dr hab. inż. Kesra Nermend, Prof. US
Katedra Metod Komputerowych
w Ekonomii Eksperymentalnej
Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytet Szczeciński

Szczecin, 12.02. 2016

RECENZJA

ROZPRAWY DOKTORSKIEJ MGR. PRZEMYSŁAWA POSPIESZNEGO

PT. „ZASTOSOWANIE TECHNIK EKSPLORACJI DANYCH DO ESTYMACJI PRACOCHŁONNOŚCI I CZASU TRWANIA PROJEKTÓW INFORMATYCZNYCH"

Napisanej pod kierunkiem dr. hab. Andrzeja Kobylińskiego, prof. nadzw. SGH

Recenzja została opracowana na zlecenie Dziekana Kolegium Analiz Ekonomicznych Szkoły Głównej Handlowej w Warszawie z dnia 14.12.2015. Dotyczy rozprawy napisanej przez mgr. Przemysława Pospiesznego pod kierunkiem dr. hab. Andrzeja Kobylińskiego, prof. nadzw. SGH, w dziedzinie nauk ekonomicznych, w dyscyplinie ekonomia. Łączna objętość pracy wynosi 246 stron.

Recenzję napisano w celu oceny kwestii czy rozprawa spełnia warunki określone w art. 13 ustawy z dnia 14 marca 2003 r. *o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki.* (Dz. U. z 2003r. Nr 65, poz. 595; z późn. zm.). Recenzja obejmuje punkty:

1. Wybór tematu pracy i zdefiniowanie problemu badawczego.
2. Cel i hipotezy rozprawy.
3. Ocena merytoryczna i formalna pracy na tle jej układu.
4. Ocena źródeł wykorzystanych w pracy.
5. Wnioski końcowe w kontekście wymogów art. 13, ust.1 Ustawy.

1. Wybór tematu pracy i zdefiniowanie problemu badawczego

W mojej opinii temat podjęty w recenzji jest bardzo ważny i wciąż aktualny. Problem estymacji pracochłonności i czasu trwania projektów informatycznych jest jednym z głównych obszarów badawczych, jakim zajmuje się inżynieria oprogramowania. Tematyka badań z tego zakresu jest przedmiotem zainteresowania licznych publikacji

z obszaru szacowania pracochłonności i czasu trwania projektów informatyczny. Od właściwego szacowania parametrów (zakres, koszt, czas) każdego przedsięwzięcia, w istotnym stopniu zależy sukces lub porażka na każdym etapie jego realizacji. Przedsiębiorstwa realizujące przedsięwzięcia informatyczne, aby być konkurencyjnymi, winny dokonywać wyboru odpowiednich metod estymacji pracochłonności i czasu realizacji projektów informatycznych posługując się przy tym adekwatnymi narzędziami informatycznymi. Biorąc pod uwagę złożoność problemu estymacji pracochłonności i czasu trwania projektów informatycznych, każda próba poszukiwania skutecznych metod, technik czy narzędzi wspierających proces estymacji parametrów projektu może być oceniona pozytywnie. W tym kontekście stwierdzam, że podjęta w pracy tematyka dotyka zagadnień niezmiennie aktualnych i ważnych. Uważam również, że propozycja Doktoranta polegająca na budowaniu agregacyjnych modeli predykcyjnych z użyciem ogólnych modeli liniowych, sieci neuronowych oraz drzew decyzyjnych CHAID do estymacji pracochłonności i czasu trwania projektów informatycznych jest słuszna.

2. **Cel i hipotezy rozprawy**

Główny cel rozprawy został zdefiniowany jako ***„budowa agregacyjnych modeli predykcyjnych z użyciem ogólnych modeli liniowych, sieci neuronowych oraz drzew decyzyjnych CHAID do estymacji pracochłonności i czasu trwania projektów informatycznych”***.

W pracy postawiono również **cele poznawcze:**

- określenie zależności pomiędzy zmiennymi opisującymi projekty informatyczne oraz ich wpływu na szacowanie pracochłonności i czasu trwania inicjatyw,
- ocena przydatności ogólnych modeli liniowych, wielowarstwowych sieci neuronowych oraz drzew decyzyjnych CHAID do estymacji pracochłonności i czasu trwania projektów informatycznych,

cele metodyczne:

- opracowanie podejścia do budowy agregacyjnych predykcyjnych modeli eksploracji danych estymujących pracochłonność i czas trwania projektów informatycznych z użyciem trzech technik regresyjnych data mining: ogólnych modeli liniowych, wielowarstwowych sieci neuronowych oraz drzew decyzyjnych CHAID,
- zaproponowanie metodyki wdrożenia zbudowanych modeli w praktyce,

oraz cel aplikacyjny:

budowa agregacyjnego modelu estymującego pracochłonność i czas trwania inicjatyw z użyciem wielobranżowej bazy historycznych projektów informatycznych, celem wstępnej kalibracji algorytmów predykcyjnych oraz oceny ich możliwości aplikacji, w rezultacie potencjalnego wdrożenia modelu w ramach procesów zarzadzania inicjatywami w rożnego typu organizacjach realizujących projekty informatyczne.

W rozprawie zostały sformułowane trzy hipotezy badawcze o następującym brzmieniu:

- predykcyjne techniki eksploracji danych (data mining) mogą znajdować zastosowanie w zarządzaniu projektami informatycznymi, wspomagając proces estymacji pracochłonności i czasu trwania inicjatyw na ich inicjalnym etapie oraz potencjalnie przyczyniać się do wzrostu prawdopodobieństwa zakończenia projektu sukcesem. Przez to stanowią one narzędzie konkurencyjne do metod tradycyjnych oraz metod wykorzystujących linie kodu źródłowego lub punkty funkcyjne",
- ogólne modele linowe, wielowarstwowe sieci neuronowe oraz drzewa decyzyjne CHAID charakteryzują się dostatecznie dobrą zdolnością predykcyjną pracochłonności i czasu trwania projektów informatycznych oraz odpornością na braki i szumy w danych, umożliwiająca potencjalne ich wdrążenie w praktyce,
- agregacyjne predykcyjne modele eksploracji danych zastosowane do estymacji projektów informatycznych na początkowym etapie umożliwią otrzymywanie dokładniejszych szacunków badanych zjawisk niż użyte indywidualnie algorytmy.

Oceniając zarówno cele pracy, jak i hipotezy badawcze, w mojej opinii cele pracy zostały sformułowane w sposób poprawny, a zarazem adekwatny do postawionego problemu badawczego. Cele pracy wskazują na charakter aplikacyjny realizowanych badań, co może prowadzić do istotnych rozwiązań ważnych – jak wcześniej wykazano – problemów. Akceptuję również postawione hipotezy i stwierdzam, że są obiektywnie weryfikowalne.

3. Ocena merytoryczna i formalna pracy na tle jej układu

Praca składa się z czterech rozdziałów (171 stron), poprzedzonych wstępem (13 stron), podsumowania (6 strony) i zakończenia. Rozprawa uzupełniona została spisem

literatury (11 stron), który zawiera podstawowe pozycje z literatury przedmiotu (tj. książki i artykuły), w tym większość w języku angielskim. Ponadto w pracy zawarto wykaz stron organizacji zajmujących się zagadnieniami będącymi tematem rozprawy. Integralną częścią pracy są załączniki (35 stron) oraz właściwe tego typu pracom spisy rysunków, wykresów i tabel (4 stron). Dodatkowo dysertacja zawiera spis treści i streszczenie w języku angielskim. W tym kształcie stanowi ona spójną całość o logicznej strukturze, zgodną z tytułem choć niepozbawioną pewnych niedociągnięć. Przykładowo, Autor na str. 10 (w pierwszym akapicie) korzysta z publikacji z roku 2011, w oparciu o którą przedstawia procentowo liczbę projektów zakończonych sukcesem zgodnie z założonym harmonogramem i budżetem. Uważam, że wartość tej pracy byłaby w istotny sposób podwyższona, gdyby zostały przedstawione bardziej aktualne dane, np. z ostatnich dwóch lat. Dalej we wstępie Autor pisze, że techniki data miningu są popularne i mają zastosowanie jako narzędzie wsparcia decyzyjnego (str. 14, akapit drugi), a na następnej stronie zaprzecza temu stwierdzeniu (str. 15, akapit drugi).

W pierwszym rozdziale, dokonano szerokiego przeglądu anglojęzycznej literatury przedmiotu, a na jej tle przeprowadzono w miarę udaną próbę uporządkowania wiedzy dotyczącej tematyki metod estymacji parametrów projektów informatycznych, w tym głównie czynników i kryteriów wpływających na zakończenie projektów informatycznych sukcesem. W podrozdziałach 1.3 i 1.4, Doktorant omawia problemy związanie z estymacją parametrów projektów informatycznych, a także dokonuje przeglądu klasycznych technik estymacji takich jak: przez analogię, szacowanie eksperckie czy dekompozycję, jak i bardziej zaawansowanych metod opartych na liniach kodu źródłowego (COCOMO/ COCOMO II, SLIM, SEER-SEM) oraz na wymiarowaniu oprogramowania z użyciem punktów funkcyjnych (IFPUG, NESMA, COSMIC). Najbardziej interesującą częścią tegoż rozdziału jest omówienie zalet i wad obecnie stosowanych technik estymacji parametrów projektów informatycznych oraz wskazanie kierunku ich poprawy w celu podniesienia potencjału estymacyjnego.

W drugim rozdziale Doktorant na drodze analizy teoretycznej uzasadnia wybór technik eksploracji danych do odkrywania wiedzy w projektach informatycznych, a następnie dokonuje przeglądu literatury z zakresu zastosowania technik data mining do estymacji parametrów projektu, jako odpowiedniego narzędzia do rozwiązania przedstawionego problemu badawczego. Istotną częścią tego rozdziału jest rekomendacja Doktoranta trzech algorytmów predykcyjnych do budowy modeli estymujących

pracochłonność i czas trwania projektów oraz rozważania na temat ograniczeń dotychczasowych badań z obszaru zastosowań modeli predykcyjnych. Pod koniec tego rozdziału Doktorant uzasadnia potrzebę opracowania podejścia agregacyjnego z użyciem kilku technik predykcyjnych, które umożliwi wdrożenie modeli predykcyjnych data mining w praktyce, jako skutecznego narzędzia do estymacji pracochłonności i czasu trwania projektów informatycznych. W podrozdziale 2.1 (str. 71) Doktorant przedstawia obszary zarządzania wiedzą w projektach informatycznych, między innymi zarządzanie ryzykiem i zasobami ludzkimi. Jak wynika z literatury, te dwa obszary są bardzo ważne i odgrywają bardzo ważna rolę w zakończeniu projektu sukcesem. W związku z tym mam pytanie w jaki sposób Doktorant zminimalizuje ryzyko oraz konflikty w zespole projektowym w razie ich wystąpienia na etapie planowania, realizacji i monitorowania. W podrozdziale 2.2. (str. 81), Doktorant używa sformułowania „brak takich rozwiązań itp." W naukach społecznych należy być ostrożnym w formułowaniu takich twardych stwierdzeń. Doktorant w podrozdziale 2.3 (str. 104) przedstawia procentowy udział technik data miningu w estymacji prac związanych z budową systemów informatycznych, z czego wynika, że zaproponowane przez doktoranta metody są najczęściej używane. Oznacza to, że dobór metod wykonany przez Doktoranta jest trafny. W podrozdziale tym (str. 105) Doktorant uzasadnia również wybór bazy ISBSG jako najbardziej kompleksowej bazy projektów informatycznych w stosunku do innych. Jest to punkt wyjścia do realizacji przez Doktoranta celu poznawczego rozprawy.

Kolejny rozdział jest poświęcony badaniom empirycznym, których celem jest stwierdzenie poprawności proponowanego rozwiązania w oparciu o modele predykcyjne pracochłonności i czasu trwania projektów informatycznych. Ocena poprawności modeli została zrealizowana zgodnie z zaleceniami literatury przedmiotu w odniesieniu do modeli predykcyjnych. Do procesu uczenia i walidacji modeli Doktorant wykorzystuje wielobranżową bazę historycznych projektów informatycznych ISBSG, zawierającą ponad 6000 projektów, charakteryzującą się dużym wolumenem dobrych jakościowo danych. Doktorant jako zmienną wyszczególnił numer porządkowy projektu. W związku z tym mam pytanie do Doktoranta, co można wywnioskować o projekcie na podstawie jego numeru porządkowego? Istotnym aspektem omówionym w podrozdziale 3.1 jest proces przygotowania danych i doboru zmiennych diagnostycznych na potrzeby budowy modeli do estymacji pracochłonności i czasu trwania inicjatyw. Za plus sposobu przygotowania danych można uznać umiejętny dobór przez doktoranta właściwych metod

do eliminacji zmiennych nominalnych, gdyż jest ona bardzo ważnym etapem budowy modeli. Najbardziej istotnym elementem omawianego rozdziału, w mojej opinii, jest agregacja modeli składających się z trzech technik predykcyjnych: ogólnych modeli linowych, sieci neuronowych oraz drzew decyzyjnych CHAID, wraz z uzasadnieniem ich wyboru, co można uznać za oryginalny wkład Autora. Zaletą prezentowanego podejścia jest to, że stosunkowo łatwo można je zaimplementować w systemie informatycznym, który byłby, w mojej opinii, pomocnym narzędziem przy planowaniu i estymacji pracochłonności i czasu trwania projektów informatycznych. Ogólnie oceniam pozytywnie zarówno procedurę badania empirycznego, jak i dobór użytych w niej metod. Przeprowadzone badania pozwalają na osiągnięcie głównego celu pracy.

W rozdziale czwartym Doktorant dokonuje ewaluacji i oceny zaproponowanych modeli indywidualnych oraz agregacyjnych pod względem dokładności estymacji oraz możliwości ich wykorzystania w praktyce. Posługuje się najczęściej spotykanymi w literaturze miarami. Na stronach 186 i 188 w tabelach 33 i 34 Doktorant prezentuje zestawienia porównania oceny między modelami indywidualnymi, a agregatowymi pod względem czasu trwania i pracochłonności projektów. Błędy modelu agregacyjnego dla pracochłonności okazały się bardzo zbliżone lub niższe od błędów modelu liniowego (który okazał się najlepszym indywidualnym modelem), co uzasadnia jego stosowanie. Dla długości trwania projektu błędy modelu agregatowego są zbliżone do najlepszego indywidualnego modelu – neuronowego. Trudno jest rozstrzygnąć, który model jest lepszy. Biorąc jednak pod uwagę większą złożoność modelu agregatowego podważa to zasadność jego stosowania w tym konkretnym przypadku. Pomimo to przeprowadzone badania wykazały wyższość modelu agregatowego, ze względu na uniwersalność. Model agregatowy w analizowanych przypadkach zawsze pozwalał na uzyskanie wyników zbliżonych do modelu najlepszego. W rozdziale czwartym Doktorant zrealizował cel aplikacyjny rozprawy.

W podsumowaniu Doktorant podkreśla przydatność oraz uniwersalność proponowanych w pracy rozwiązań. Pisze o możliwości wykorzystania ich w organizacjach zajmujących się produkcją systemów informatycznych. Jednocześnie Doktorant przedstawia ograniczenia pracy, które mogą zostać wyeliminowane w wyniku podjęcia dalszych badań. Przedstawia również wnioski z przeprowadzonych badań oraz odniesienie do założonych celów i hipotez badawczych.

Z formalnego punktu widzenia praca jest poprawna, napisana dobrym językiem. Nie mam istotnych zastrzeżeń redakcyjnych do tekstu. Autor dowiódł umiejętności konstruowania tekstu naukowego. Dostrzegam wszak w rozprawie pewne błędy i uchybienia, które pragnę odnotować w celu ich wyeliminowania, gdyby Doktorant zdecydował się na jej opublikowanie. Oto one:

- w tekście znajduje się bardzo dużo błędów typograficznych polegających na pozostawieniu na końcu wiersza spójników (tzw. wiszących spójników),
- pozostawianie pustych miejsc przed elementami graficznymi w pracy (np. str. 70, 114, 126, 129, 121) – przecież po to są one numerowane, aby umieszczać je w pobliżu odwołania, niekoniecznie bezpośrednio po odwołaniu,
- w przypadku, kiedy tabela zajmuje więcej niż jedną stronę, nagłówek tabeli powinien być powtarzany na każdej nowej stronie (np. str. 107, 119, 120, 125, 129, 141, 145, 148, 149, 152, 164, 165, 166, 171, 176, 177, 182),
- struktura pracy byłaby bardziej przejrzysta, gdyby doktorant przy numeracji rysunków i tabel, w każdym rozdziale zaczynał numerację od początku, gdzie pierwsza cyfra oznaczałby numer rozdziału zaś druga cyfra numer rysunku,
- w tekście występują niejasności np. na str. 33 napisane jest: „ gdzie zakres jest nie jest w pełni znany " ?,
- Autor często niepotrzebnie stosuje duże odstępy miedzy numeracjami tekstów i kolejnymi akapitami np. str. 33, 34, 35, 40, 41, 42, 43, 44, 47, itp.,
- str. 89. Autor pisze: dla zbioru n składnik losowy jest opisany przez wartości losowe $Y_1,\ldots,Y_n$. Skoro n jest zbiorem to nie może być numerem ostatniego elementu,
- str. 89. Wartości losowe zostały oznaczone dużymi literami, a we wzorze (2.1) występują jako małe litery,
- wzory (2.1)-(2.4) nie mają powołań na literaturę,
- wzór (2.1) – brak wyjaśnienia co oznacza b, c i d,
- wzór (2.2) – brak wyjaśnienia co oznacza p,
- wzory (2.5)-(2.7) nie jest opisane co oznacza μ_i,
- wzór (2.7) jest zapisany w sposób niejasny. Ze sposobu zapisu można wnioskować, że logarytm jest liczony z $\mu_i\mu_i$. Z drugiej strony odstęp przed μ_i sugeruje, że wartość logarytmu jest mnożona przez μ_i. Aby uniknąć takich wątpliwości Autor nie

powinien stosować odstępów, ale nawiasy jak we wzorze (2.6), albo μ_i przenieść przed logarytm (jeżeli logarytm jest liczony z μ_i), albo napisać μ_i^2 (jeżeli logarytm jest liczony
z μ_i^2),

- rysunek 10. Na rysunku nie zaznaczono wag,
- wzory (2.8) i (2.9). Autor wcześniej pisze, że jest n sygnałów. Wzory jednak wskazują, że sygnałów jest n+1 i zmienia się od 0 do n,
- dlaczego $f(P_j)$ jest równe $f\left(\sum_{i=0}^{n} x_i w_{ij}\right)$ a nie $f\left(\sum_{i=0}^{n} x_i w_{ij} \mu_i\right)$ jak by to wynikało ze wzoru (2.8),
- pisząc wzory należy używać przeznaczonych do tego nawiasów, czyli tak $f\left(\sum_{i=0}^{n} x_i w_{ij}\right)$ a nie tak $f(\sum_{i=0}^{n} x_i w_{ij})$.

Pragnę podkreślić, że w mojej opinii wymienione uchybienia redakcyjne w żaden sposób nie zmniejszają wartości merytorycznej recenzowanej pracy.

4. Ocena źródeł wykorzystanych w pracy

Bibliografia obejmuje 124 źródeł, w tym głównie anglojęzycznych (źródła polskojęzyczne stanowią ok. 14 %). Tematyka wszystkich odnosi się do problematyki poruszanej w pracy. Wśród pozycji literatury przedmiotu ok. 33% stanowią publikacje z ostatnich 6 lat (2010-2015), co świadczy o tym, że Doktorant jest na bieżąco z najnowszymi doniesieniami w zakresie objętym pracą. Szkoda, że powołał się na tylko na dwie swoje publikacje, w tym jedną publikację, której jest współautorem - trudno bowiem uwierzyć, że nie ma ich więcej zważywszy na fakt, że badania o tak szerokim zakresie z pewnością trwały na tyle długo, że mogły się pojawić stosowne opracowania naukowe. Literatury polskojęzycznej jest stosunkowo mało do liczby ośrodków i naukowców zajmujących się zagadnieniami związanymi z obszarem zarządzania projektami informatycznymi oraz inżynierii oprogramowania. Mam tu na myśli choćby: ośrodek poznański z zespołem profesora J. Węglarza, wrocławski z zespołem Z. Huzara, czy szczeciński z zespołem profesora Z. Szyjewskiego. Mimo przedstawionych niedociągnięć, dobór i wykorzystanie źródeł w pracy oceniam dobrze.

5. Wnioski końcowe w kontekście wymogów art. 13, ust.1 Ustawy

Przedstawiona do recenzji rozprawa doktorska pt. „*Zastosowanie technik eksploracji danych do estymacji pracochłonności i czasu trwania projektów informatycznych*" przedstawia oryginalne i interesujące badania. Autor podjął w niej istotny naukowo problem. Przeprowadzając badania wykazał się:

- dobrą znajomością literatury z zakresu podjętego tematu, co może być podstawą w przyszłości do prowadzenia samodzielnych badań naukowych w obszarze ekonomii,
- dużą wiedzą z zakresu funkcjonowania przedsiębiorstwa z branży informatycznej, w szczególności pod kątem planowania procesów produkcyjnych systemów informatycznych,

ponadto:

- przedstawił oryginalną autorską koncepcję opracowania modeli predykcyjnych, jako alternatywne podejście do estymacji pracochłonności i czasu trwania projektów informatycznych w stosunku do tradycyjnych metod szacowania, opartych na wiedzy eksperckiej, liniach kodu źródłowego i punktach funkcyjnych,
- przedstawiona przez Autora koncepcja modeli może być pomocna w procesie estymacji parametrów projektów tak, aby zminimalizować ryzyko zaniechania realizacji rozpoczętych już projektów przez inwestora z powodu nie oczekiwanych korzyści biznesowych,
- wykorzystał możliwości jakie dają metody data miningu do opracowania zestawu modeli liniowych, neuronowych oraz drzew decyzyjnych do szacowania pracochłonności i czasu trwania inicjatyw,
- Autor przez agregację opracowanych modeli zapewnił uzyskanie dokładniejszych szacunków oraz możliwości przeciwdziałania wystąpienia nadmiernego dopasowania danego algorytmu do danych.

W mojej opinii wyczerpuje to wymagania art. 13 Ustawy, zatem stwierdzam, że recenzowana praca, może być podstawą do ubiegania się o nadanie stopnia naukowego doktora nauk ekonomicznych, w dyscyplinie naukowej ekonomii. W związku z powyższym wnioskuję o dopuszczenie mgr. Przemysława Pospiesznego do publicznej obrony rozprawy pt. „*Zastosowanie technik eksploracji danych do estymacji pracochłonności i czasu trwania projektów informatycznych*" napisanej pod kierunkiem dr. hab. Andrzeja Kobylińskiego, prof. nadzw. SGH.