PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 525 • 2018

ISSN 1899-3192
e-ISSN 2392-0041

Krzysztof Błoński, Ewa Putek-Szeląg
Uniwersytet Szczeciński
e-mails: krzysztof.blonski@usz.edu.pl; ewa.putek-szelag@usz.edu.pl

WYKORZYSTANIE METODY *PROPENSITY SCORE MATCHING* W BADANIACH TYPU *DESK RESEARCH*

THE USE OF THE PROPENSITY SCORE MATCHING METHOD IN DESK RESEARCH

DOI: 10.15611/pn.2018.525.14
JEL Classification: C13, C14

Streszczenie: Celem artykułu jest przedstawienie możliwości zastosowania *Propensity Score Matching* (PSM) do analizy danych zastanych w ramach badań typu *desk research*. W artykule przedstawiono koncepcję wykorzystania tej metody do łączenia dwóch zbliżonych baz danych zastanych. Artykuł ma charakter badawczy. Źródłem danych zastanych są wyniki Europejskiego Sondażu Socjologicznego oraz Europejskiego Badania Jakości Życia. W postępowaniu badawczym korzystano również z literatury z zakresu socjologii i statystyki, dostępnej w postaci książek i artykułów. Za pomocą metody PSM wybrano z dwóch wymienionych baz danych jednostki najbardziej podobne pod względem wybranych zmiennych. Celem takich działań będzie połączenie wymienionych baz danych, co według autorów ma zwiększyć możliwość przeprowadzania analiz. W artykule zastosowano również regresję logistyczną, na podstawie której policzone zostały prawdopodobieństwa.

Słowa kluczowe: dane zastane, *desk research*, statystyczny wpływ netto.

Summary: The aim of the article is to present the features of Propensity Score Matching (PSM) useful in analyzing the existing data in desk research. The article will present the concept of using this method to connect two similar databases. The article is of research nature. The source of existing data will be the results of the European Sociological Survey and European Quality of Life Survey. The research also included the use of literature in the field of sociology and statistics available in the form of books and articles. By means of the PSM method, the most similar, in terms of selected variables, units from the mentioned two databases will be selected. The purpose of such activities will be to combine the mentioned databases, which, according to the authors, is to increase the possibility of conducting analyses. In the article also logistic regression is used, based on which the probabilities were calculated.

Keywords: existing data, desk research, Propensity Score Matching.

1. Wstęp

Rola danych wtórnych (zastanych) w procesie badawczym jest jasno określona – rozpoczynają fazę gromadzenia danych. Analiza danych zastanych jest więc nieodzownym etapem badań i jednocześnie punktem wyjścia do dalszych działań [Mynarski 2000, s. 11-15]. W literaturze przedmiotu pojawiają się sugestie, że analiza danych zastanych pozostaje niedocenianym elementem realizacji badań [Bednarowska 2015, s. 18]. Za tą tezą przemawiają m.in. wnioski dotyczące częstości ich stosowania w badaniach społecznych [Smith 2008, s. 173] czy stosowane przez praktyków nazewnictwo – „zakurzona technika badawcza" [PTBRiO 2013, s. 26].

Dane zastane w wąskim rozumieniu to każdego rodzaju materiały, których cel powstawania nigdy nie był celem badawczym, ale korzystają z nich różni badacze. Natomiast dane zastane w szerokim rozumieniu to każdego rodzaju materiały niewywoływane przez badacza, który na nich pracuje niezależnie od tego, czy wcześniej były wywoływane w celach badawczych [Makowska (red.) 2013]. Szczegółowy opis źródeł, podziału danych zastanych, jak i ich wady i zalety prezentuje m.in. *Analiza danych zastanych* [Makowska (red.) 2013, s. 16 i n.).

Analizę danych zastanych wykorzystuje się przede wszystkim w ramach badań typu *desk research*, analizy treści czy wtórnej analizy statystycznej. Oprócz wymienionych zastosowań można wskazać również badania historyczno-porównawcze, analizę dyskursu, *case study*, analizę obrazu czy badania ewaluacyjne [Makowska (red.) 2013].

Desk research to badanie polegające na poszukiwaniu danych jakościowych lub ilościowych, które mogą być przydatne z punktu widzenia poruszanego tematu [Makowska (red.) 2013, s. 18]. Do najczęściej wymienianych zalet badania tego typu zalicza się: kwestię dostępności danych, koszty wykonania analiz [Hofferth 2005, s. 893], możliwości analiz na większych próbach (przy założeniu ich dostępności) [Frankfort-Nachmias, Nachmias 2001, s. 321] oraz brak wpływu badacza na przedmiot badania [Babbie 2003, s. 341]. Można do tej listy dodać również argument, iż *desk research* pozwala na szersze porównanie różnych wyników badania dotyczących tego samego lub podobnego obszaru badawczego, dzięki czemu istnieje możliwość wzbogacenia dotychczasowych mechanizmów wnioskowania [Bednarowska 2015].

Bazując na tej przesłance, można wskazać, że celem artykułu jest przedstawienie możliwości zastosowania *Propensity Score Matching* (statystyczny wpływ netto – PSM) do analizy danych zastanych w ramach badań typu *desk research*. Źródłem danych zastanych będą wyniki Europejskiego Sondażu Socjologicznego (European Social Survey – ESS) oraz Europejskiego Badania Jakości Życia (European Quality of Life Survey – EQLS).

2. Metodyka badań ESS i EQLS oraz zakres przeprowadzonych analiz

Europejski Sondaż Społeczny to akademicki projekt badań społecznych realizowanych w Europie od 2001 roku. Celem badania ESS jest obserwacja zmian społecznych zachodzących w Europie, tj. postaw wobec kluczowych problemów, zmian w systemach wartości i zachowań. Część pytań wykorzystywanych w poszczególnych rundach sondażu jest taka sama, część pytań (moduły rotacyjne) jest zmienna. W analizie wykorzystano odpowiedzi na pytania pochodzące zarówno z części stałej, jak i rotacyjnej. Dotychczas przeprowadzono osiem rund badań – ostatnią w 2016 roku. Do realizacji celu artykułu wykorzystano wyniki badań z 2012 roku (runda 6). Przeprowadzono wtedy blisko 55 tysięcy wywiadów w 29 krajach. W Polsce badaniami zostało objętych 1898 osób.

European Quality of Life Surveys (EQLS) to badania przeprowadzane okresowo co 4 lata (od 2003 roku; ostatnie badanie przeprowadzono w 2016 roku[1]) w oparciu o wystandaryzowany kwestionariusz wywiadu. Celem badań EQLS jest uchwycenie poglądów, postaw i doświadczeń osób dorosłych zamieszkałych w Europie w celu dokonania oceny ich jakości życia. Dokonując operacjonalizacji celu badania, można wskazać dalsze zagadnienia będące obszarem zainteresowania. Są to m.in.: zatrudnienie, dochody, wykształcenie, warunki mieszkalne, rodzina, zdrowie, równowaga między pracą a życiem osobistym. Dodatkowo badania obejmują również pomiar subiektywnej oceny poziomu zadowolenia osobistego, a także postrzeganą jakość życia społeczeństwa. W przypadku tych danych zastanych do realizacji celu badania wykorzystano wyniki przedostatniego badania z roku 2012. Wyniki są reprezentatywne dla każdego kraju uczestniczącego w badaniu. Badanie w Polsce przeprowadzono na próbie 2262 respondentów.

Statystyczny wpływ netto to metoda zaproponowana przez P. Rosenbauma i D. Rubina [1983, s. 41-55]. Umożliwia ona redukcję obciążenia selekcyjnego przy szacowaniu przeciętnego efektu oddziaływania na jednostki poddane interwencji (*Average Treatment Effect on Treated* – ATT). Polega ona na dopasowywaniu do grupy poddanej interwencji takiej grupy kontrolnej, wyselekcjonowanej z puli kontrolnej osób niepoddanych oddziaływaniu, że rozkłady charakterystyk wektora w obu grupach będą zbalansowane X [Denkowska 2015, s. 62-63]. Metoda PSM jest jedną z zalecanych przez Komisję Europejską metod przeprowadzania ewaluacji projektów i programów współfinansowanych ze środków unijnych [European Commission 2014, s. 6-7, za: Denkowska 2016, s. 67]. W badaniach ekonomicznych była stosowana m.in. do oceny efektywność aktywnych programów rynku pracy [Wiśniewski, Zawadzki (red.), 2011, s. 139-150; Śliwicki 2014, s. 27-40], natomiast w badaniach

[1] W momencie przekazywania artykułu do wydawnictwa (przełom roku 2017 i 2018) nie były dostępne wyniki badań z ostatniej rundy. Według zapowiedzi mają one być udostępnione w marcu 2018 roku.

edukacyjnych zastosowano ją do szacowania wpływu niskiego poziomu wykształcenia na wykluczenie w wyróżnionych obszarach życia [Panek, Zwierzchowski 2015, s. 22-26]. W przypadku badań społecznych można ją zastosować m.in. do oszacowania efektu wpływu statusu społeczno-zawodowego oraz wykluczenia społecznego na dostępność i gotowość współpracy jednostek badania w badaniu terenowym [Rószkiewicz 2017] czy do łączenia baz danych zastanych. W tym artykule zostanie przedstawiona koncepcja wykorzystania tej metody do łączenia dwóch zbliżonych baz danych zastanych.

Algorytm postępowania w ramach metody PSM zawiera pięć etapów. Są to [Trzciński 2009, s. 31 i n.; Denkowska 2015, s. 63]:

1. Szacowanie wartości *propensity score.*

2. Wybór metody dopasowania grupy respondentów jednej bazy danych do grupy respondentów z drugiej bazy danych na podstawie oszacowanych wartości *propensity score* (w literaturze przedmiotu wskazywane są cztery potencjalne metody: najbliższego sąsiada, z limitem, z promieniem, Kernela).

3. Sprawdzenie wspólnego obszaru określoności – jest to jeden z etapów oceny dopasowania grupy kontrolnej. Na tym etapie możliwe jest stosowanie różnych zaleceń, np. wizualne porównanie rozkładów *propensity score* w różnych grupach, wyznaczenie wartości maksymalnych oraz minimalnych w obu grupach czy też zastosowanie metody *trimming*.

4. Ocena zbalansowanych zmiennych – drugi z etapów oceny dopasowania grupy kontrolnej, w ramach którego możliwe jest stosowanie zarówno numerycznych, jak i graficznych metod diagnostycznych. Do podstawowych numerycznych metod oceny zbalansowania zalicza się metodę badania standaryzowanych różnic średnich autorstwa Rosenbauma i Rubina [1983]. Jeżeli zbalansowania zmiennych nie można uznać za satysfakcjonujące, to wówczas należy zastosować inne metody dopasowywania grupy kontrolnej (oznacza to cofnięcie się do etapu drugiego) bądź ewentualnie powrócić do etapu estymacji modelu regresji logistycznej (powrót do etapu pierwszego), wprowadzając do modelu interakcje oraz zmienne ilościowe podniesione do kwadratu.

5. Estymacja efektu przyczynowego – etap ten jest realizowany w momencie otrzymania satysfakcjonującego zbalansowania wszystkich zmiennych, interakcji oraz zmiennych w wyższych potęgach uwzględnionych w modelu.

Technika ta jest sposobem na redukcję ilości cech/wymiarów, za pomocą których możemy opisać obserwacje w zbiorze danych. Wymiary te zostają sprowadzone do jednego syntetycznego wskaźnika, definiowanego czasem jako skłonność do partycypacji w warunku interwencji [Haber (red.) 2007, s. 187]. Prowadzi to do uzyskania wektora balansującego obie grupy. Każdy uzyskany wynik w dwóch grupach możemy interpretować jako pewnego rodzaju poziom podobieństwa względnego.

3. Zastosowanie PSM w badaniach typu *desk research*

W pierwszym etapie wybrane zostały z dwóch baz takie same zmienne opisujące zmienną niezależną:

1. wiek,
2. płeć (mężczyzna – 0; kobieta – 1),
3. wykształcenie (podstawowe i zawodowe, średnie, wyższe),
4. miejsce zamieszkania (wieś, małe miasto, duże miasto),
5. zadowolenie (skala pomiaru o rozpiętości 1-10, gdzie 1 – bardzo niezadowolony, zaś 10 – bardzo zadowolony).

Za zadowolone osoby uznane zostały te, które zaznaczyły 8 lub więcej punktów w zmiennej zadowolenie, a za niezadowolone te, które zaznaczyły do 3 punktów. Na skutek tych działań w bazie EQSL zostały 1242 ankietowane osoby, a w bazie ESS – 1128 osób.

W następnym etapie wykorzystano regresję logistyczną, która określona jest równaniem:

$$P(Y = 1|x_1, x_2, \dots, x_k) = P(X) = \frac{e^{a_o + \sum_{i=1}^{k} a_i x_i}}{1 + e^{a_o + \sum_{i=1}^{k} a_i x_i}},$$

gdzie: Y oznacza zmienną dychotomiczną przyjmującą wartości: 1 – najczęściej dla zdarzeń pożądanych, jak np. przeżycie, sukces; 0 – w przeciwnym przypadku, np. zgon lub porażka; a_i, $i = 0,\dots, k$ są współczynnikami regresji, $x_1, x_2,\dots, x_k$ to zmienne niezależne (ilościowe lub jakościowe).

Dzięki regresji logistycznej uzyskano przewidywane prawdopodobieństwo przynależenia obserwacji do danej grupy. Technika PSM, w tym artykule, została wykorzystana do utworzenia pierwszej grupy (na podstawie bazy EQSL), składającej się z jednostek w jak największym stopniu podobnych do tych, które znalazły się w drugiej grupie (ESS). Dopasowanie jednostek odbywa się w oparciu o wartość tylko jednej zmiennej – *propensity score*.

W przypadku zmiennych z bazy EQSL istotnymi zmiennymi okazały się wiek i wykształcenie, natomiast w przypadku zmiennych ESS – wiek, wykształcenie oraz miejsce zamieszkania. Mimo że zmienna miejsce zamieszkania nie jest istotna w przypadku wyjaśniania poziomu satysfakcji z danych zawartych w bazie EQSL, to przy szacowaniu modelu została ona uwzględniona, aby z takich samych zmiennych niezależnych zbudowane zostały modele. Wyniki zastosowania regresji logistycznej przedstawiają tabele 1 i 2.

W kolejnym kroku na podstawie zbudowanych modeli regresji zostały policzone prawdopodobieństwa zdarzeń. Metoda *Propensity Score Matching* posłużyła do odpowiedniego dopasowania do przypadków z grupy ESS przypadku lub przypadków z grupy EQSL. Do grupowania wykorzystano metodę najbliższego sąsiedz-

Tabela 1. Wyniki estymacji modelu logitowego z zero-jedynkowymi zmiennymi objaśniającymi bez uwzględnienia efektów interakcji dla zmiennych z bazy EQLS

Zmienna/poziom zmiennej	Ocena parametru β	Błąd standardowy	Chi-kwadrat Walda	p	Iloraz szans (e^{β})
Wyraz wolny	**2,694**	**0,568**	**22,456**	**0,000**	**14,785**
Małe miasto	–0,050	0,459	0,012	0,913	0,951
Duże miasto	–0,274	0,472	0,337	0,561	0,760
Wykształcenie średnie	0,416	0,222	3,523	0,061	1,517
Wykształcenie wyższe	**1,609**	**0,394**	**16,695**	**0,000**	**4,999**
Wiek	**–0,021**	0,005	15,513	0,000	0,979

Źródło: opracowanie własne.

Tabela 2. Wyniki estymacji modelu logitowego z zero-jedynkowymi zmiennymi objaśniającymi bez uwzględnienia efektów interakcji dla zmiennych z bazy ESS

Zmienna/poziom zmiennej	Ocena parametru β	Błąd standardowy	Chi-kwadrat Walda	p	Iloraz szans (e^{β})
Wyraz wolny	2,700	0,298	82,003	0,000	14,875
Małe miasto	**-0,473**	**0,224**	**4,446**	**0,035**	**0,623**
Duże miasto	**-0,979**	**0,230**	**18,170**	**0,000**	**0,376**
Wykształcenie średnie	0,203	0,199	1,036	0,309	1,225
Wykształcenie wyższe	**1,021**	**0,285**	**12,859**	**0,000**	**2,777**
Wiek	**-0,013**	**0,005**	**8,015**	**0,005**	**0,987**

Źródło: opracowanie własne.

twa. Z przyporządkowanych par obserwacji wybrano te, dla których osobie z grupy pierwszej przyporządkowana została osoba z grupy drugiej. W wyniku przeprowadzonego badania zostało wyłonionych 58 par.

W celu sprawdzenia, czy zastosowana metoda dała pozytywne wyniki, w pierwszej kolejności zostały przedstawione struktury grup.

Grupy pod względem poziomu zadowolenia i płci są podobne, natomiast pod względem miejsca zamieszkania i wykształcenia sytuacja przedstawia się inaczej. W grupie EQSL najwięcej osób mieszkało w małym mieście, a w grupie ESS na wsi. Z kolei w pierwszej grupie najwięcej zostało wybranych osób z wykształce-

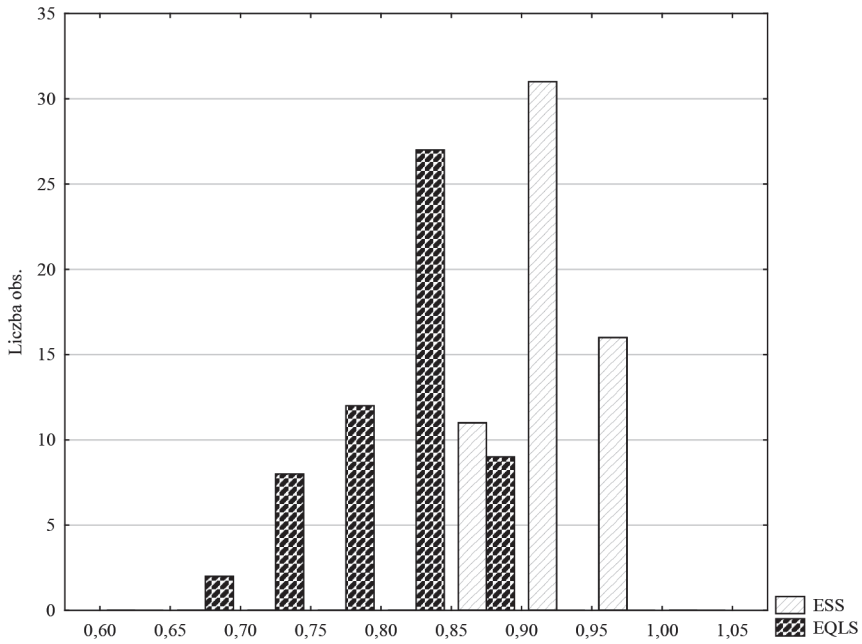
niem średnim, a w drugiej prawie po tyle samo z podstawowym i zawodowym oraz średnim.

Tabela 3. Klasyfikacja wybranych do grupy osób ze względu na wybrane cechy

Grupa	Poziom zadowolenia		Płeć		Miejsce zamieszkania			Wykształcenie		
	0 (niezadowolenie)	1 (zadowolenie)	mężczyzna	kobieta	wieś	małe miasto	duże miasto	podstawowe i zawodowe	średnie	wyższe
Grupa EQSL	8	50	24	34	2	39	17	6	39	13
Grupa ESS	6	52	31	27	23	19	16	23	21	14

Źródło: opracowanie własne.

Dodatkowo autorzy przedstawili rozkłady prawdopodobieństw w badanych grupach (rys. 1).



Rys. 1. Rozkłady wartości *propensity score* dla 58 par z podziałem na grupy EQSL i ESS

Źródło: opracowanie własne.

Niestety, rozkłady prawdopodobieństw w badanych grupach istotnie się różnią. W celu sprawdzenia zastosowano test istotności różnic grup zależnych, gdyż w opi-

sywanym przypadku autorom zależało na znalezieniu jednostek jak najbardziej podobnych ($t = 15{,}73811$; $p = 0{,}000$).

4. Zakończenie

Współcześnie, ze względu na postępującą otwartość nauki, a więc i wzrastający dostęp do wyników badań, można prowadzić analizy, korzystając z coraz większego zasobu danych wtórnych. Jednak często zdarzają się sytuacje, że w kilku bazach znajdują się interesujące autora informacje. Autorzy artykułu za pomocą metody PSM chcieli wybrać jednostki, najbardziej podobne pod względem wybranych zmiennych, z dwóch wymienionych baz danych, aby w następnej kolejności móc połączyć dwie bazy.

Niestety, wyniki uzyskane w tym przypadku metodą PSM nie są zadowalające, gdyż:

- pierwotne bazy składały się z ponad tysiąca obserwacji, a zbiór podobnych obserwacji zawierał niecałe 60 obserwacji;
- dopasowanie wartości *property score* do wybranych obserwacji w dwóch badanych grupach było różne;
- dopasowanie grup pod względem wybranych zmiennych również nie było podobne.

Według autorów otrzymany wynik nie przesądza o przydatności lub nie wykorzystanej metody. Należałoby przeprowadzić analizę z wykorzystaniem innych zmiennych lub innych baz danych.

Literatura

Babbie E., 2003, *Badania społeczne w praktyce*, Wydawnictwo Naukowe PWN, Warszawa.

Bednarowska Z., 2015, *Desk research – wykorzystanie potencjału danych zastanych w prowadzeniu badań marketingowych i społecznych*, Marketing i Rynek, nr 5.

Denkowska S., 2015, *Wybrane metody oceny jakości dopasowania w Propensity Score Matching*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 384.

Denkowska S., 2016, *Zastosowanie analizy wrażliwości do oceny wpływu nieobserwowanej zmiennej w Propensity Score Matching*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 427.

European Commission, 2014, *The Programming Period 2014–2020. Guidance Document on Monitoring and Evaluation. European Regional Development Fund and Cohesion Fund. Concepts and Recommendations*, http://ec.europa.eu/regional_policy/sources/docoffic/2014/working/wd_2014_en.pdf (15.12.2017).

Frankfort-Nachmias Ch., Nachmias D., 2001, *Metody badawcze w naukach społecznych*, Wydawnictwo Zysk i S-ka, Poznań.

Haber A. (red.), 2007, *Ewaluacja ex-post. Teoria i praktyka badawcza*, Agencja Rozwoju Przedsiębiorczości, Warszawa.

Hofferth S.L., 2005, *Secondary data analysis in family research*, Journal of Marriage and Family, vol. 67, no. 4.

Makowska M. (red.), 2013, *Analiza danych zastanych*. *Przewodnik dla studentów*, Wydawnictwo Naukowe Scholar, Warszawa.

Mynarski S., 2000, *Praktyczne metody analizy danych rynkowych i marketingowych*, Wyd. Zakamycze, Kraków.

Panek T., Zwierzchowski J., 2015, *Opis metodologii badawczej współzależności pomiędzy wykluczeniem społecznym a edukacją*, Instytut Badań Edukacyjnych, Warszawa, http://eduentuzjasci.pl/publikacje-ee-lista/analizy/1157-opis-metodologii-badawczej-wspolzaleznosci-pomiedzy-wykluczeniem-spolecznym-a-edukacja.html (3.01.2018).

PTBRiO, 2013, *Badania marketingowe*, Rocznik Polskiego Towarzystwa Badaczy Rynku i Opinii, XVII.

Rosenbaum P.R., Rubin D.B., 1983, *The central role of propensity score in observational studies for casual effects*, Biometrika, vol. 70, no. 1.

Rószkiewicz M., 2017, *Ocena efektu wpływu statusu społeczno-zawodowego oraz wykluczenia społecznego na wskaźnik odpowiedzi wśród polskich gospodarstw domowych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia, nr 30-31.

Smith E., 2008, *Using Secondary Data in Educational and Social Research*, McGraw-Hill Education, Maidenhead.

Śliwicki D., 2014, *Zastosowanie estymatorów jądrowych do szacowania efektywności aktywnych programów rynku pracy*, Acta Universitatis Nicolai Copernici. Ekonomia XLV, nr 1. DOI: http://dx.doi.org/10.12775/AUNC_ECON.2014.002.

Trzciński R., 2009, *Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa.

Wiśniewski Z., Zawadzki K. (red.), 2011, *Efektywność polityki rynku pracy w Polsce*, Wojewódzki Urząd Pracy, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.