dr hab. Tomasz Żądło, prof. UE
Katedra Statystyki, Ekonometrii i Matematyki
Uniwersytet Ekonomiczny w Katowicach

Katowice, 22 sierpnia 2023

Recenzja rozprawy doktorskiej
mgra Pawła Lańducha
pt. „Wykorzystanie technik imputacyjnych w szacowaniu informacji wynikowych oraz w analizie struktury danych w statystyce przedsiębiorstw"
napisanej pod kierunkiem dra hab. Andrzeja Młodaka, prof. AK

1. Ocena problemu badawczego i wyboru tematu rozprawy

Tematyka pracy jest aktualna i bardzo ciekawa. Rozważana problematyka jest obecnie jedną z najbardziej dynamicznie rozwijających się w ramach metody reprezentacyjnej. Choć pewne propozycje metodologiczne w zakresie wnioskowania na podstawie prób nielosowych oraz łączenia i imputacji danych w badaniach reprezentacyjnych są znane od dawna, to rozwój tych metod jest niezwykle szybki, a ich znaczenie rośnie. W stale zmieniającej się gospodarce możliwość uzyskania w krótkim czasie i przy niskim koszcie informacji w postaci oszacowań charakterystyk podpopulacji przedsiębiorstw charakteryzujących się wysoką dokładnością ma ogromne znaczenie dla decydentów. Bez wątpienia podjęty problem badawczy może być podstawą przygotowania rozprawy doktorskiej w dziedzinie nauk ekonomicznych, w dyscyplinie ekonomia.

Tytuł pracy obejmuje „szacowanie informacji wynikowych" oraz „analizę struktury danych". Terminy te powinny być omówione przez Doktoranta, gdyż rozważania w rozprawie są ograniczone do szacowania wyłącznie wartości globalnej w podpopulacjach. Bez tych wyjaśnień pojawiają się wątpliwości, czy tytuł pracy nie definiuje problemu badawczego znacznie szerzej, niż jest on rozważany w pracy.

2. Ocena treści rozprawy

Praca liczy 124 strony i uwzględnia wprowadzenie, pięć kilkunastostronicowych rozdziałów, zakończenie, spis literatury, wykaz tablic, wykaz rysunków i wykresów oraz aneks. Choć tematyka pracy jest bardzo interesująca, to praca jest napisana niestarannie, zawiera błędy merytoryczne a z opisu procedury przeprowadzonych badań symulacyjnych wynika, że zostały one przeprowadzone nieprawidłowo. Brak precyzji opisu zagadnień teoretycznych sprawia, że nie ma pewności, czy prezentowane rozwiązania są w pełni autorskie, czy też są modyfikacją metod prezentowanych w literaturze (i jak znaczącą), czy też tylko zastosowaniem metod prezentowanych w literaturze do rozważanego zbioru danych. Stąd niezbędne jest doprecyzowanie treści rozprawy i sformułowań takich jak na s. 89

w zakończeniu („zaproponowane w niniejszej pracy metody [...]") mogących sugerować, że w pracy prezentowane są nowe, w pełni autorskie rozwiązania.

Cel pracy jest różnie definiowany w streszczeniu na s. 4 oraz we **wprowadzeniu** na stronach 11-12. Na s. 4 jest przedstawiony zbyt ogólnie. Cel zapisany na stronach 11-12 jest prezentowany wraz ze zbędnym we wstępie dość szczegółowym opisem jego realizacji w pracy. Ponadto nie jest określony precyzyjnie jako „ocena efektywności zastosowania metod imputacji masowej". Dalsza lektura wstępu (s. 13) i pracy, sugeruje, że Doktorant miał na myśli porównanie własności estymatorów wartości globalnej w warstwach (w tym tych wykorzystujących metody imputacji) za pomocą wariancji i względnego średniego błędu szacunku nazywanego przez Autora współczynnikiem zmienności estymatora, co jest dosłownym tłumaczeniem z języka angielskiego. Tak rozumiany cel byłby określony precyzyjnie, ale nieprawidłowo. Błędem jest porównanie własności estymatorów tylko przy pomocy tych mierników, gdyż to oznacza, że Doktorant pomija obciążenie zaproponowanych w pracy estymatorów, które nie są nieobciążone. Gdyby przyjąć, że takie podejście jest prawidłowe, to lepszym rozwiązaniem od prezentowanych w pracy byłby estymator wartości globalnej przyjmujący wartość 1 zł dla każdej próby – miałby on zerową wariancję. Należy porównywać dokładność estymacji a jej miernikiem jest błąd średniokwadratowy, a nie wariancja. Oprócz opisu celu pracy na stronach 12 i 13 pojawia się też „główny cel analizy", a na s. 56 nieprecyzyjnie określone: „cel badania" („jakościowa ocena możliwości [...]") oraz „ostateczny cel dociekań" („estymacja wartości parametru [...]"). Również w dalszej części rozprawy Autor wielokrotnie zapisuje te same treści. Na s. 14 pojawia się „główna hipoteza badawcza", którą Autor przepisuje na stronach 56 i 89 zamiast tylko odwołać się do tej przedstawionej we wprowadzeniu, oraz hipoteza dodatkowa, którą Autor ponownie prezentuje na s. 89 oraz na s. 57 nazywając ją – w tym drugim przypadku – „hipotezą pomocniczą". Hipoteza badawcza („imputacja masowa [...] może być użytecznym narzędziem") powinna być doprecyzowana. Podobnie nieprecyzyjnie zapisano hipotezę dodatkową. Ponadto należy podkreślić, że we wstępie Autor wielokrotnie niejasno opisuje szczegółowe zagadnienia metodologiczne, które będą prezentowane w dalszej części pracy. Lepszym rozwiązaniem byłby krótszy i bardziej ogólny wstęp z ewentualnymi powołaniami na szczegółowe zapisy w rozdziałach późniejszych.

W **rozdziale pierwszym** za literaturą omówiono zagadnienie badań statystycznych przedsiębiorstw. Rozważano ich cechy, wybrane aspekty metodologiczne, przedstawiono rodzaje badań przedsiębiorstw i wspomniano o badaniach wielookresowych. Nie uwypuklono tu żadnych autorskich propozycji, modyfikacji lub sugestii o charakterze metodologicznym. Niezbędne jest wprowadzenie korekt wymienionych w ocenie formalnej pracy.

Tematyka **rozdziału drugiego** dotyczy integracji danych. Autor omawia za literaturą bardzo ważne z punktu widzenia badań statystycznych i nabierające w ostatnim czasie znaczenia w statystyce publicznej zagadnienia takie jak parowanie statystyczne, deterministyczne i probabilistyczne łączenie rekordów oraz mikrointegrację. W rozdziale Doktorant nie akcentuje własnych propozycji. Konieczne jest wprowadzenie zmian sugerowanych w ocenie formalnej pracy.

Rozdział trzeci dotyczy imputacji masowej ze szczególnym uwzględnieniem łączenia prób losowych oraz losowych z nielosowymi. Ważnym z punktu widzenia dalszej części pracy jest podrozdział 3.4, gdyż tam opisano problem integracji próby losowej i nielosowej rozważany w badaniach symulacyjnych. Kluczowe zagadnienia są prezentowane za Yang, Kim, & Hwang (2021, s. 53–55), ale powołanie na ten artykuł (nie licząc źródła tabeli 6) pojawia się na tych stronach tylko raz – na stronie 53. To sprawia, że kwestia autorstwa prezentowanych tam treści (w tym założeń 1 i 2 na s. 54) jest niejasna. Rozdział trzeci powinien też uwzględniać szersze studia literaturowe, które można byłoby zacząć choćby na podstawie bibliografii wielokrotnie cytowanego przez Doktoranta artykułu Yang, Kim, & Hwang (2021).

Na stronie 49 prezentowany jest estymator Horvitza-Thompsona wartości globalnej w populacji oraz estymator jego wariancji (oba bez powołania na literaturę), przy czym błędnie zapisano, że jest to jego wariancja. Skoro Autor rozważa w pracy problem estymacji wartości globalnej w podpopulacji, to powinien zostać zaprezentowany (dodatkowo lub wyłącznie) estymator Horvitza-Thompsona wartości globalnej w podpopulacji. Ponadto w przypadku estymatorów nieobciążonych zapis powinien uwzględniać (por. Särndal, Swensson, & Wretman, 1992, s. 42–43):

- postać estymatora wraz z informacją, że jest nieobciążony,
- postać wariancji estymatora,
- postać estymatora wariancji lub estymatorów wariancji oraz prezentację ich własności (np. informację, że są nieobciążone lub asymptotycznie nieobciążone, lub że są obciążone, ale badania symulacyjne prezentowane w literaturze wskazują na określone zalety).

Wzory, opis własności estymatora, opisy własności estymatorów wariancji o charakterze teoretycznym (np. nieobciążoność) lub wynikających z badań symulacyjnych powinny być poparte powołaniami na literaturę.

Na stronie 50 w wierszu 6 licząc od dołu strony jest prezentowany estymator wartości globalnej w populacji o postaci zbliżonej do postaci estymatora Horvitza-Thompsona. Autor powinien zaprezentować (dodatkowo lub wyłącznie) postać estymatora wartości globalnej w podpopulacji, gdyż taki problem jest rozważany w pracy. Ta uwaga dotyczy też innych estymatorów prezentowanych w rozprawie. Ponadto w przypadku estymatorów obciążonych zapis powinien uwzględniać:

- postać estymatora wraz z informacją, że jest obciążony (oraz, jeśli to prawda, że np. jest asymptotycznie nieobciążony lub zgodny, lub że badania symulacyjne prezentowane w literaturze wskazują na określone zalety),
- (przybliżoną) postać błędu średniokwadratowego estymatora, jeśli jest znana (jeśli nie jest znana, to należy o tym wspomnieć),
- postać estymatora lub estymatorów błędu średniokwadratowego oraz prezentację ich własności (np. informację, że jest nieobciążony lub asymptotycznie nieobciążony, lub że jest obciążony, ale badania symulacyjne prezentowane w literaturze wskazują na określone zalety).

Podobnie jak w przypadku estymatorów nieobciążonych wzory i opisy własności powinny być poparte powołaniami na literaturę. Jeśli rozważany obciążony estymator jest asymptotycznie nieobciążony, to zamiast błędu średniokwadratowego i jego estymatora można przedstawić postać wariancji (ewentualnie przybliżoną postać wariancji) i postać estymatora lub estymatorów wariancji jak np. w Deville & Särndal (1992, s. 379–380). Wówczas jednak niezbędne jest zdefiniowanie, co oznacza „asymptotycznie", jakie założenia zostały przyjęte oraz sprawdzenie w badaniach symulacyjnych, czy – mimo asymptotycznej nieobciążoności – obciążenie estymatora w rozważanych przypadkach, w tym dla rozważanej liczebności próby, nie jest duże. W cytowanym przez Doktoranta artykule Yang, Kim, & Hwang (2021) tak wygląda prezentacja np. estymatora wykorzystującego wartości badanej zmiennej pochodzące od najbliższego sąsiada. Powyższe uwagi powinny zostać uwzględnione w przypadku wszystkich estymatorów prezentowanych w pracy.

Gdyby do estymacji błędu średniokwadratowego estymatora prezentowanego na stronie 50 i podobnych, użyto estymatorów wariancji podobnych do estymatorów wariancji stosowanych w przypadku estymatora Horvitza-Thompsona, oznaczałoby to zignorowanie faktu, że rzeczywiste wartości badanej zmiennej są zastąpione wartościami imputowanymi, a więc zignorowanie obciążenia estymatorów wartości globalnej. Autor może przedstawić takie podejście, koniecznie pisząc o jego wadach, ale jeśli celem takiego rozwiązania będzie analiza jego własności w badaniu symulacyjnym (zob. wzory (3) i (4) w dalszej części recenzji). Należałoby jednak też spróbować rozważyć estymatory błędu średniokwadratowego i nie ignorować obciążenia estymatora wartości globalnej. Można byłoby tu wykorzystać prosty estymator błędu średniokwadratowego dany wzorem (3.2.16) w Rao & Molina (2015, s. 44), gdzie estymator syntetyczny byłby zastąpiony jednym z rozważanych przez Doktoranta obciążonych estymatorów wykorzystujących ideę imputacji.

Zgodnie z tytułem podrozdziału 3.4 powinien on omawiać problem integracji próby losowej i nielosowej. Kończy się on prezentacją postaci estymatora wartości globalnej w populacji wykorzystującego tylko informację z próby losowej, a brakuje postaci estymatora wartości globalnej wykorzystującego ideę imputacji dla tego problemu. Podrozdział ten w obecnej jego formie wygląda na niedokończony.

Rozdział czwarty powinien zostać podzielony na dwie części w celu poprawy struktury pracy. Ogólne rozważania teoretyczne dotyczące imputacji masowej prezentowane w podrozdziale 4.3 należałoby przesunąć do rozdziału trzeciego. Tam też powinny się znaleźć opisy wykorzystywanych estymatorów przygotowane zgodnie z wytycznymi zaprezentowanymi powyżej. Można wtedy również rozważyć zmianę tytułu tego rozdziału, aby lepiej oddawał jego nową zawartość. Natomiast opis procedury badania symulacyjnego znajdujący się w rozdziale czwartym znalazłby się w obecnym rozdziale piątym (który byłby w nowej wersji pracy rozdziałem czwartym).

Obecnie opis zastosowania metod imputacji masowej uwzględnia:

- pierwszy podrozdział 4.3.2 (nie ma podrozdziału 4.3.1, są natomiast dwa podrozdziały 4.3.2), którego zawartość w większości pochodzi z Yang, Kim, & Hwang (2021, s. 34),

w szczególności opis „etapu 1" i „etapu 2", lecz nie zostało to czytelnie określone przez Autora,

- drugi podrozdział 4.3.2, którego idea pochodzi z Yang, Kim, & Hwang (2021, s. 36), lecz nie zostało to jasno zapisane (w tym sformułowanie „w niniejszej pracy metodę tę zastosowano w następujący sposób" nie uwzględnia powołania na literaturę, opis „etapu 1" i etapu 2" również bez powołania na literaturę),
- podrozdziały 4.3.3 i 4.3.4, które są prezentowane, jakby były autorskimi modyfikacjami Doktoranta polegającymi na wykorzystaniu omawianych tam metod imputacji w metodzie integracji próby losowej z nielosową (brak jednak jednoznacznej deklaracji Autora w tym zakresie popartej powołaniami na literaturę).

Treść wymienionych powyżej podrozdziałów wymaga doprecyzowania. W szczególności warto byłoby zaprezentować wyniki szerszych studiów literaturowych, aby na ich tle wyraźniej uwypuklić autorskie modyfikacje.

Poziom szczegółowości opisu części dotyczącej opisu projektu badania symulacyjnego powinien być taki, aby pozwolić czytelnikowi mającemu dostęp do danych wejściowych użytych w rozprawie na samodzielne odtworzenie przeprowadzonych badań. Powinien on uwzględniać:

a) czytelny, szczegółowy opis generowania sztucznej populacji naśladującej oryginalny zbiór danych,
b) czytelny opis użytego schematu losowania próby losowej i sposobu doboru próby nielosowej,
c) odwołania do numerów wzorów prezentujących wykorzystywane estymatory wartości globalnej w podpopulacjach, które będą przedstawione w nowej wersji rozdziału trzeciego, a których własności będą analizowane w badaniu symulacyjnym np. z wykorzystaniem omawianych poniżej mierników (1) i (2),
d) odwołania do numerów wzorów prezentujących wykorzystywane estymatory błędów średniokwadratowych rozważanych estymatorów, które będą przedstawione w nowej wersji rozdziału trzeciego, a których własności będą analizowane w badaniu symulacyjnym np. z wykorzystaniem omawianych poniżej mierników (3) i (4),
e) w przypadku estymatora Horvitza-Thompsona odwołania do numerów wzorów prezentujących jego estymatory wariancji, które będą przedstawione w nowej wersji rozdziału trzeciego, a których własności będą analizowane w badaniu symulacyjnym np. z wykorzystaniem omawianych poniżej mierników (5) i (6),
f) odwołania do wzorów prezentujących przedziały ufności (jeśli przedziały ufności będą rozważane w badaniu symulacyjnym),
g) opis każdego kroku badania symulacyjnego, w tym informacja o statystykach, których wartości są liczone w każdej iteracji,
h) wzory przedstawiające mierniki pozwalające na symulacyjne porównanie dokładności estymatorów wartości globalnej,
i) wzory przedstawiające mierniki pozwalające na symulacyjne porównanie dokładności estymatorów błędów średniokwadratowych.

Część z wymienionych punktów jest ujęta w pracy, ale precyzja opisu nie jest wystarczająca (nie pozwala na odtworzenie badania przy założeniu, że czytelnik ma dostęp do rozważanych danych). Miernikami wymienionymi w punktach h) oraz i) mogłyby być:

- przybliżona symulacyjnie wartość względnego obciążenia estymatora (ang. Relative Bias):

$$RB^{(sym)}(\hat{\theta}) = \frac{\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} - \theta}{|\theta|}100\%, \quad (1)$$

gdzie B to liczba powtórzeń w badaniu symulacyjnym, $\hat{\theta}^{(b)}$ wartość estymatora wartości globalnej w warstwie w b-tym kroku badania symulacyjnego, θ – szacowany parametr, czyli wartość globalna w warstwie,

- przybliżona symulacyjnie wartość względnego RMSE estymatora (ang. Relative RMSE)

$$RRMSE^{(sym)}(\hat{\theta}) = \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \theta\right)^2}}{|\theta|}100\%, \quad (2)$$

gdzie wszystkie oznaczenia wyjaśniono pod wzorem (1),

- przybliżona symulacyjnie wartość względnego obciążenia estymatora błędu średniokwadratowego estymatora:

$$RB^{(sym)}\left(\widehat{MSE}\left(\hat{\theta}\right)\right) = \frac{\frac{1}{B}\sum_{b=1}^{B}\widehat{MSE}^{(b)} - MSE^{(sym)}}{MSE^{(sym)}}100\%, \quad (3)$$

gdzie B to liczba powtórzeń w badaniu symulacyjnym, $\widehat{MSE}^{(b)}$ wartość estymatora błędu średniokwadratowego rozważanego estymatora wartości globalnej w b-tym kroku badania symulacyjnego, $MSE^{(sym)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \theta\right)^2$ jest przybliżonym symulacyjnie błędem średniokwadratowym rozważanego estymatora a oznaczenia wyjaśniono pod wzorem (1),

- przybliżona symulacyjnie wartość względnego RMSE estymatora błędu średniokwadratowego estymatora:

$$RRMSE^{(sym)}\left(\widehat{MSE}\left(\hat{\theta}\right)\right) = \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{MSE}^{(b)} - MSE^{(sym)}\right)^2}}{MSE^{(sym)}}100\%, \quad (4)$$

gdzie wszystkie oznaczenia wyjaśniono pod wzorem (3).

W przypadku estymatora Horvitza-Thompsona, który w obecnej wersji pracy jako jedyny jest estymatorem nieobciążonym, zamiast mierników (3) i (4) byłyby rozważane:

- przybliżona symulacyjnie wartość względnego obciążenia estymatora wariancji estymatora:

$$RB^{(sym)}\left(\widehat{D}^2(\hat{\theta})\right) = \frac{\frac{1}{B}\sum_{b=1}^{B}\widehat{D}^{2(b)} - D^{2(sym)}}{D^{2(sym)}}100\%, \quad (5)$$

gdzie B to liczba powtórzeń w badaniu symulacyjnym, $\widehat{D}^{2(b)}$ wartość estymatora wariancji estymatora wartości globalnej w warstwie w b-tym kroku badania

symulacyjnego, $D^{2(sym)} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}\right)^2$ jest przybliżoną symulacyjnie wariancją rozważanego estymatora,

- przybliżona symulacyjnie wartość względnego RMSE estymatora wariancji estymatora:

$$RRMSE^{(sym)}\left(\widehat{D}^2(\hat{\theta})\right) = \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{D}^{2(b)} - D^{2(sym)}\right)^2}}{D^{2(sym)}}100\%, \qquad (6)$$

gdzie wszystkie oznaczenia wyjaśniono pod wzorem (5).

Gdyby w badaniu symulacyjnym analizowano również własności przedziałów ufności, to wzory pozwalające na wyznaczenie przedziałów ufności dla danej próby musiałyby być ujęte w pracy w nowej wersji rozdziału trzeciego (teraz nie są, choć wyniki są prezentowane). Wówczas przy prezentacji wyników należałoby ująć mierniki zazwyczaj wykorzystywane w literaturze w tym przypadku takie jak:

- odsetek iteracji, w których przedział ufności pokrył nieznaną w praktyce (ale znaną w badaniu symulacyjnym) wartość parametru (ang. coverage rate)
- oraz przybliżoną symulacyjnie wartość oczekiwaną długości przedziału ufności.

Podobnie zaprojektowane jest badanie symulacyjne prezentowane w Yang, Kim, & Hwang (2021), która to publikacja jest wielokrotnie cytowana w rozprawie, ale Doktorant nie korzysta z niej do zaprojektowania własnego badania. W szczególności w badaniach Monte Carlo w tym artykule przybliżane są symulacyjnie wartości obciążeń analizowanych estymatorów i ich średnie błędy szacunku, co pozwala na wyznaczenie błędów średniokwadratowych. Aby była możliwa weryfikacja prawidłowości przeprowadzonych badań, niezbędne jest też zaprezentowanie programu wykorzystanego do ich przeprowadzenia jako załącznika pracy lub uwzględnienie linku do repozytorium z tym programem. W kolejnych akapitach zostaną dodatkowo przedstawione bardziej szczegółowe uwagi dotyczące rozdziału czwartego, pokazujące konieczność kompleksowego przeredagowania tekstu tego rozdziału zgodnie z powyższymi sugestiami dotyczącymi prezentacji wzorów opisujących estymatory oraz opisu badania symulacyjnego.

Prezentowany na s. 61 sposób generowania wartości badanej zmiennej nie jest jasny. Przedstawione modele nie uwzględniają parametru stojącego przy zmiennej objaśniającej. Mimo deklaracji, że dane „generowano warstwowo", to parametry modeli przyjmują takie same wartości w warstwach. Nie wiadomo też, jakie wartości parametrów przyjęto, na jakiej podstawie i dlaczego takie same w przypadku obu zmiennych objaśnianych, jak podano we wzorach prezentowanych na s. 61. Składnik losowy oznaczany przez $\epsilon_{h,i}$ jest wprowadzany dwukrotnie z różnymi wariancjami zakładanymi arbitralnie. W drugim przypadku (ostatni akapit na s. 61) Doktorant uzasadnia wybór wartości parametru wartością przyjętą w artykule Yang, Kim, & Hwang (2021), gdzie rozważano inny model, inne dane i badaną zmienną w innych jednostkach. Skoro Doktorant wzoruje sztuczną populację na próbkowych danych rzeczywistych, to naturalnym wyborem byłoby przyjęcie za wartość wariancji generowanego składnika losowego wartości wariancji resztowej uzyskanej na podstawie danych rzeczywistych i przy założeniu rozważanego modelu. Można byłoby też dodatkowo rozważyć

inny rozkład niż normalny z odpowiednio dobranymi parametrami. Nie wiadomo też, czym jest „wartość z modelu" zapisana słownie w ostatnim wzorze na s. 61 (ten termin pojawia się też w innych miejscach pracy). Powinna być zapisana odpowiednim wzorem. Jeśli byłaby to wartość badanej zmiennej wygenerowana z wykorzystaniem modeli prezentowanych w wierszach 6, 9 i 10 na s. 61 oznaczałoby to, że dwukrotnie są uwzględniane składniki losowe generowane z niezależnych rozkładów normalnych z zerową wartością oczekiwaną i różnymi wariancjami. Ponadto wygląda na to, że sposób generowania danych zakłada, że przychody ze sprzedaży w jednym przedsiębiorstwie nie zależą od przychodów ze sprzedaży w innym przedsiębiorstwie (podobnie wynagrodzenia brutto).

Opis badania symulacyjnego w rozdziale 4 jest nieprecyzyjny. Na s. 62 w wierszach 6-7 napisano, że próba losowa (próba A) to próba warstwowa, przy czym w każdej warstwie stosowano „schemat PPS". Takich schematów w wydanej już 40 lat temu monografii Brewer & Hanif (1983) jest kilkadziesiąt, a część z nich oraz nowsze są oprogramowane np. w pakiecie `sampling` programu R. Są wśród nich też takie, dla których są znane prawdopodobieństwa inkluzji drugiego rzędu. Nie wiadomo, który z tych schematów losowania jest wykorzystywany przez Doktoranta. Należy też zwrócić uwagę, że wiele programów komputerowych pozwala na losowanie kolejnych elementów do próby w każdym kolejnym kroku z prawdopodobieństwami proporcjonalnymi do wartości cechy dodatkowej, ale nie są to algorytmy zapewniające prawdopodobieństwa inkluzji pierwszego rzędu proporcjonalne do wartości cechy dodatkowej. Na s. 62 Doktorant niejasno opisuje dobór elementów populacji do próby nielosowej (próby B) – nie definiuje kryterium ujęcia elementu populacji w tej próbie. Ponadto nie zostały wyjaśniane wszystkie oznaczenia, użyte we wzorach na s. 62. Nie podano również, czym są parametry we wzorze w pierwszym wierszu na stronie 63. Wzory prezentowane na stronie 62 i 63 sugerują, że dobór elementów populacji do próby nielosowej ma charakter losowy. Nie wiadomo też, czy ma on miejsce w każdej iteracji badania symulacyjnego, czy jednorazowo przed pierwszą iteracją.

Nie mamy pewności, jak wyglądały kolejne kroki iteracyjnej procedury badania, krótko wymienione na s. 64. Skrócony opis badania symulacyjnego pojawia się po raz pierwszy już na s. 16 i sugeruje, że zostało ono przeprowadzone nieprawidłowo. Wynika z niego, że miernikiem użytym do porównań dokładności estymatorów nie był błąd średniokwadratowy estymatora ani nawet wariancja estymatora (która byłaby użyta nieprawidłowo ze względu na obciążenie estymatorów), ale wartość oczekiwana estymatora względnego średniego błędu szacunku. Następnie ten problem jest rozważany na stronie 64. Choć oznaczenia użyte w ostatnim wzorze na tej stronie nie są wyjaśnione, to biorąc pod uwagę opis na stronie 16 należy przypuszczać (mimo że nad „var" Doktorant nie zapisał „daszka", który jest w pierwszym wzorze na stronie 64), że miał to być wzór na wspomnianą symulacyjnie przybliżoną wartość oczekiwaną estymatora względnego średniego błędu szacunku. Opis projektu badania symulacyjnego powinien zostać przedstawiony tak, jak opisano powyżej.

W rozprawie daje się zauważyć brak zrozumienia problemu obciążenia estymacji. W rozdziale trzecim na stronie 48 w pierwszych 4 wierszach napisano o konieczności oceny precyzji (estymacji wariancji estymatora), ignorując problem obciążenia estymatora, a że

obciążenie jest problemem Autor stwierdza np. na s. 51: „[...] w przypadku próby nielosowej potraktowanie jej jak jako losowej obciąża wyniki". Podobnie na s. 65 podaje za Yang, Kim, & Hwang (2021), że asymptotyczne obciążenie estymatora wartości globalnej w populacji wykorzystującego informacje o badanej zmiennej od najbliższego sąsiada może być zignorowane (jednak – co ważne – przy pewnych założeniach), ale nie podaje, że Autorzy pokazują w badaniach symulacyjnych, że obciążenie może być bardzo wysokie. Moduł obciążenia w niektórych przypadkach był tam większy a nawet kilkadziesiąt razy większy od średniego błędu szacunku. Ponadto obciążenie tego estymatora wartości globalnej jest rzędu zależącego od liczebności próby nielosowej, a w rozprawie rozważane są estymatory wartości globalnej nie w populacji, ale w podpopulacjach. To oznacza, że można spodziewać się jeszcze wyższych obciążeń. W przypadku wnioskowania na podstawie raz wylosowanej próby (nie badań symulacyjnych) Autor powinien pisać o estymacji dokładności estymatora (estymacji błędu średniokwadratowego estymatora). Natomiast w przypadku badań symulacyjnych powinien rozważać dokładność estymatora (zob. wzór (2)) i dokładność estymatorów błędu średniokwadratowego (zob. wzór (4)).

Na s. 56 w wierszach 20-21 Doktorant podkreśla, że ważnym elementem jakości informacji prezentowanych w statystyce publicznej jest precyzja oszacowań, ponownie pomijając fakt, że jeśli wykorzystywane są estymatory obciążone (jak te analizowane w rozprawie), to powinna być oceniana nie precyzja, a dokładność estymacji. Nawet gdyby wszystkie rozważane w pracy estymatory wartości globalnej były nieobciążone i wystarczające byłoby porównanie precyzji, to opis nadal nie byłby prawidłowy. Autor na s. 56 w wierszach 21-23 deklaruje, że „w obecnym rozdziale omówione zostaną rezultaty przeprowadzonej analizy porównawczej wariancji estymatora Horvitza-Thompsona po wykonaniu imputacji masowej każdej z metod". Wyniki prezentowane są jednak w kolejnym rozdziale. Choć stwierdza, że porównywana będzie wariancja, to na s. 64 prezentuje estymator wariancji (wariancja to nie to samo co estymator wariancji). To, że wartość estymatora wariancji estymatora A jest mniejsza od wartości estymatora wariancji estymatora B, nie znaczy, że wariancja estymatora A jest mniejsza od wariancji estymatora B. Nawet gdyby porównania w dalszej części pracy były wykonane dla jednej próby, a nie w badaniach symulacyjnych, to wykorzystanie prezentowanego na s. 64 estymatora wariancji estymatora Horvitza-Thompsona zaproponowanego w Deville & Tillé (2005) dla planu losowania zbalansowanego w celu estymacji precyzji różnych estymatorów dla planu losowania próby warstwowej z różnymi prawdopodobieństwami inkluzji wymaga uzasadnienia. Po pierwsze nie napisano, którym wzorem w artykule Deville & Tillé (2005) zapisano prezentowany w pracy estymator wariancji. Po drugie rozważane w pracy estymatory wartości globalnej mają postać podobną do estymatora Horvitza-Thompsona, ale nim nie są. Po trzecie Doktorant nie podaje, czy Deville & Tillé (2005) informują, że można stosować ten estymator wariancji dla innych planów losowania niż losowanie zbalansowane. W innych artykułach Prof. Tillé można znaleźć estymatory wariancji estymatora Horvitza-Thompsona wykorzystujące tylko informacje o prawdopodobieństwach inkluzji pierwszego rzędu i możliwe do zastosowania dla dowolnych planów bezzwrotnego losowania próby z różnymi prawdopodobieństwami inkluzji pierwszego

rzędu. Podsumowując Doktorant do oceny wariancji proponowanego estymatora A dla planu losowania B (z dodatkowym mechanizmem imputacji) proponuje bez uzasadnienia wykorzystać estymator wariancji estymatora C stosowany dla planu losowania D.

Badania symulacyjne prezentowane w **rozdziale piątym** muszą zostać przeprowadzone ponownie i uwzględniać uwagi do rozdziału 4 zaprezentowane powyżej. Ponadto Doktorant porównuje własności proponowanych metod wyłącznie z estymatorem Horvitza-Thompsona, pisząc na stronie 88, że warunkiem wykorzystania rozważanych w pracy metod „było uzyskanie poziomów precyzji porównywalnych do złotego standardu, jakim jest próba losowa i schemat losowania". Doktorant chce więc porównywać precyzję zamiast dokładności (i ignorować obciążenie). Ponadto porównuje własności estymatora Horvitza-Thompsona wykorzystującego tylko informacje o badanej zmiennej z proponowanymi metodami wykorzystującymi, przynajmniej pośrednio, wartości zmiennej dodatkowej. Stąd w badaniach symulacyjnych powinno się też uwzględnić klasyczne metody szacowania charakterystyk podpopulacji wykorzystujące informacje o zmiennych dodatkowych. Po pierwsze należałoby wziąć pod uwagę jeden z estymatorów kalibrowanych prezentowanych w Deville & Särndal (1992). Po drugie powinny być uwzględnione metody wykorzystywane w podejściu modelowym w metodzie reprezentacyjnej, choćby najlepszy liniowy nieobciążony predyktor prezentowany przez Żądło (2008, s. 96) a dany tam wzorem (4.6), w ogólniejszej formie przedstawiony w Royall (1976, s. 658) i Valliant, Dorfman, & Royall (2000, s. 30). Przy przyjętych tam założeniach predyktor ten w przypadku problemu predykcji wartości globalnej w podpopulacji jest wyznaczany jako suma dwóch elementów – sumy wartości badanej zmiennej w próbie w podpopulacji (tu można uwzględnić zarówno próbę losową i nielosową) i sumy wartości teoretycznych badanej zmiennej dla niewylosowanych elementów podpopulacji.

Przyjęta przez Doktoranta liczba powtórzeń w badaniu symulacyjnym (wynosząca 500) wydaje się niewielka w porównaniu z przyjmowaną w literaturze. Doktorant nie uzasadnia, że jest ona wystarczająca pod względem jakiegoś ustalonego kryterium. Można byłoby przykładowo przyjąć, że liczba iteracji jest wystarczająca, gdy symulacyjne względne obciążenia (wyznaczane zgodnie ze wzorem (1)) rozważanych nieobciążonych estymatorów Horvitza-Thompsona w przypadku estymacji wartości globalnej w każdej z warstw nie przekraczają określonej, bardzo niskiej wartości.

Część wyników symulacyjnych jest prezentowana **w załączniku**. Nagłówek czwartej kolumny każdej z tabel jest zapisany nieprecyzyjnie jako „oszacowana wartość zmiennej". Tabele wynikowe powinny zawierać wartości mierników danych wzorami od (1) do (6). Nie wiadomo też, jakim wzorem były dane rozważane przedziały ufności i jak wyznaczano wartości dolnego i górnego końca przedziału ufności w każdej iteracji badania symulacyjnego. W tabelach podano ich konkretne wartości, podczas gdy w każdej iteracji symulacji wyniki te, jeśli były to faktycznie końce przedziałów ufności, musiały być różne.

Bibliografia wymaga korekty. Brakuje jednolitego zapisu bibliograficznego monografii, rozdziałów w monografiach i artykułów w czasopismach. Przykładowo rozważmy tylko pierwsze trzy publikacje na s. 99, które są artykułami w czasopismach. W przypadku tylko tych

trzech publikacji Doktorant nie stosuje jednolitego zapisu tytułu czasopisma, jednolitego zapisu tomu i numeru czasopisma, jednolitego zapisu numeru stron, znaki interpunkcyjne też nie są stosowane jednolicie (np. po dacie wydania i na końcu zapisu bibliograficznego).

Literatura

Brewer, K. R. W., & Hanif, M. (1983). *Sampling With Unequal Probabilities*. New York: Springer. https://doi.org/10.1007/978-1-4684-9407-5

Deville, J.-C., & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. https://doi.org/10.1080/01621459.1992.10475217

Deville, J.-C., & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, *128*(2), 569–591. https://doi.org/10.1016/j.jspi.2003.11.011

Rao, J. N. K., & Molina, I. (2015). *Small area estimation. Second edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Royall, R. M. (1976). The Linear Least-Squares Prediction Approach to Two-Stage Sampling. *Journal of the American Statistical Association*, *71*(355), 657–664. https://doi.org/10.1080/01621459.1976.10481542

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach* (1st edition). New York: Wiley-Interscience.

Yang, S., Kim, J. K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, *47*(1), 29–58.

Żądło, T. (2008). *Elementy statystyki małych obszarów z programem R*. Katowice: Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego.

3. Ocena formalnej strony rozprawy

W wielu miejscach rozprawy **brakuje powołań na literaturę** np. na s. 25 w wierszach 14-15 („alternatywna metoda PPS bez zwracania ze stałą licznością próby" – nie ma pewności, czy jest to metoda omawiana w monografii podanej w zdaniu wcześniejszym i na której spośród ponad 700 stron), na s. 29 w wierszu 2, na s. 32 w wierszu 17, na s. 36 w wierszu 14 (także w innych miejscach pracy znajdują się listy punktowane prezentowane za literaturą lub inspirowane literaturą, ale bez powołań na odpowiednie publikacje). Gdy Doktorant powołuje się już na literaturę, to zazwyczaj nie podaje numerów stron nawet w przypadku monografii (np. źródło tabeli 3 na s. 34). Zdarzają się też powołania na publikacje, których nie uwzględniono lub zapisano inaczej w bibliografii np. na s. 31 „Lynn (2008)" i na s. 35: „(ESS ADMIN, 2020)". Zdarzają się też nieprawidłowo zapisane powołania na literaturę, np. na s. 44 są w całości zapisane w nawiasach, chociaż powinny być zapisane standardowo. Powyżej

wymieniono tylko kilka przykładów, ale niezbędne jest poprawienie całej pracy pod tym względem.

W pracy **wzory** zapisywane są niestarannie: brakuje wyśrodkowania, znaków interpunkcyjnych po wzorach (np. s. 38), a brak numeracji wzorów wpływa negatywnie na czytelność powołań na wcześniejsze fragmenty pracy. Na precyzję rozważań pozytywnie wpłyną powołania na literaturę wraz z numerem strony, z której dany wzór pochodzi, uwzględnianie w nawiasie przed dwukropkiem bezpośrednio nad wzorem. Dzięki temu unikamy wątpliwości, kto jest autorem danego wzoru. Autor nie zawsze wyjaśnia wszystkie oznaczenia w prezentowanych wzorach. Przykładowo w ostatnim wzorze na stronie 64 pojawiają się symbole $\hat{Y}_{h,i}$ oraz $var(\hat{Y}_{h,i})$, które nie są wyjaśnione ani poniżej, ani wcześniej. Zdarza się też, że stosowane są podobne, ale nieidentyczne oznaczenia bez wyjaśnień, z czego wynika różnica. Na przykład na s. 64 badana zmienna jest zapisywana dużą literą (jak w statystyce matematycznej oznacza się zmienne losowe), a na kolejnej stronie już małą (jak oznacza się realizacje zmiennych losowych). Często też nazwy odpowiednich statystyk czy mierników nie są prawidłowe. Przykładowo w ostatnim wierszu na stronie 63 napisano, że „wariancja estymatora wyrażona jest wzorem". Po pierwsze, precyzja zapisu wymaga ujęcia informacji, jakiego parametru estymator jest rozważany lub którym wzorem jest dany ten estymator. Po drugie na stronie 64 w pierwszym wzorze przedstawiono nie wariancję estymatora (jak deklaruje Doktorant), ale estymator wariancji estymatora. Ponadto na s. 65 napisano „wartość globalna zostaje wyznaczona za pomocą wzoru", a poniżej prezentowany jest estymator, a nie – zgodnie z opisem – parametr, czyli że wzór przedstawia nie wartość globalną, ale estymator wartości globalnej. Podobne błędne zapisy pojawiają się w innych częściach pracy.

Brak precyzji opisu jest istotnym problemem w pracy. Wcześniej zwrócono na to uwagę w przypadku prezentacji kluczowych rozważań teoretycznych, ale przykłady obejmują też między innymi:

- s. 5 w wierszu 3: „500 symulacji",
- w ostatnim wierszu na s. 11: „z wygenerowanej w ten sposób populacji", podczas gdy Doktorant nie przedstawia, w jaki sposób populacja została wygenerowana (sugeruję odwołanie do odpowiedniego podrozdziału),
- s. 12 w wierszu 11: „Za pomocą imputacji wykonana będzie estymacja wartości globalnych",
- s. 12 i 13: „Wyznaczenie wartości globalnych [...] w oparciu o estymator [...]" (estymator nie pozwala na wyznaczenie nieznanych w praktyce wartości parametrów populacji, ale na ich oszacowanie),
- na stronie 21 w wierszach 10-15: „tytuł książki [...] mogący stanowić uzasadnienie dla postrzegania dynamiki populacji przedsiębiorstw jako [...]",
- s. 25 w wierszu 7: „10% próba" zamiast „próba o liczebności stanowiącej 10% liczebności populacji",
- s. 25 w wierszu 17: „Nieobciążony estymatorem wartości w populacji jest [...]" (brak określenia szacowanego parametru),

- s. 26 w wierszach 16-17: „asymetrii rozkładu populacji przedsiębiorstw" (termin asymetria rozkładu dotyczy zmiennej, a nie zbioru),
- s. 30 w wierszach 4 i 5 licząc od dołu strony: „w przywołanym już rozporządzeniu 2019/2152",
- s. 33 w wierszach 21-22: „oszacowania łącznego rozkładu parametrów" (w podejściu randomizacyjnym szacowane charakterystyki populacji i podpopulacji nie są losowe),
- s. 34: „Uwaga: Szary obszar oznacza zmienne nieobserwowane łącznie w obu zbiorach" (podczas gdy np. wartości zmiennych $Y_1, \ldots, Y_Q$ nie są obserwowane w zbiorze B, ale są obserwowane w zbiorze A),
- s. 35 w wierszach 12 i 13: „parametry rozkładu wielkości prawdopodobieństw można szacować o klasyczne tablice kontyngencji",
- s. 35 w wierszu 3 licząc od dołu strony: „definicja ta pochodzi z projektu Eurostatu",
- na s. 44 Doktorant pisze o losowaniu dwustopniowym a ma raczej na myśli losowanie dwufazowe stosowane w przypadku braków odpowiedzi (opis nie jest jasny; pojęcie losowania dwustopniowego, a także dwuetapowego pojawia się również w innych miejscach rozprawy),
- s. 46 w wierszu 10: „na poziomie niższych domen",
- na stronach 47 i 49 pojawiają się pojęcia estymatora bezpośredniego i estymacji bezpośredniej bez wcześniejszego zdefiniowania tych pojęć,
- na s. 49 Doktorant używa pojęć prawdopodobieństwa pierwszego rzędu i prawdopodobieństwa drugiego rzędu zamiast prawdopodobieństwa inkluzji pierwszego rzędu i prawdopodobieństwa inkluzji drugiego rzędu,
- błędnie zapisana definicja prawdopodobieństw inkluzji rzędu drugiego na s. 49 (pojawia się zmienna losowa J_i),
- s. 50 w wierszu 6 licząc od dołu strony wprowadzono wagi $d_{B,i}$, podczas gdy dodatkowy subskrypt *B* jest zbędny, gdyż sumowanie odbywa się po zbiorze *B* (a na stronie 49 zdefiniowano wielkości d_i),
- s. 51 w wierszu 12: „obciąża wyniki",
- cel zapisany na s. 53 w wierszach 10-11 nie jest zapisany prawidłowo,
- s. 54 w pierwszym zdaniu pierwszego akapitu: „funkcją rozkładu",
- na s. 61 w wierszach 4 i 5 licząc od dołu stronu brakuje uzasadnienia zdania „zdecydowano się na takie parametry szumu z uwagi na [...] asymetrię rozkładu cech",
- wielokrotnie w pracy poruszana jest kwestia asymetrii rozkładu rozważanych cech, ale brak jest informacji, jakie są wartości współczynników asymetrii dla rozważanych zmiennych w poszczególnych warstwach i w całej populacji (choć podaje wartości współczynników zmienności w rozdziale 5),
- w podrozdziale 4.2 na s. 63 w wierszach 10-11: „przyjmuje się założenia nr 1 i nr 2 sformułowane w pkt. 4.3" (nie sprecyzowano, co oznacza „pkt. 4.3")
- wzór na s. 63 w wierszu 13 nie jest zapisany prawidłowo,
- s. 64: „asymptotyczne obciążenie rozumiane jest tutaj w ten sposób, że wartość oczekiwana obciążenia zmierza do zera",

- na s. 66 stwierdzono, że „liczba *k* jednak nie może być zbyt duża - należy bowiem zapewnić możliwość efektywnej kontroli obciążenia estymatora" a następnie na stronie 67, że „można zastosować metodę opartą na praktycznej ocenie danych", ale nie podano żadnych szczegółów,
- greckie litery są stosowane w jednym miejscu pracy do oznaczenia parametrów modelu a w innym miejscu jako symbole estymatorów parametrów modelu,
- założono, że składnik losowy w modelu prezentowanym w drugim równaniu na stronie 68 ma rozkład normalny standardowy,
- w pierwszym wierszu na stronie 69 wspomina się o węzłach wewnętrznych i granicznych, ale te terminy nie zostały wcześniej wprowadzone,
- s. 70 w pierwszym zdaniu drugiego akapitu: „próby nielosowej – tzn. niestatycznego zbioru danych",
- s. 74: „niestatystyczne źródło danych" (brak informacji o fachowej literaturze z zakresu metody reprezentacyjnej, gdzie ten termin jest definiowany),
- na s. 90 w wierszu 4 napisano „zaproponowanie [...] wybranych czterech metod imputacji masowej", co sugeruje, że Doktorant jest autorem prezentowanych metod,
- s. 90 w wierszu 12: „przy przyjęciu ilościowej [...] próby",
- w pracy używane jest słowo obciążenie zarówno jako „obciążenie estymatora" jak i „obciążenie sprawozdawcze" (sugeruję używać zawsze precyzyjnego dwuwyrazowego określenia).

Pojawiają się też **problemy z tłumaczeniami** terminów statystycznych. Zdarza się, że są one częściowe lub dosłowne, a Doktorant nie korzysta z tych, które można znaleźć w polskiej literaturze. Jako przykłady można wymienić:

- „wygładzanie jądra",
- „model superpopulacji",
- „selekcja",
- „metoda PPS" lub „schemat PPS", lub „losowanie PPS",
- w pracy różnie tłumaczone jest angielskie słowo „semiparametric",
- „współczynnik zmienności estymatora",
- „SRSWOR",
- ponadto na s. 8 napisano „statystyki oficjalnej, w Polsce określanej jako statystyka publiczna", a następnie w pracy wielokrotnie korzysta z określenia statystyka oficjalna.

Doktorantowi nie udało się ustrzec też przed **błędami ortograficznymi i interpunkcyjnymi**, w tym:

- na s. 43 w wierszu 3 licząc od dołu stronu (przedrostek mikro zapisuje się łącznie z rzeczownikiem; ten sam błąd pojawia się też w innych miejscach pracy),
- s. 51 w wierszu 22 („nie" z imiesłowami przymiotnikowymi piszemy łącznie),
- s. 68 w wierszu 12 („by" z osobowymi formami czasownika piszemy łącznie).
- w wielu miejscach pracy brakuje przecinków, przykładowo na s. 15 w wierszu 11, na s. 20 w drugim wierszu podrozdziału 1.2, na s. 25 w wierszu 20, na s. 28 w wierszu 5,

- zbędne przecinki np. po inicjalnym składniku zdania na s. 29 w ostatnim wierszu (też w innych miejscach pracy), s. 38 w wierszu 9, na s. 20 przecinki przed „itd.",
- brak znaków interpunkcyjnych po wzorach (np. na s. 38),
- nieprawidłowe stosowanie kropki np. zbędna kropka w środku zdania na s. 10 w wierszu 21, brak kropki na końcu zdania na s. 17 w wierszu 4 licząc od dołu strony, kropka na początku wiersza 18 na s. 38, dwie kropki na s. 45 w wierszu 6,
- kropka kończąca zdanie powinna znajdować się po cudzysłowie (np. s. 29 w wierszu 3), licząc od dołu strony; ten sam błąd pojawia się też w innych miejscach pracy),
- w przypadku dzielenia wyrazów zapisywanych z łącznikiem, łącznik powinien zostać zdublowany – powinien być zapisany na stronie 38 na końcu wiersza 4 i na początku wiersza 5; też na s. 55).

Pojawiają się też **błędy i usterki o charakterze edycyjnym**, takie jak:

- duża liczba pojedynczych liter na końcu wiersza w całej pracy np. pięć na stronie 10,
- różne formatowanie akapitów (z i bez wcięcia akapitowego) np. na s. 4 i 10,
- po tytułach rozdziałów i podrozdziałów Doktorant stawia kropkę, a nawet dwie kropki (np. s. 45),
- nie podjęto próby standaryzacji wielkości czcionki używanej na rysunkach (np. bardzo mała czcionka na rysunku 1 na s. 18 i bardzo duża na rysunku 2 na s. 20),
- zbędny znak „|" na s. 19 w wierszu 5,
- brak spacji na s. 20 w wierszu 6 licząc od dołu strony,
- spacje przed znakami interpunkcyjnymi (np. na stronach 41 i 49),
- są dwa podrozdziały 4.3.2,
- s. 5, wiersze 13-15 (zdanie należy poprawić),
- s. 11, wiersze 20-21 (początek zdania należy poprawić),
- rysunek 3 na s. 22 mógłby zostać przedstawiony w opracowaniu własnym Doktoranta w języku polskim,
- sugeruję zmianę formy zapisu listy punktowanej, która zaczyna się na stronie 23, a kończy na stronie 29 (lub 30),
- s. 25 w wierszu 22: „do operatu mogą dostać dopisane",
- s. 29 w wierszu 10: „chęć dostarczenia jak więcej informacji [...]",
- s. 32 w wierszu 20: „lub/i" (wystarczy „lub"),
- s. 37 w wierszu 9: „literatura czyni takie podział",
- tytuł podrozdziału „2.3.2 Łączenie probabilistyczne" na s. 39 w wierszu 17 nie jest odpowiednio sformatowany i wyodrębniony z tekstu (nie ma go też w spisie treści),
- strona 41, zapis wierszy 10-12,
- pierwsze zdanie na s. 42 jest wyjaśnieniem oznaczeń użytych we wzorach prezentowanych na s. 41 – powinno być kontynuacją zdania ze s. 41 i poprzedzone słowem „gdzie",
- w tabeli 3 kolorem szarym zaznaczono braki danych, a w tabeli 4 kolorem szarym oznaczone dostępne dane (wymagane ujednolicenie),
- różna wielość czcionki w przypisach (s. 27 i s. 48),

- s. 48 w wierszach 16-17: „obie metody mają dobre właściwości oceny",
- s. 52 w wierszach 2-3: „Autorzy przywołanej wyżej pracy podkreślają w swoim materiale" lepiej zastąpić przez „Chen i in. (2022) podkreślają",
- s. 53 w wierszu 4: „Yang i in. (2021) jako jaką z jedną metod [...]",
- s. 53 w wierszu 11: jest cudzysłów górny, ale nie ma dolnego,
- s. 54 drugie zdanie pierwszego akapitu wymaga korekty (w tym czasownik powinien być w liczbie mnogiej a słowo „odpowiednio" bezpośrednio po czasowniku),
- s. 60 w wierszu 7: „akutalizowana",
- s. 60, wiersze 15-17 (należy poprawić zdanie),
- na s. 63 w wierszu 21 jest „masowe", a powinno być „masowej",
- ostatni wzór prezentowany na stronie 64: symbol „*" nie jest znakiem mnożenia,
- s. 64: tytuł podrozdziału 4.3.2 jest w ostatnim wierszu,
- s. 65 pierwsze zdanie podrozdziału 4.3.2 należy poprawić,
- s. 73 tytuł tabeli 10 jest prezentowany na dole strony a cała tabela na stronie kolejnej,
- s. 75 w wierszu 4: „nanajbliższych",
- s. 78 w wierszu 7: „wzięty pod w przypadku",
- dolna połowa strony 78 jest pusta,
- na s. 79 (i kolejnych) Doktorant prezentuje jeden wykres na danej stronie i zostawia pozostałą część strony pustą,
- s. 85 w wierszu 8: „Jednym z nich z względy jakościowe".

Wśród innych usterek dotyczących strony technicznej i językowej pracy można wymienić:

- powtórzenia w 2 i 3 zdaniu wprowadzenia i na s. 15 w wierszach 24 i 25,
- s. 27: „tym niemniej" jest rusycyzmem (lepiej „niemniej jednak"),
- fragment zdania na s. 40 w wierszach 3 i 4 licząc od dołu strony wymaga korekty,
- s. 42 w wierszu 2 od dołu strony: „porównania [...] jest",
- s. 47 w wierszach 19-20: „w tym imputację [...] oparta na [...]",
- s. 56 w wierszu 6 od dołu strony: „przedstawionych we wprowadzeniu celu",
- s. 60 w wierszu 6: pleonazm „w miesiącu styczniu" (podobnie w innych miejscach pracy).

Praca wymaga gruntownej, fachowej korekty edycyjnej.

4. Konkluzja

Choć problem badawczy rozważany w rozprawie jest bardzo interesujący i ważny z punktu widzenia prowadzenia badań reprezentacyjnych o charakterze ekonomicznym, przedstawione powyżej krytyczne uwagi uniemożliwiają zaakceptowanie pracy w jej obecnym kształcie. W szczególności brak precyzji opisu uniemożliwia pełną ocenę oryginalności rozwiązania problemu badawczego. Oprócz koniecznego przeredagowania pracy warto byłoby w tym zakresie również uwzględnić na końcu każdego rozdziału dodatkowy, krótki podrozdział podsumowujący dany rozdział ze szczególnym uwzględnieniem bardzo precyzyjnego opisu własnego wkładu Doktoranta. Pozwoliłoby to na dokładną ocenę oryginalności rozwiązania problemu naukowego przez Autora. Pominięcie problemu obciążenia, brak rozróżnienia

pomiędzy oszacowaniem precyzji estymatora a precyzją estymatora jak również pomiędzy precyzją estymacji a dokładnością estymacji mogą podawać w wątpliwość spełnienie wymaganego ustawowo warunku stawianego rozprawom doktorskim dotyczącego wykazania ogólnej wiedzy teoretycznej Doktoranta w rozważanym zakresie. Ponadto błędy popełnione przy projektowaniu i realizacji badania symulacyjnego nie pozwalają na stwierdzenie, że dysertacja w jej obecnej formie świadczy o umiejętności samodzielnego prowadzenia badań przez Doktoranta. Niezbędne jest też usunięcie licznych błędów warsztatowych i korekta typograficzna.

Można stwierdzić, że rozprawa ma potencjał, aby stać się pracą naukową w postępowaniu awansowym, ale po jej uzupełnieniu i poprawieniu. Po naniesieniu zmian proszę o przesłanie do recenzji pracy w wersji elektronicznej z czytelnie zaznaczonymi fragmentami, które zostały poprawione lub dodane.

Wnioskuję o uzupełnienie i poprawienie przez mgra Pawła Lańducha rozprawy pt. „Wykorzystanie technik imputacyjnych w szacowaniu informacji wynikowych oraz w analizie struktury danych w statystyce przedsiębiorstw".

T. Zgdło