu, który został użyty do dopasowania.
- Midastouch – metoda ta opiera się na algorytmie klastrowania, który pozwala na zgrupowanie rekordów podobnych do rekordu z brakiem danych. Wartość brakująca jest następnie zastępowana medianą zmiennych objaśniających w grupie klastrowej.

- Sample – metoda ta polega na wygenerowaniu losowych wartości z rozkładu próbkującego dla zmiennej objaśniającej.
- CART (ang. *Classification and Regression Trees*) – metoda ta polega na zbudowaniu drzewa decyzyjnego, które pozwala na przewidywanie wartości brakującej na podstawie wartości innych zmiennych objaśniających.
- RF (ang. *Random Forest*) – metoda ta polega na zbudowaniu wielu drzew decyzyjnych, które pozwalają na przewidywanie wartości brakującej na podstawie wartości innych zmiennych objaśniających.
- Norm.nob. – metoda ta pozwala na imputację wartości brakujących za pomocą losowych wartości wygenerowanych z normalnego rozkładu oraz losowych wartości z rozkładu jednostajnego między wartością minimalną a maksymalną zmiennej.

Drugą metodą imputacji wykorzystaną w badaniu jest imputacja braków danych za pomocą wytrenowanych sieci neuronowych. Sieci neuronowe to rodzaj modelu uczenia maszynowego, który naśladuje funkcjonowanie ludzkiego mózgu i jest stosowany do rozwiązywania różnych problemów, takich jak klasyfikacja, regresja, rozpoznawanie obrazów i języka naturalnego. Sieci neuronowe są coraz bardziej popularnym narzędziem w dziedzinie uczenia maszynowego ze względu na ich zdolność do wykrywania złożonych wzorców i uczenia się na podstawie dużej ilości danych. Jednak ich skuteczność zależy od wielu czynników, takich jak: jakość danych, architektura sieci i parametry treningowe. Dlatego ważne jest, aby dokładnie przetestować i dostosować sieć do konkretnego problemu. Dużym problemem jest dobór odpowiedniej architektury sieci, gdyż nie ma jednoznacznych przesłanek, jak dana architektura dla danych zmiennych powinna wyglądać. Istnieją jedynie podpowiedzi, że liczba neuronów w poszczególnych warstwach nie powinna przekraczać liczby zmiennych, a głębokość sieci nie powinna przekraczać 3.

3. Metodyka badań własnych

Dane wykorzystane do analizy zostały pobrane ze strony https://www.openml.org/. Zbiór danych „Credit Approval" został udostępniony przez Rossa Quinlana na platformie UCI w 1987 roku. Zawiera informacje dotyczące wniosków o kredyt, w których atrybuty, takie jak: wiek, dochód, historia kredytowa itp. zostały zanonimizowane, aby chronić poufność danych. Zbiór ten zawiera 690 obserwacji i 16 zmiennych, ciągłych i nominalnych, występują również braki danych. Głównym celem analizy tego zbioru danych jest przewidzenie, czy wniosek o kredyt zostanie zatwierdzony, czy też odrzucony na podstawie dostępnych informacji o wnioskodawcach.

Zbiór zawiera 15 zmiennych nazwanych od A1 do A15 oraz zmienną wynikową, gdzie „+" znaczy przyznanie karty kredytowej, a „-" nieprzyznanie karty kredytowej. W analizie zostały uwzględnione tylko zmienne numeryczne, a więc A2, A3, A8, A11, A14, A15. Zmienna wynikowa została zmieniona z „+" i „-" na 1 i 0, gdzie

1 oznaczało przyznanie karty, a 0 oznaczało nieprzyznanie karty. Ostatecznie zbiór zawierał 6 zmiennych decyzyjnych, jedną zmienną wynikową oraz 690 obserwacji. W zmiennych A14 oraz A2 występowały braki danych. Aby lepiej zilustrować badane zagadnienie, zostały wygenerowane w sposób losowy braki danych również w pozostałych 4 zmiennych. Usunięte w ten sposób wartości rzeczywiste zostały zapomniane i nie uwzględniono ich w dalszych analizach. Z powodu ochrony danych osobowych wszystkie zmienne od A1 do A15 nie są opisane i nie wiadomo, jaką cechę opisują.

Do uzupełnienia braków danych wykorzystano procedurę MICE oraz sztuczne sieci neuronowe. W przypadku procedury MICE jakość przeprowadzonej imputacji została oceniona miarą MAE (ang. *Mean Absolute Error*), która mierzy średnią wartość bezwzględną błędu między rzeczywistą wartością zmiennej objaśnianej a jej wartością przewidywaną przez model. Miara ta była wyliczana osobno dla każdej ze zmiennej jedynie na podstawie tych obserwacji, które posiadały rzeczywiste wartości. Im lepiej model dopasował się do istniejących już obserwacji, tym lepiej estymuje wartości brakujących obserwacji. Ponieważ MAE zwraca wartości w postaci bezwzględnej odpowiadającej wartościom danej zmiennej, a wartości zmiennych różnią się istotnie między sobą, miara MAE została dodatkowo podzielona przez średnią wartość dla danej zmiennej. W ten sposób spróbowano uzyskać wyniki przeprowadzonej imputacji w przedziale 0-1, gdzie 0 oznaczało idealną imputację, a 1 słaby poziom imputacji. Warto zaznaczyć, że miernik ten może przyjąć wartości wyższe od 1, ale oznaczałoby to, że średni błąd był wyższy niż średnia wartość dla danej zmiennej, co świadczyłoby o bardzo słabej dokładności przeprowadzonej imputacji. Do przeprowadzenia imputacji wielokrotnej wykorzystano pakiet MICE w Rstudio.

Procedura uzupełniania braków za pomocą sieci neuronowych wyglądała następująco:

- Na starcie usunięto wierszami wszystkie braki danych.
- Dokonano standaryzacji zmiennych.
- Podzielono zbiór bez braków na zbiór treningowy i testowy.
- Stworzono pierwszą sieć neuronową, której zadaniem było przewidywanie wartości pierwszej zmiennej A2 na podstawie pozostałych.
- Dokonano oceny zbudowanej sieci na podstawie zbiorów treningowego i testowego i przeprowadzono poszukiwanie optymalnej architektury sieci.
- Dokonano ostatecznej oceny sieci oraz predykcji brakujących wartości dla zmiennej A2, wyciągając ze zbioru oryginalnego te wiersze, w których brak danych występował jedynie dla zmiennej A2.
- Do zbioru danych bez braków dodano obserwacje z uzupełnionymi wartościami w zmiennej A2, rozszerzając zbiór danych bez braków.
- Powtarzano kroki dla każdej zmiennej z brakami danych aż do osiągnięcia 680 kompletnych obserwacji. Pozostałe 10 obserwacji były to obserwacje, w których występował pewien układ braków danych, w którym brak danych występował

dla danej obserwacji nie tylko w jednej zmiennej, ale przynajmniej w dwóch. Klasyczna sieć neuronowa nie jest w stanie nauczyć się przewidywania wartości 2 różnych zmiennych, dlatego pozostałe 10 obserwacji zostało usuniętych z dalszych analiz.

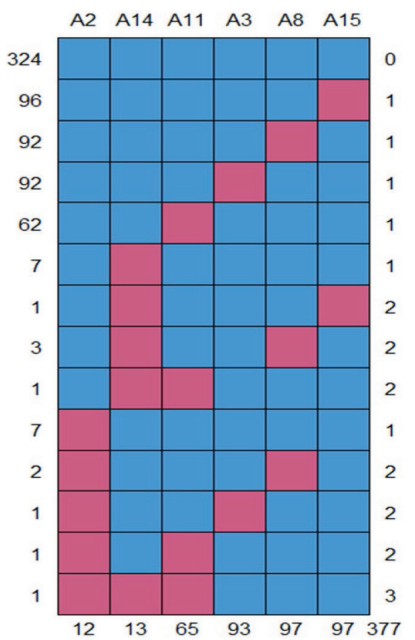
Do trenowania i testowania sieci skorzystano z biblioteki neuralnet dostępnej w Rstudio.

Ostatni etapem badania to porównanie obu metod imputacji na podstawie błędów MAE dzielonych przez wartość średnią ze zbioru oraz wybór tej metody, która dla badanego zbioru zwróciła lepsze wyniki.

4. Wyniki badań własnych

Poprawne przeprowadzenie imputacji braków danych wymaga wnikliwego zbadania badanego zbioru. Aby zdać sobie sprawę, z jakim zbiorem mamy do czynienia, zbadano najpierw strukturę braków danych. Struktura ta jest przedstawiona na rysunku 1.

Czerwone kwadraty pokazują, gdzie występują poszczególne braki danych. Widać, że wierszy, w których nie występuje żaden brak danych, jest tylko 324. Braki w samej zmiennej A15 i A8 występują 97 razy, w zmiennej A3 – 93 razy, w zmiennej A11 – 65 razy, w zmiennej A14 – 13 razy i w zmiennej A2 – 7 razy. Macierz ta pokazuje również, ile razy braki danych dla danej zmiennej wystąpiły tylko dla niej, a ile razy wraz z brakiem danych w jednej zmiennej wystąpiły również w innej. Można zauważyć, że braki w zmiennej A15 wystąpiły 96 razy jako niezależne braki danych, natomiast raz wystąpiły wraz z zmienną A14.



Rysunek 1. Układy braków danych w badanym zbiorze

Źródło: opracowanie na podstawie badań własnych.

Usunięcie obserwacji z brakami danych spowodowałoby pomniejszenie zbioru danych o ponad 50%, co powodowałoby sporą utratę informacji. Aby tego uniknąć, dokonano pierwszej próby imputacji braków danych metodą imputacji wielokrotnych. Wyniki imputacji poszczególnymi metodami dla poszczególnych zmiennych przedstawiono w tabeli 1.

Tabela 1. Błędy znormalizowane MAE dla określonych metod imputacji wielokrotnych w poszczególnych zmiennych

	Metoda					
Zmienna	PMM	midastouch	sample	CART	RF	norm.nob
A2	0,199	0,198	0,206	0,166	0,175	0,200
A3	0,509	0,493	0,555	0,492	0,489	0,557
A8	0,574	0,559	0,687	0,548	0,544	0,696
A11	0,782	0,750	0,816	0,731	0,634	1,164
A14	0,453	0,458	0,487	0,437	0,425	0,474
A15	1,020	0,840	0,864	0,628	0,587	3,176

Źródło: opracowanie na podstawie badań własnych.

Dla każdej zmiennej wybrano najlepszą pod względem wybranej miary metodę. Dla zmiennej A2 najlepsza okazała się metoda CART, natomiast dla pozostałych zmiennych metoda RF. Warto zauważyć, że wszystkie wartości znormalizowanego MAE wychodzą niepokojąco wysokie. Świadczy to o tym, że program miał problem z dobrym dopasowaniem się do danych, a imputacje mogą być obarczone dużym błędem.

Drugi sposób imputacji jest oparty na sieciach neuronowych. Pierwsza sieć neuronowa przewiduje zmienną A2 na podstawie pozostałych zmiennych. Przebadano wszystkie logiczne architektury sieci, czyli takie, gdzie liczba neuronów nie przekraczała liczby zmiennych decyzyjnych w sieci, a jej maksymalna głębokość wynosiła 3. Następnie wybrano tę architekturę, dla której błąd prognozy w zbiorze testowym jest najmniejszy. Tabela 2 przedstawia 10 najlepszych architektur sieci neuronowych wraz z błędami prognozy na zbiorach treningowych i testowych.

Z tabeli wynika, że dla pierwszej sieci najlepszą strukturą będzie struktura złożona z 2 warstw ukrytych, gdzie w każdej jest po 1 neuronie. Sieć o takiej strukturze została wykorzystana do uzupełnienia 7 brakujących wartości w zmiennej A2 i powiększenia zbioru wyjściowego do budowania kolejnej sieci neuronowej.

Druga sieć neuronowa powstała w celu przewidywania zmiennej A3 na podstawie pozostałych zmiennych. W poszukiwaniu optymalnej struktury sieci ponownie zbadano wszystkie logiczne architektury sieci. W tabeli 3 przedstawiono 10 struktur sieci, które charakteryzują się najmniejszym błędem na zbiorze testowym.

Tabela 2. Wartości błędów predykcji na zbiorach treningowych i testowych dla różnych struktur sieci neuronowej przewidującej zmienną A2

Pozycja	architectures_A2	MAE_train_A2	MAE_test_A2
1	1-1	0,2526	0,2876
2	1-3	0,2411	0,2928
3	1	0,2551	0,2963
4	1-2	0,2307	0,2983
5	1-5	0,2221	0,3093
6	1-1-5	0,1886	0,3132
7	1-1-4	0,1918	0,3149
8	1-4	0,2195	0,3190
9	1-4-2	0,1494	0,3200
10	3-1	0,1048	0,3210

Źródło: opracowanie na podstawie badań własnych.

Tabela 3. Wartości błędów predykcji na zbiorach treningowych i testowych dla różnych struktur sieci neuronowej przewidującej zmienną A3

Pozycja	architectures_A3	MAE_train_A3	MAE_test_A3
1	1-1-2	0,6486	0,7466
2	1-3	0,7529	0,7737
3	1-4	0,7013	0,7910
4	1-5	0,6969	0,8049
5	1-1	0,7879	0,8115
6	1-3-5	0,5876	0,8176
7	1-2	0,7601	0,8189
8	1	0,7983	0,8193
9	1-2-5	0,5532	0,8245
10	1-5-1	0,3947	0,8278

Źródło: opracowanie na podstawie badań własnych.

W przypadku drugiej sieci neuronowej wyniki predykcji są znacznie gorsze niż w przypadku predykcji dla zmiennej A2. Najniższą wartością błędu na zbiorze testowym charakteryzowała się sieć o strukturze z 3 warstwami ukrytymi, po 1 neuronie na 2 pierwszych warstwach i 2 na ostatniej. Sieć o takiej strukturze została wykorzystana do uzupełnienia 92 brakujących wartości w zmiennej A3 i powiększenia zbioru wyjściowego do budowania kolejnej sieci neuronowej.

Trzecia sieć neuronowa miała za zadanie estymować wartości zmiennej A8 na podstawie pozostałych zmiennych. Analiza optymalnej struktury sieci wyłoniła 10 najlepszych architektur, przedstawionych w tabeli 4.

Tabela 4. Wartości błędów predykcji na zbiorach treningowych i testowych dla różnych struktur sieci neuronowej przewidującej zmienną A8

Pozycja	architectures_A8	MAE_train_A8	MAE_test_A8
1	1-1-1	0,7621	0,7951
2	1-3	0,7552	0,8136
3	1-4	0,7355	0,8154
4	1	0,8733	0,8229
5	1-4-1	0,6126	0,8602
6	1-1-4	0,6905	0,8606
7	1-5	0,6950	0,8822
8	1-1-2	0,6890	0,8861
9	1-1	0,8095	0,8915
10	1-3-1	0,6112	0,9139

Źródło: opracowanie na podstawie badań własnych.

Ponownie z tabeli widać, że mimo przeprowadzenia strojenia parametrów sieci, nawet najlepsza architektura sieci neuronowej pod względem minimalizacji błędu prognozy zwraca wysokie wartości błędu. Najlepszą strukturą sieci okazała się 3-warstwowa po 1 neuronie w każdej. Sieć o takiej strukturze została wykorzystana do uzupełnienia 92 brakujących wartości w zmiennej A8 i powiększenia zbioru wyjściowego do budowania kolejnej sieci neuronowej.

Czwarta sieć neuronowa powstała w celu estymowania wartości zmiennej A11 na podstawie pozostałych zmiennych. Błędy zbiorów treningowych i testowych 10 najlepszych struktur sieci neuronowej przedstawiono w tabeli 5.

Tabela 5. Wartości błędów predykcji na zbiorach treningowych i testowych dla różnych struktur sieci neuronowej przewidującej zmienną A11

Pozycja	architectures_A11	MAE_train_A11	MAE_test_A11
1	1-1	1,1173	1,0368
2	1	1,1517	1,1797
3	1-1-1	0,9667	1,2453
4	1-3	1,0552	1,2648
5	1-2-1	0,8814	1,2889
6	1-1-2	0,9797	1,2951
7	1-4	1,0054	1,2970
8	3-2-2	0,4992	1,3012
9	1-2-3	0,8503	1,3153
10	2-2-4	0,5516	1,3241

Źródło: opracowanie na podstawie badań własnych.

Ponownie widać, że sieć miała duże problemy z nauczeniem się zależności w danym zbiorze. Mimo zastosowania optymalnej architektury sieci, wartości błędu na zbiorze testowym są bardzo wysokie. Widać również, że szybko, wraz ze spadkiem wartości błędu na zbiorze treningowym, błąd na zbiorze testowym rośnie, co oznacza, że zwiększanie liczby neuronów w poszczególnych warstwach będzie nieuchronnie prowadziło do przeuczenia sieci. Najlepszą architekturą okazała się 2-warstwowa, z 1 neuronem w każdej. Sieć o takiej strukturze została wykorzystana do uzupełnienia 62 brakujących wartości w zmiennej A11 i powiększenia zbioru wyjściowego do budowania kolejnej sieci neuronowej.

Przedostatnia już sieć neuronowa miała za zadanie estymować wartości zmiennej A14. Dziesięć najlepszych struktur sieci przedstawiono w tabeli 6.

Tabela 6. Wartości błędów predykcji na zbiorach treningowych i testowych dla różnych struktur sieci neuronowej przewidującej zmienną A14

Pozycja	architectures_A14	MAE_train_A14	MAE_test_A14
1	1-2	0,5898	0,6273
2	1-3	0,5850	0,6333
3	1	0,6146	0,6353
4	1-1	0,6139	0,6382
5	1-1-1	0,5440	0,6535
6	1-1-2	0,5473	0,6615
7	1-5	0,5532	0,6636
8	1-3-5	0,5316	0,6649
9	1-1-3	0,5324	0,6816
10	1-4	0,5696	0,6972

Źródło: opracowanie na podstawie badań własnych.

Najlepsze wyniki sieć osiągnęła dla 2-warstwowej struktury, z 1 neuronem na pierwszej i 2 na drugiej. Ponownie wynik imputacji zostawia wiele do życzenia. Sieć o takiej strukturze została wykorzystana do uzupełnienia 7 brakujących wartości w zmiennej A14 i powiększenia zbioru wyjściowego do budowania kolejnej sieci neuronowej.

Ostatnia sieć neuronowa przewiduje wartości zmiennej A15 na podstawie pozostałych zmiennych. Jest to sieć, która ma największy zbiór danych, na których może się uczyć zależności. Wyniki poszukiwania optymalnej struktury sieci przedstawiono w tabeli 7.

Imputacja dla zmiennej A15, mimo że sieć posiadała największy zbiór ze wszystkich sieci, wypadła najgorzej. Prawdopodobnie jest to spowodowane bardzo małym zbiorem, jak na potrzeby sieci neuronowych, przez co sieć nie była w stanie się nauczyć zależności miedzy zmiennymi. Kolejnym powodem może być fakt, że poprzed-

nie imputacje również niosły ze sobą spory błąd predykcji, co może powodować dodatkowe trudności. Najlepszą architekturą spośród zbadanych okazała się zwykła sieć jednowarstwowa z jednym neuronem. Ponownie widać zjawisko przeuczenia sieci, gdy zwiększamy ilość warstw i neuronów w warstwach. Utworzony model sieci neuronowej wykorzystano do uzupełnienia pozostałych 96 braków w zmiennej A15.

Tabela 7. Wartości błędów predykcji na zbiorach treningowych i testowych dla różnych struktur sieci neuronowej przewidującej zmienną A15

Pozycja	architectures_A15	MAE_train_A15	MAE_test_A15
1	1	1,521848	1,436233
2	1-1	1,5619041	1,456152
3	1-2-2	1,3270052	1,523583
4	1-2	1,710745	1,538516
5	1-3-3	1,1887738	1,605239
6	1-1-2	1,071265	1,611048
7	1-3-4	1,3300673	1,624031
8	1-1-4	1,2584002	1,627393
9	1-3-5	1,2674116	1,630391
10	1-4	1,8493138	1,654853

Źródło: opracowanie na podstawie badań własnych.

Ostatecznie uzyskano zbiór 680 obserwacji. Pozostałe 10 obserwacji posiadało taki układ braków danych, że w jednej obserwacji występowały jednocześnie braki w 2 lub więcej zmiennych. Ponieważ klasyczna sieć neuronowa nie jest w stanie przewidywać wartość 2 lub więcej zmiennych, obserwacje te zostały wyrzucone.

Porównanie dokładności poszczególnych imputacji dla każdej ze zmiennych metodami imputacji wielokrotnej oraz imputacji przy użyciu sieci neuronowej przedstawiono w tabeli 8.

Tabela 8. Porównanie wyników imputacji metodami imputacji wielokrotnej i imputacji z wykorzystaniem sieci neuronowych

	Imputacja	
Zmienna	wielokrotna	sieciami neuronowymi
A2	0,166	0,288
A3	0,489	0,747
A8	0,544	0,795
A11	0,634	1,037
A14	0,425	0,627
A15	0,587	1,436

Źródło: opracowanie na podstawie badań własnych.

W przypadku badanego zbioru przyjęty błąd prognozy dla każdej ze zmiennych z brakami danych wychodził mniejszy w przypadku imputacji metodą imputacji wielokrotnych wedle algorytmu MICE. Nie mniej jednak należy zauważyć, że obie metody imputacji obarczone są dużym błędem prognozy. Algorytm MICE najgorzej poradził sobie ze zmienną A11, natomiast sieć neuronowa zdecydowanie gorzej poradziła sobie z przewidywaniem zmiennej A15. W przypadku tego zbioru trzeba stwierdzić, że lepszą metodą uzupełniania braków danych jest metoda imputacji wielokrotnej MICE.

5. Zakończenie

Artykuł miał na celu pokazanie dwóch różnych metod imputacji braków danych w przykładowym zbiorze danych z brakami. Pokazana została struktura braków danych w zbiorze i na tej podstawie przeprowadzono procedurę imputacji braków danych oraz porównano otrzymane wyniki.

Imputacja za pomocą sieci neuronowych wypadła gorzej dla każdej zmiennej, trzeba natomiast pamiętać, że analizowany zbiór posiadał mało zmiennych i mało obserwacji, przez co sieć neuronowa miała duży problem, żeby poprawnie nauczyć się zależności między zmiennymi. Dodatkowym problemem przy imputacji przy sieciach była duża skłonność sieci do przeuczenia, gdzie na zbiorze treningowym sieć dopasowywała się coraz lepiej do danych, natomiast na zbiorze testowym błąd predykcji był coraz wyższy. Metoda imputacji klasycznymi sieciami neuronowymi ma również tę wadę, że sieć nie potrafi przewidywać układów braków danych, w których w jednej obserwacji występują braki danych w 2 bądź większej liczbie zmiennych. Taka sytuacja wystąpiła w badanym zbiorze 10 razy, przez co finalny zbiór bez braków danych, który zwróciły sieci neuronowe, posiadał 680 obserwacji. Kolejnym problemem w sieciach neuronowych jest poszukiwanie optymalnej architektury sieci. W przypadku zbioru danych z niewielką liczbą zmiennych nie jest to bardzo problematyczne, natomiast wraz ze wzrostem liczby zmiennych drastycznie rośnie liczba kombinacji, jakich można użyć w architekturze sieci. W takim przypadku przebadanie wszystkich kombinacji staje się bardzo uciążliwe i czasochłonne i trzeba zastosować metody heurystyczne przy poszukiwaniu sieci, które prawdopodobnie nie pozwolą uzyskać optymalnych wyników.

Imputacja metodą MICE pozwoliła uzyskać lepsze wyniki pod względem minimalizacji średniego błędu predykcji dla każdej badanej zmiennej. Dodatkowo imputacja za pomocą tej metody zwróciła pełny zbiór danych – 690 obserwacji. Metoda ta była szybsza w zastosowaniu i generowała mniej problemów decyzyjnych. Jedyną kwestią, którą należało rozstrzygnąć był dobór odpowiedniej metody imputacji.

Nie oznacza to jednak, że metoda imputacji za pomocą procedury MICE zawsze zwróci lepsze wyniki. Wszystko zależy od struktury badanego zbioru, liczby braków danych, wielkości zbioru, liczby zmiennych itd. Metoda imputacji z użyciem sieci neuronowych może okazać się lepsza w przypadku większych zbiorów, w których

sieć będzie miała wystarczająco dużo danych treningowych, aby nauczyć się zależności pomiędzy zmiennymi. Należałoby zbadać obie metody na innym, większym zbiorze danych z niewielką liczbą zmiennych, ale z dużą liczbą obserwacji.

Literatura

Allison, P. D. (2001). *Missing data*. Sage Publications.

Buuren, S. van. (2012). *Flexible Imputation of Missing Data.* Chapman & Hall Book.

Cheng-Xian Li, S. i Marlin, B. (2020). Learning from irregularly-sampled time series: A missing data perspective. *Proceedings of the 37th International Conference on Machine Learning* (s. 5937--5946). PMLR.

Cox, G. M. i Cochran, W. G. (1957). *Experimantal Designs*. John Wiley and Sons.

Grochowina, D. (2014). Wpływ metod imputacji na skuteczność klasyfikacyjną modelu Logitowego zastosowanego do prognozowania upadłości przedsiębiorstw. *Ekonomia*, *45*(2), 187-203.

Hand, D. J., Adèr, H. J. i Mellenbergh, G. J. (2007). *Missing data.* Springer.

Kasprzyk, D. i Kalton, G. (1982). *Imputing for missing survey responses.* American Statistical Association.

Kulpa T. (2013). Metody uzupełniania brakujących danych na przykładzie liczby zarejestrowanych pojazdów. *Transport Miejski i Regionalny* (10).

Little, R. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*(420), 1227-1237.

Little, R. J. (2002). Statistical Analysis with Missing Data. John Wiley & Sons.

Misztal, M. (2011). Próba oceny wpływu wybranych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, (176), 246-253.

Pokropek, A. (2018). Wybrane statystyczne metody radzenia sobie z brakami danych. *Polskie Forum Psychologiczne, 23*(2), 291-310.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-592.

Modern Methods of Imputation of Missing Data – Comparison of Selected Methods

Abstract: The article describes the results of the imputations carried out with the multiple imputation methods of the MICE procedure and using Neural Networks, on a small set containing 690 observations and 6 variables. The author then compares the obtained results and chooses MICE imputation as a better method for filling in data gaps in the studied set.

Keywords: imputation, neural networks, data gaps, MICE algorithm