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

Katowice, 29 lipca 2024

Recenzja poprawionej 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

Podtrzymuję moją ocenę tematyki pracy jako aktualnej, bardzo ciekawej i wpisującej się w aktualne trendy rozwoju metody reprezentacyjnej obejmujące wspomaganie badań próbkowych innymi źródłami informacji. Podjęty problem badawczy bez wątpienia może stanowić podstawę przygotowania rozprawy doktorskiej w dziedzinie nauk ekonomicznych, w dyscyplinie ekonomia. W szczególności warto podkreślić rosnące znaczenie tej tematyki zarówno w zakresie teoretycznym, jak i praktycznym, w tym w prezentowanym w pracy zagadnieniu badań przedsiębiorstw prowadzonych w ramach statystyki publicznej.

2. Ocena zmian merytorycznych w poprawionej wersji rozprawy

Jakość pracy po poprawie jest lepsza zarówno w zakresie merytorycznym, jak i edycyjnym, co jednak nie oznacza, że jest ona pozbawiona usterek i błędów.

Kluczowy błąd w poprzedniej wersji rozprawy polegający na analizie precyzji zamiast dokładności rozważanych estymatorów został poprawiony. Doktorant uwzględnia już obciążenie estymacji w prowadzonych analizach. Cel pracy również został poprawiony, choć użyte w nim (i w innych miejscach rozprawy) sformułowanie „estymatory imputacji masowej" powinno być zastąpione przez precyzyjne określenie np. „estymatory wartości globalnej w podpopulacjach wykorzystujące imputację masową" (gdyż to nie imputacja masowa jest szacowana, ale wartość globalna w podpopulacjach). Jednak mimo zawartej w poprzedniej recenzji uwagi dotyczącej konieczności doprecyzowania hipotezy

badawczej i hipotezy dodatkowej, nie wprowadzono żadnych istotnych zmian w tym zakresie. Podobnie stało się z uwagą, że treść hipotez prezentowana we wstępie jest kopiowana w dalszej części pracy zamiast wprowadzenia odpowiednich odwołań.

W zakresie rozdziału pierwszego i drugiego rozprawy poprzednia recenzja zawierała głównie uwagi o charakterze technicznym i edytorskim, które zostaną omówione w kolejnej części niniejszego opracowania. Rozdział trzeci istotnie poszerzono zgodnie z sugestią uwzględnienia bardziej szczegółowych studiów literaturowych, jednak kwestia autorstwa założeń 1 i 2 (na co zwracano uwagę w recenzji poprzedniej wersji rozprawy) prezentowanych obecnie na stronie 61 moim zdaniem nadal nie jest jasna. W poprzedniej recenzji zawarta była uwaga dotycząca konieczności uwzględnienia bardziej szczegółowej prezentacji rozważanych metod estymacji (m.in. uwzględnienie opisu własności oraz postaci estymatorów wariancji lub błędu średniokwadratowego). Choć uwaga ta została uwzględniona w przypadku estymatora Horvitza-Thompsona (HT) danego wzorem (9) i estymatora danego wzorem (15) (zob. opis poniżej wzoru (15)), lecz nie w przypadku wszystkich metod estymacji i predykcji prezentowanych w pracy. Zgodnie z sugestią przedstawiono postać estymatora HT wartości globalnej nie w populacji, ale w podpopulacji, gdyż to ten problem jest studiowany w pracy, a także postać jego wariancji i estymatora wariancji (zob. s. 53). Nie wprowadzono jednak tej prostej modyfikacji, prezentując estymator dany wzorem (12) na stronie 54 (nadal jest to estymator wartości globalnej w populacji a nie w podpopulacji). Podobnie ma to miejsce w przypadku innych estymatorów przedstawianych w pracy. Wprowadzono korekty w podrozdziale 3.4 tak, aby jego zawartość odpowiadała tytułowi oraz przeniesiono zgodnie z sugestią odpowiednią część poprzedniego rozdziału 4 do obecnego rozdziału 3, dodając podrozdział 3.5. Ponadto Autor wprowadził pewne modyfikacje w podrozdziałach prezentujących wykorzystywane metody imputacji (obecnie są to podrozdziały 3.5.1-3.5.3), co było sugerowane w poprzedniej recenzji. Mimo sugestii, aby pozostałą część poprzedniego rozdziału 4 połączyć z poprzednim rozdziałem 5, Doktorant zdecydował się nie wprowadzać tej korekty, co w mojej ocenie niekorzystnie wpływa na strukturę pracy.

W części teoretycznej uwzględniono sugerowane w poprzedniej recenzji: najlepszy liniowy nieobciążony predyktor oraz estymator kalibracyjny. Należy uznać to za istotną zmianę. Szkoda jedynie, że dodane w nowej wersji pracy opisy predyktora i estymatora kalibracyjnego nie są precyzyjne. Prezentując predyktor, nie wyjaśniono wszystkich oznaczeń użytych we wzorze (21) i nie przedstawiono, jak szacowana jest jego dokładność, na który to problem zwracano już uwagę w poprzedniej recenzji. Opisano, jak wyprowadzona jest postać estymatora kalibracyjnego, ale nie przedstawiono odpowiedniego wzoru. Ponadto w tym przypadku również nie podjęto problemu estymacji dokładności. Należy też dodać, że z kodu prezentowanego na stronie 149 wynika, że w celu uzyskania wartości predyktora są szacowane parametry modelu, gdzie zmienną objaśnianą jest badana zmienna po transformacji logarytmicznej, a następnie stosowana jest transformacja odwrotna. Stąd

wynika, że rozważany w badaniu symulacyjnym predyktor nie jest prezentowanym w pracy predyktorem należącym do klasy najlepszych liniowych nieobciążonych predyktorów, gdyż są to predyktory definiowane dla modeli liniowych.

W poprzedniej recenzji zwrócono uwagę, że w pracy pojawia się wzór (obecnie s. 74 wiersz 7 licząc od dołu strony) zawierający określenie „wartość z modelu", które miało być zapisane wzorem w nowej wersji pracy – tej zmiany jednak nie uwzględniono. Podobnie na stronie 74 nadal arbitralnie przyjmowane są wartości odchyleń standardowych składników losowych (1 oraz 0,5), podczas gdy można było je ustalić z wykorzystaniem danych rzeczywistych. Na stronach 77 i 78 napisano, że będą badane obciążenie i dokładność estymatorów MSE i przedstawiono odpowiednie mierniki dane wzorami (33) i (34). Następnie, w pewnej sprzeczności z powyższym, na s. 78 pojawia się argumentacja przeciwko zastosowaniu estymatorów wariancji, które mogłyby zostać wykorzystane w przypadku losowania systematycznego. Należy też podkreślić, że poza estymatorem wariancji (11), w pracy nie prezentowano innych wzorów opisujących estymatory wariancji lub estymatory MSE pozostałych prezentowanych w pracy estymatorów i predyktora. W szczególności Doktorant nie podjął się zapisu i analizy własności zaproponowanego w poprzedniej recenzji bardzo prostego estymatora błędu średniokwadratowego inspirowanego estymatorem MSE estymatorów syntetycznych. Stąd po lekturze rozdziału 4, Czytelnik nie ma pewności, czy ten problem będzie analizowany w dalszej części rozprawy, a jeśli tak (a jest analizowany w rozdziale 5) to na podstawie jakich wzorów.

Poprawiono opis badania symulacyjnego prezentowany obecnie w podrozdziale 4.2. Była to bardzo poważna usterka poprzedniej wersji rozprawy, teraz opis ten jest czytelny. Oprócz rozważanych wcześniej estymatorów uwzględniono inne klasyczne metody estymacji i predykcji, co należy uznać za znaczącą zaletę. Mocną stroną projektu badania symulacyjnego jest też uwzględnienie różnych wariantów dotyczących wielkości próby nielosowej, co świetnie komponuje się z aktualnymi rozważaniami prowadzonymi w literaturze w zakresie wspomagania estymacji danymi pochodzącymi z rejestrów.

Zgodnie z opisem zaprezentowanym w rozprawie na stronach 77-78, zaprojektowane badanie symulacyjne można podzielić na dwie części. Pierwsza to analiza porównawcza własności estymatorów wartości globalnej z wykorzystaniem mierników danych wzorami (31) i (32), a druga to analiza porównawcza własności estymatorów MSE z wykorzystaniem mierników danych wzorami (33) i (34). W ponownie przeprowadzonym badaniu symulacyjnym w większości przypadków prawidłowo wyznaczono wartości obciążeń i RRMSE estymatorów, co umożliwia analizę porównawczą własności wykorzystywanych metod. Wyjątkiem są wyniki dla obciążonego estymatora kalibracyjnego wartości globalnej przychodów ze sprzedaży. Uzyskane symulacyjne wartości obciążeń w zaprojektowanym badaniu wyniosły dokładnie 0, czyli nie były zgodne ze znanymi własnościami teoretycznymi. Ponadto wartości RRMSE tego estymatora też wyniosły 0, co oznacza, że dla każdej próby wylosowanej

w badaniu wartość estymatora była równa wartości parametru, który jest szacowany. Taka sytuacja nie jest możliwa w praktyce badań. Błąd wynika z faktu (s. 148, wiersz 2 licząc od dołu strony), że Doktorant w równaniu kalibracyjnym błędnie uwzględnia zmienną badaną zamiast dodatkowej, przyjmując, że znane są wartości globalne w warstwach zmiennej, które przecież szacuje (w tym przypadku przychodów ze sprzedaży). Gdy są szacowane wartości globalne wynagrodzenia, wyniki są już inne, ale zgodnie z opisem na s. 103 w wierszu drugim licząc od dołu strony, nadal zmienną uwzględnianą w kalibracji są przychody ze sprzedaży, które są traktowane przez Doktoranta jako jedna z dwóch badanych zmiennych, a nie jako zmienna dodatkowa. Poza tym wyniki należy uznać za ciekawe. Przykładowo okazało się, że w wielu przypadkach (strony 88-92) estymator HT, gdzie zakłada się, że wartości badanej zmiennej są znane w próbie losowej, charakteryzuje się tylko nieco wyższą dokładnością niż rozważane estymatory, które wykorzystują informacje o badanej zmiennej wyłącznie spoza próby losowej. Najsłabszą częścią nowej wersji rozprawy jest druga część badania symulacyjnego dotycząca analizy własności estymatorów błędów średniokwadratowych. Jak wspomniano powyżej, poza jednym wzorem (zob. wzór (11)) oraz opisem pod wzorem (15) Autor nie omawia w rozprawie problemu estymacji dokładności i precyzji, a rozważa ten problem w badaniu symulacyjnym. W aneksie, gdzie zaprezentowano skrypty wykorzystywane w badaniu symulacyjnym, też nie znalazłem odpowiednich funkcji pozwalających na ocenę dokładności estymacji. Biorąc dodatkowo pod uwagę skrypty prezentowane na stronach 156-161, które służyły do przeprowadzenia obliczeń, można stwierdzić, że wszystkie wyniki zaprezentowane w tabelach 20-29 oraz na w tabelach 40-49 są nieprawidłowe. Prezentowane wartości nie są wartościami mierników, które jak twierdzi Doktorant, zostały wyznaczone z wykorzystaniem wzorów (33) i (34).

W związku z powyższymi uwagami nasuwają się następujące pytania:

1. Dlaczego w badaniu symulacyjnym zdecydowano się na losowanie próby w warstwach z wykorzystaniem schematu losowania systematycznego z prawdopodobieństwami inkluzji proporcjonalnymi do cechy dodatkowej (a nie na przykład innego schematu losowania z prawdopodobieństwami inkluzji proporcjonalnymi do cechy dodatkowej) skoro Autor krytycznie ocenia możliwości oceny precyzji estymatora HT w tym przypadku (s. 78, wiersze 5-12)?
2. W pracy przedstawiono wyniki pokazujące, że w wielu przypadkach estymator HT, gdzie zakłada się, że wartości badanej zmiennej są znane w próbie losowej, charakteryzuje się tylko nieco wyższą dokładnością niż rozważane estymatory, które wykorzystują informacje o badanej zmiennej wyłącznie spoza próby losowej, co należy uznać za wielką zaletę tych metod. Od jakich czynników zależy wielkość utraty dokładności w rozważanym badaniu symulacyjnym?

3. Aby wartości oszacowań parametrów populacji lub podpopulacji można było uznać za wartościową i wiarygodną informację, niezbędny jest akceptowalny poziom szacowanej dokładności tych ocen. Stąd przedmiotem zainteresowania nie są wyłącznie metody estymacji parametrów populacji i podpopopulacji, ale też metody szacowania ich dokładności. W jaki sposób szacować dokładność estymatorów rozważanych w pracy w praktyce badań reprezentacyjnych?

3. Ocena zmian technicznych w poprawionej wersji rozprawy

Jakość edycyjna pracy po poprawie jest istotnie lepsza, ale nie wprowadzono wszystkich wymienionych w poprzedniej recenzji poprawek, także tych podanych bardzo szczegółowo. Przykładowo w poprzedniej recenzji prosiłem o podanie powołań na literaturę do następujących fragmentów (podałem tam numery stron i numery wierszy):

- znajdującego się w poprzedniej pracy na s. 25 w wierszach 14-15 „alternatywna metoda PPS bez zwracania ze stałą licznością próby" (w poprawionej wersji pracy s. 24, wiersze 24-25),
- znajdującego się w poprzedniej pracy na s. 29 w wierszu 2 (w poprawionej wersji pracy s. 28 wiersz 16)
- znajdującego się w poprzedniej pracy na s. 32 w wierszu 17 (w poprawionej wersji pracy s. 33, wiersz 16)

Żadna z powyższych trzech korekt nie została wprowadzona. W poprzedniej recenzji zwracałem też uwagę, że „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". W przypadku większości wzorów Doktorant nie wprowadził korekt zgodnie z tą sugestią. Jako przykład korekt, które nie zostały wprowadzone, mogą też służyć kropki, które nadal pojawiają się po tytułach rozdziałów i podrozdziałów. Mimo uwagi dotyczącej używania frazy „niestatystyczne źródło danych", termin „niestatystyczny" pojawia się jednak w pracy na s. 79 w wierszu 5 licząc od dołu strony. Także nie wszystkie błędy interpunkcyjne zostały poprawione. Znacząco poprawiono bibliografię, lecz nadal zdarzają się usterki np. na stronie 132 w publikacjach trzeciej i piątej (obie to artykuły w czasopismach) inaczej zapisano numery tomów i numery stron.

W nowych częściach rozprawy i poprawionych fragmentach też zdarzają się błędy techniczne. Przykładowo kilkukrotnie pojawia się fraza „wyprowadzać twierdzenie" zamiast „udowadniać twierdzenie". W objaśnieniach do wzoru (18) na s. 65, ϵ nazwano resztą zamiast składnikiem losowym. Zdarzają się usterki edycyjne, np. na s. 47 w czwartym wierszu licząc od dołu strony jest słowo „postawę" zamiast „podstawę", s. 56 wiersz 5 „ocena poprawnych tych założeń", s. 56 wiersz 9 „dodatkowe" zamiast „dodatkowo". Niektóre listy punktowane są niewłaściwie sformatowane np.

na s. 56, zdarzają się powtórzenia np. s. 58 wiersze 9-11, rozdziały 3 i 4 nie zawierają podsumowania, które wprowadzono w rozdziałach 1 i 2. Podsumowania nie zawierają też uwypuklonych autorskich propozycji, o co proszono w poprzedniej recenzji (ostatecznie są one jednak zaprezentowane w zakończeniu). W recenzji proszono, aby poprawić znajdujący się na s. 54 w pierwszym zdaniu pierwszego akapitu (teraz strona 60, drugi wiersz od dołu) termin „funkcja rozkładu" – poprawiono go na „warunkową gęstość gęstości". W aneksie w opisie skryptów błędnie podano numery wzorów – na stronie 152 w wierszu 2 powinno być powołanie na wzór (31) a nie (30), podobnie w każdym kolejnym. W nazwach skryptów na stronach 156 i 159 też są błędy – nie istnieją mierniki nazywane przez Autora „obciążeniem błędu średniokwadratowego" oraz „RMSE błędu średniokwadratowego".

4. Konkluzja

Podsumowując należy stwierdzić, że jakość rozprawy uległa poprawie zarówno pod względem merytoryczno-analitycznym jak i edytorskim. Jednak nadal są widoczne liczne usterki, niestety także w zakresie uwag szczegółowo opisanych w poprzedniej recenzji. Choć Doktorant uzyskał wartościowe i ciekawe wyniki porównań własności rozważanych estymatorów, to wyniki z zakresu badania własności estymatorów błędów średniokwadratowych rozważanych estymatorów są nieprawidłowe. To w mojej ocenie jest najistotniejszą wadą rozprawy w jej obecnej postaci. Jednak mimo to, biorąc pod uwagę niewielką, ale autorską modyfikację metody imputacji opartej na *k* najbliższych sąsiadach, zaprojektowane badanie symulacyjne oraz uzyskane wyniki z zakresu porównań własności rozważanych estymatorów (które mają kluczowe znaczenie), można ocenić rozprawę jako spełniającą wymogi ustawowe. Można uznać, że praca jest oryginalnym rozwiązaniem problemu badawczego, Doktorant wykazał się ogólną wiedzą teoretyczną w dyscyplinie naukowej, a dysertacja świadczy o umiejętności samodzielnego prowadzenia badań przez Doktoranta.

Wnioskuję o dopuszczenie mgra Pawła Łańducha, Autora rozprawy pt. „Wykorzystanie technik imputacyjnych w szacowaniu informacji wynikowych oraz w analizie struktury danych w statystyce przedsiębiorstw" do dalszych etapów przewodu doktorskiego.

T. Żądło