

Mikołaj Jasiński
Marek Bożykowski
Tomasz Zajac
Marek Styczeń
Albert Izdebski
Uniwersytet Warszawski

DOKŁADNIEJ, RZETELNIEJ, TANIEJ. BADANIA OPARTE NA REJESTRACH PUBLICZNYCH JAKO SZANSA DLA BADAŃ SPOŁECZNYCH W POLSCE

*Polska socjologia empiryczna bada głównie świadomość ludzi,
a lekceważy pozaświadomościową stronę życia społecznego.*

(Sulek 2002: 121)

Badania społeczne dotyczące różnorodnych zjawisk (między innymi z zakresu procesów edukacyjnych, rynku pracy, demografii, migracji, epidemiologii) mogą być prowadzone na podstawie informacji z rejestrów administracyjnych. Znajomość potencjału oraz ograniczeń tego podejścia stwarza możliwość pogodzenia pozornie sprzecznych postulatów – prowadzenia tanich, a jednocześnie efektywnych badań złożonych zjawisk społecznych. Korzystanie z informacji z rejestrów może pomóc rozwiązać szereg metodologicznych problemów, takich jak malejący poziom realizacji sondażowych badań reprezentacyjnych, nierzetelność zbieranych informacji itp.

Celem artykułu jest prezentacja specyfiki metodologii badań statystycznych wykorzystujących informacje z rejestrów. Jest to jeden z efektów projektu badawczego zrealizowanego przez Autorów. Rozważania metodologiczne uzupełniają przykładowe wyniki badań ścieżek edukacyjnych studentów oraz historii zatrudnienia absolwentów Uniwersytetu Warszawskiego z wykorzystaniem danych z rejestrów uczelni i Zakładu Ubezpieczeń Społecznych, przeprowadzonych w ramach projektu. Autorzy wyjaśniają również problemy technologiczne, administracyjne oraz organizacyjne, których przewyciężenie jest warunkiem prowadzenia badań wykorzystujących informacje z rejestrów. Wnioski metodologiczne przedstawione w artykule rozwijają propozycje badaczy skandynawskich najbardziej doświadczonych w dziedzinie badań wykorzystujących informacje z rejestrów administracyjnych.

Główne pojęcia: metodologia badań statystycznych; badania losów absolwentów; rynek pracy; rejestr administracyjny; Zakład Ubezpieczeń Społecznych.

M. Jasiński: Instytut Socjologii, e-mail: mikoj@is.uw.edu.pl; M. Bożykowski: Instytut Socjologii, e-mail: m.bozykowski@is.uw.edu.pl; T. Zajac: Instytut Socjologii, e-mail: zajact@is.uw.edu.pl; M. Styczeń: Instytut Socjologii, e-mail: stymar@is.uw.edu.pl; A. Izdebski: Instytut Socjologii, e-mail: izdebskia@is.uw.edu.pl

Wprowadzenie

Jednym z zasadniczych celów badań społecznych jest, obok analiz opinii, badanie faktów, zdarzeń składających się na historie obiektów tworzących różne populacje, np. ruchliwości zawodowej, historii edukacyjnych. Rejestry administracyjne są znaną formą zapisu faktów. Towarzyszyły organizowaniu się społeczności i tworzeniu instytucji. Dzięki temu stanowią dogodne źródło prowadzenia badań aktywności społecznej. Przykładowo, księgi parafialne czy portowe księgi cła palowego są wykorzystywane w badaniach historycznych. Rejestry są cennym źródłem także dla socjologa.

Socjologia ma długą tradycję badań bazujących na danych urzędowych. Do najbardziej znanych należą klasyczne studia Emile'a Durkheima z 1897 roku nad społecznymi uwarunkowaniami samobójstw (Durkheim 2006) oraz słynne badania Paula Lazarsfelda nad psychospołecznymi skutkami bezrobocia prowadzone w Marienthalu (Jahoda i in. 1974), w których wykorzystał między innymi dane z miejscowej biblioteki¹.

Częstym nieporozumieniem jest utożsamianie badań opartych na rejestrach z analizą statystyk publicznych, która skądinąd może być źródłem ciekawych wniosków, jak np. w badaniach Durkheima. Należy jednak pamiętać, że statystyki publiczne są przedstawiane w formie zagregowanej, co naraża badacza na popełnienie tzw. błędu ekologicznego, tj. próby wyciągania wniosków o jednostkach na niższym poziomie zagregowania (np. osobach) na podstawie prawidłowości zaobserwowanych na poziomie agregatu (np. powiatu). Popełnienie tego typu błędu zarzuca się właśnie Durkheimowi (zob. m.in. Freedman 2001: 4027–4030). Analizie zależności samobójstw od wyznania na zbiorze osób – badania faktycznie opartego na rejestrach, nie zaś statystykach publicznych – nie można by postawić takiego zarzutu, jak studiom Durkheima.

W dobie rozwoju elektronicznych rejestrów administracyjnych zarządzanych przez sprawne systemy informatyczne można by oczekiwać rozkwitu badań zjawisk społecznych odwołujących się do tych źródeł. Tymczasem w Polsce dominującą od lat metodą badawczą niezmiennie pozostają badania ankietowe. Nie pozostaje to bez wpływu na pole zainteresowań socjologa-empiryka, który koncentruje się przede wszystkim na badaniu świadomości ludzi (por. Sułek 2002: 123). Lekceważy przy tym badanie faktów albo bada je przy użyciu metod sondażowych, narażając się na szereg trudnych do rozwiązania problemów, między innymi niedokładność odpowiedzi (w tym skłonność do zaokrągleń) oraz kłopoty z odtwarzaniem faktów z przeszłości.

¹ Interesujące przykłady badań bazujących na danych urzędowych znajdzie Czytelnik między innymi w Sułek 2002.

Problemem, który od dawna szczególnie znacząco wpływa na jakość wyników badań sondażowych nie tylko w Polsce (zob. np. Goyder i McKenzie Leiper 1985), jest drastycznie obniżony poziom realizacji badań. Przykładowo, w ciągu ostatnich 20 lat poziom realizacji Polskiego Generalnego Sondażu Społecznego obniżył się z przeszło 80% w 1992 roku do zaledwie 45% w 2010 roku². Już niewielkie odsetki odmów udziału w badaniu lub nieobecności respondentów podczas wizyt ankieterów mogą znacząco skrzywić wyniki ze względu na nielosowy charakter tych zjawisk. Opisane w literaturze przedmiotu (Lissowski 1971; Jasiński 2004) poprawne metodologicznie sposoby odciążenia wyników są kosztowne i trudne w realizacji.

Istnieją metody badań wolne od opisanych powyżej wad. Ich podstawą są informacje zbierane na potrzeby administracji. Przykładowo, można oprzeć analizy na statystykach publicznych. Jak już wspomnieliśmy, wadą tego rozwiązania jest to, że wykorzystuje uprzednio zagregowane wyniki, przez co nie istnieje możliwość badania zależności na poziomie indywidualnym. Można także analizować dokumenty urzędowe. Te zawierają wprawdzie dane na temat indywidualów, jednak oparcie badań na pismach urzędowych naraża wielu problemów. Przede wszystkim dokumenty musiałyby w pierw zostać zakodowane w formie umożliwiającej analizy – należałoby je sprowadzić do określonego zestawu zmiennych. Proces taki byłby jednak bardzo pracochłonny.

Zasadniczą zaletą badań bazujących na informacjach pobranych z rejestrów administracyjnych jest to, że znacznie redukują koszty zbierania danych – dane już zostały zebrane przez odpowiednie instytucje dla celów administracyjnych, niezależnie od prowadzenia badań.

Metodologia badań opartych na danych z rejestrów administracyjnych szczególnie dynamicznie rozwija się w krajach skandynawskich. Funkcjonują tam sprawdzone rozwiązania prawne, organizacyjne i metodologiczne, dzięki którym dane administracyjne, po odpowiednim przetworzeniu, służą celom badawczym. Dostępnych jest szereg publikacji tamtejszych badaczy (Korpi 1989; Knudsen 1993; Harala 1995; Longva i in. 1998; Wallgren i Wallgren 2007) i skandynawskich urzędów statystycznych prezentujących doświadczenia i uwagi metodologiczne pomocne przy ustalaniu procedur badawczych. W niniejszym artykule przedstawiamy najważniejsze aspekty tego podejścia na podstawie doświadczeń Pracowni Ewaluacji Jakości Kształcenia na Uniwersytecie Warszawskim, które dobrze korespondują z wnioskami ze wspomnianych publikacji (Jasiński i in. 2012; Zajac 2011, 2012; Bożykowski i in. 2014). PEJK UW jest pierwszą akademicką jednostką badawczą w Polsce,

² Por. <http://pgss.iss.uw.edu.pl/index.php?show=wprowadzenie.html>

która rozwija systematyczne badania procesów kształcenia na bazie rejestrów administracyjnych.

Wykorzystanie rejestrów administracyjnych jako źródła danych do badań statystycznych powinno łączyć użyteczność analiz z bezpieczeństwem informacji. Sprostanie temu postulatowi wymagało opracowania odpowiednich procedur metodologicznych i organizacyjnych. Niezbędne było jasne odróżnienie istotnych cech rejestru od klasycznego zbioru danych statystycznych, rozpoznanie mocnych i słabych stron wniosku bazującego na rejestrach i wreszcie prześledzenie logiki pracy badawczej z danymi pochodzącymi z rejestrów. Po zaprezentowaniu metodologicznych aspektów badań statystycznych opartych na rejestrach administracyjnych przedstawimy przykładowe wyniki uzyskane w ramach prac PEJK UW.

Definicja rejestru

Rejestr ma być z założenia kompletną listą obiektów określonego rodzaju (np. ludzi, budynków, książek, zdarzeń), która pozwala na jednoznaczny identyfikację każdego zawartego w nim obiektu (por. Dz.U. 1995 Nr 88 poz. 439: art. 2 ust. 12 i art. 13 ust. 1; UNECE 2007: 15; Wallgren i Wallgren 2007: 4). Wspomniana identyfikowalność obiektów zawartych w rejestrze jest związana z inną, definicyjną cechą rejestru – aktualizowaniem informacji w nim zawartych wraz ze zmianami atrybutów tych obiektów.

Rejestry są tworzone przez administracje różnego rodzaju instytucji (urzędy, przedsiębiorstwa, szkoły, organizacje społeczne) w związku ze świadczonymi przez nie usługami czy powstającymi tam dobrami, a nie na potrzeby badań statystycznych (por. Starr 1983: 30).

Ta cecha jest szczególnie warta podkreślenia ze względu na to, że konstytuuje pojęcia rejestrów administracyjnych oraz danych statystycznych. **Administracja korzysta z rejestrów przy podejmowaniu decyzji mających konsekwencje dla konkretnych obiektów zawartych w danym rejestrze, natomiast badacze korzystają z baz danych opartych na informacjach z rejestrów do analiz, których wyniki nie mają jednostkowych konsekwencji** (por. Wallgren i Wallgren 2007: 11). Zrozumienie tej różnicy jest kluczem do opracowania właściwych (tj. respektujących prywatność) procedur udostępniania informacji pochodzących z rejestrów do badań statystycznych.

Zalety i wady rejestrów administracyjnych jako źródła danych do badań statystycznych

Rejestry administracyjne, jak wspomnieliśmy w części wprowadzającej, są niezwykle bogatym źródłem danych dla badaczy społecznych. Podczas badań opartych na informacjach pochodzących z rejestrów administracyjnych należy mieć świadomość słabych i mocnych stron tego podejścia (por. Hodgins 1998: 31; Smith i in. 2004: 8; UNECE 2007: 45–46; Wallgren i Wallgren 2007: 63–65).

Do głównych zalet tego typu badań należy możliwość prowadzenia obliczeń dla całej populacji, podczas gdy badania reprezentacyjne prowadzi się z reguły na wylosowanej próbie elementów populacji. W przypadku badań wyczerpujących możliwe jest porównywanie wąskich podpopulacji, np. sędziów, absolwentów określonego kierunku studiów z wybranego rocznika, osób na urloпах macierzyńskich. Badania reprezentacyjne zwykle na to nie pozwalają. Bardzo mały odsetek obiektów należących do danej podpopulacji przekłada się na niewielką liczebność takich obiektów w wylosowanej próbie. Badania wyczerpujące pozwalają ponadto na poznanie prawdziwych i dokładnych wartości wybranych parametrów w badanej populacji. Ewentualne zniekształcenia związane są z niedoskonałościami rejestrów. Niedoskonałości te opisujemy w dalszej części tekstu. Istnieje możliwość przynajmniej częściowego rozpoznania wielkości i kierunku tych zniekształceń, a w rezultacie uwzględnienia ich podczas analiz. W przypadku badań sondażowych niepewność uzyskanych wyników jest wpisana w samą logikę badania. Odbieganie uzyskanych oszacowań od wartości parametrów populacyjnych bierze się między innymi z losowości doboru próby oraz zniekształceń wynikających z trudności dotarcia do badanych.

Korzystając z rejestrów łatwiej prowadzić analizę dynamiki badanych zjawisk. W wielu rejestrach przy poszczególnych zapisanych zdarzeniach widnieje także ich data. Pozwala to na zbadanie zmian intensywności określonego zjawiska w czasie, z częstotliwością zależną od sposobu kodowania czasu zdarzeń w rejestrach. Częstotliwość ta jest większa niż przyjęto oczekiwać w analizach bazujących na danych sondażowych. Co więcej, dokładność zapisu danych w rejestrach jest zasadniczo niezmienna w czasie, a w przypadku badań sondażowych precyzja odtwarzania zdarzeń zależy od tego, ile czasu od nich minęło. Oczywiście, granice możliwości analiz wyznaczane są przez okres objęty rejestrem. Z jednej strony oznacza to, że niemożliwe jest prowadzenie w ten sposób badań dotyczących okresu, w którym rejestr nie funkcjonował, jednak z drugiej – pozwala na prowadzenie analiz retrospektywnych. Analizy takie mogą obejmować okresy znacznie wcześniejsze niż moment sformułowania pytań badawczych.

W odróżnieniu od badań sondażowych, badania oparte na informacjach pochodzących z rejestrów nie borykają się z zasady z problemem odmów udziału w badaniu (z wyjątkiem np. osób uchylających się od obowiązku rejestracji). Obecność danych w rejestrach administracyjnych jest albo w pewnych populacjach obowiązkowa (np. PESEL dla obywateli polskich), albo zarejestrowanie w rejestrze definiuje daną populację. Przykładami drugiej sytuacji są rejestry studentów uczelni (osoba, która chce studiować, musi się najpierw zarejestrować w rejestrze swojej uczelni), rejestry zawodowe (np. Centralny Rejestr Lekarzy, Krajowy Rejestr Adwokatów i Aplikantów Adwokackich) czy rejestry różnych organizacji. Poza tym, w przypadku rejestrów nie spotyka się odmów odpowiedzi spowodowanych drażliwością pytań, niechęcią badanych do udzielania odpowiedzi, niedokładnością ich pamięci czy znudzeniem związanym z długością kwestionariusza. Podanie kompletnych i prawdziwych informacji jest w przypadku rejestrów z zasady w interesie wypełniającego – albo jest podstawą do uzyskania przez niego jakichś pożytków (np. uzyskanie zwrotu podatków), albo jest wymuszone przez przepisy grożące karą w wypadku wprowadzenia fałszywych danych lub zatajenia informacji (np. wskutek niezarejestrowania umowy o pracę w ZUS). Braki danych w rejestrach mają inną naturę niż w badaniach sondażowych. W przypadku rejestrów są one nie tyle konsekwencją interakcji z ankierem, ile przewidywanych skutków udzielenia danej odpowiedzi, np. dzięki pominięciu w zeznaniu podatkowym niektórych transakcji (a tym samym zaniżeniu wielkości obrotu) przedsiębiorca może uniknąć rejestrowania się jako płatnik podatku VAT.

Informacje gromadzone w rejestrach są z reguły dokładniejsze niż te zbierane metodą sondażową. Rejestry są wynikiem realizacji rozmaitych obowiązków urzędowych (wypełnianie formularzy podatkowych, ubezpieczeniowych, rejestracja pojazdów itp.). Z zasady informacje te przekazywane są według wystandaryzowanych, rzadko zmieniających się reguł. Kontrola poprawności wypełnionych formularzy rejestrowych prowadzona przez instytucje zarządzające rejestrami dodatkowo zmniejsza ryzyko występowania błędów. Znika także problem związany z tendencją badanych do zaokrągleń, np. w przypadku pytań o zarobki.

Sformalizowanie relacji między udzielającym informacji a instytucją zarządzającą rejestrem redukuje problem „efektu ankietera”, znaczący w przypadku wywiadów osobistych i analizy bezpośredniej. Sprzyja temu unifikacja urzędników i procedur administracyjnych określana przez przepisy ustaw i regulaminów.

Nie należy także zapominać o tym, że koszty badań opartych na rejestrach są wyraźnie mniejsze niż w przypadku badań ankietowych. Wynika to w dużej mierze z tego, że informacje są zbierane na potrzeby określonych instytucji niezależnie od badacza, zatem nie musi on ponosić kosztów związanych z konstruowaniem narzędzia pomiarowego, losowania próby, zbierania

informacji w terenie itp. Przykładem badania, w którym podjęto tematykę obecnie poddającą się analizie na podstawie informacji czerpanych z rejestrów, jest panelowe badanie historii życia Polaków w okresie transformacji ustrojowej przeprowadzone przez Zespół Badania Ruchliwości Społecznej Instytutu Socjologii Uniwersytetu Warszawskiego.

W latach 1987, 1991 i 1998 przeprowadzono trzy pomiary losów życiowych losowej próby 3000 mężczyzn i 2000 kobiet urodzonych w latach 1948–1957, spisując relacje o stanie aktualnym i zapamiętanej przeszłości. Badanie dotyczyło między innymi historii kształcenia się, historii pracy zawodowej i historii rodziny. Pozwoliło na sformułowanie niezwykle interesujących wniosków charakteryzujących dynamikę zarobków i zatrudnienia w czasie kluczowych dla naszego kraju przemian społeczno-gospodarczych przełomu lat osiemdziesiątych i dziewięćdziesiątych (Pohoski i Styczeń 2000). Niestety, ze względu na koszty do ostatniego pomiaru wylosowano tylko próbkę liczącą 1336 mężczyzn, z których skutecznie przebadano 85%. Dalszych pomiarów panelu nie prowadzono.

Znaczna część problematyki podjętej we wspomnianym badaniu mogłaby być obecnie analizowana na bazie informacji pochodzących z rejestrów administracyjnych. Co więcej – badanie takie byłoby znacznie tańsze, oparte na większej liczbie pomiarów i mogłoby objąć większą zbiorowość (nawet populację), a wskaźniki byłyby wyznaczane na podstawie źródeł dokładniejszych niż ludzka pamięć.

Warto dodać, że ograniczenie źródeł informacji do rejestrów pozwala na zapewnienie badanym pełnej anonimowości. Osobisty, telefoniczny czy pisemny kontakt badanego z badaczem czy ankierem nie jest konieczny. Dodatkowo istnieje technologiczna możliwość prowadzenia analiz statystycznych bez udostępniania danych o pojedynczych jednostkach obserwacji.

Oczywistą wadą badań opartych na informacjach pochodzących z rejestrów jest ograniczenie źródeł do istniejących rejestrów. Nie można zatem zbadać w ten sposób tego, co nie jest odnotowywane w żadnym rejestrze. Przykładem są opinie, o których (przynajmniej w państwach demokratycznych) administracja nie gromadzi informacji w rejestrach. Aby poznać opinie badanych, trzeba posłużyć się innymi metodami.

Należy pamiętać, że rejestry są konstruowane tak, aby były użyteczne przede wszystkim dla korzystających z nich instytucji, nie zaś do celów badawczych. Stąd, jeśli jakaś informacja nie jest niezbędna dla funkcjonowania instytucji i podejmowania przez nią decyzji, to prawdopodobnie nie będzie przez nią zbierana lub będzie zbierana w sposób niestaranny³. Przykłado-

³ Osobnym, ciekawym problemem badawczym są uwarunkowania zmian przyjmowanych w rejestrach definicji, dopuszczalnych kategorii odpowiedzi oraz zakresu zbieranych informacji. Tematyka ta nie będzie jednak podejmowana w niniejszym artykule.

wo, rejestry Zakładu Ubezpieczeń Społecznych nie obejmują umów o dzieło, ponieważ nie są od nich odprowadzane składki do ZUS. Podobnie, informacje o związkach nieformalnych nie są rejestrowane w urzędach stanu cywilnego.

Istotnym źródłem błędów, zarówno w badaniach sondażowych, jak i w przypadku badań bazujących na rejestrach administracyjnych, jest problem niekompletności informacji. Przyczyny niekompletności danych z badań sondażowych (odmowy odpowiedzi na pytania, ucieczki w kategorię „trudno powiedzieć”, odpowiedzi niewiarygodne, błędy kodowania itp.) są odmienne niż w przypadku badań opartych na rejestrach. Niekiedy ze względu na niedoskonałości rejestrów informacje w nich zgromadzone nie pokrywają się ze stanem faktycznym – brakuje w nich danych o niektórych obiektach lub zawierają informacje o obiektach, które nie posiadają cech kwalifikujących je do figurowania w danym rejestrze. Różnice te biorą się z trzech rodzajów źródeł.

Po pierwsze, część obiektów nigdy nie zostaje zarejestrowana pomimo istniejącego obowiązku administracyjnego. Przykładem są nieruchomości zbudowane bez potrzebnych zezwoleń, niezgłoszone do ewidencji gruntów i budynków. Podobnie, w żadnym rejestrze nie figuruje informacja o zatrudnianiu pracowników „na czarno”. Poza tym możliwa jest sytuacja, gdy obiekt figuruje w rejestrze wbrew przepisom, np. osoby figurujące w rejestrach na podstawie sfałszowanych dokumentów.

Po drugie, informacje o obiektach mogą wpływać do rejestru z opóźnieniem. Wynika to na ogół z naturalnych problemów komunikacyjnych. Informacja o liczbie ludności w danym momencie z zasady odbiega nieco od stanu faktycznego.

Trzecia sytuacja dotyczy rejestrów, do których nie ma obowiązku się zgłaszać, np. rejestr bezrobotnych w Polsce nie zawiera informacji o wszystkich osobach spełniających ustawowe warunki uznania za osobę bezrobotną. W tej sytuacji badacz ma jednak świadomość charakteru braków danych. Może tę wiedzę uwzględnić przy prowadzonych analizach i interpretacjach.

Osobnym, wartym podkreślenia problemem jest możliwość występowania błędów pomiaru zarówno w rejestrach, jak i w danych zbieranych metodą sondażową. W przypadku rejestrów przyczyną występowania nieprawdziwych informacji mogą być na przykład oszustwa podatkowe, a w przypadku badań sondażowych – niezgodne z prawdą odpowiedzi badanych (kłamstwa, zaokrąglenia, zafałszowania spowodowane kłopotami z pamięcią itp.).

Przedstawione powyżej braki danych i zniekształcenia informacji w rejestrach są też znaczącym źródłem obciążeń w przypadku badań sondażowych. Należy bowiem pamiętać, że operaty używane do losowania prób w badaniach sondażowych są z zasady rejestrami. W związku z tym wady rejestrów przekładają się na skrzywienia wyników badań sondażowych.

Logika badań opartych na rejestrach

Wbrew temu, co piszą Anders i Britt Wallgren (2007: 63), nie istnieje odrębna metodologia badań statystycznych bazujących na rejestrach. Fakt istnienia zgromadzonych już danych w rejestrach administracyjnych nie oznacza, że postępowanie badawcze oparte na informacjach z rejestrów miałyby nie wpisywać się w kanon klasycznego podejścia do badań statystycznych, stosowanego np. w badaniach sondażowych. Nadal konieczne jest poprzedzenie analizy sformułowaniem pytań badawczych i konceptualizacją oraz operacjonalizacją pojęć.

W obu przypadkach pierwszym etapem jest postawienie pytań badawczych. Postulat ten, oczywisty w przypadku sondażu, może budzić wątpliwości przy badaniach bazujących na rejestrach zawierających już przecież zebrane wcześniej informacje. Sformułowanie hipotez i pytań badawczych powinno być w tym przypadku poprzedzone jedynie ogólnym rozpoznaniem zakresu informacji zbieranych przez instytucje (np. struktura formularzy ZUS, zeznań podatkowych, protokołów egzaminacyjnych uczelni itp.). Poznanie ogólnej charakterystyki struktury dostępnych rejestrów administracyjnych umożliwi wybór konkretnych rejestrów jako źródeł danych do przyszłych analiz.

Podczas badań opartych na rejestrach, na etapie konceptualizacji, konieczne jest uwzględnienie ewentualnych odmienności definicji pojęć przyjętych na użytek badania od definicji przyjętych w rejestrach. Na przykład, urzędowa definicja osoby bezrobotnej nie tylko może odbiegać od definicji przyjętej przez badacza, ale także może zmieniać się w czasie w zależności od zmian polityk państwa i zmian przepisów prawnych. Poza tym różne instytucje mogą jednocześnie stosować odmienne definicje tego samego terminu. Jasne określenie wspomnianych różnic pozwoli na ustalenie zawartości informacyjnej oraz ograniczeń zmiennych wytworzonych w dalszych etapach na bazie rejestrów.

W badaniach opartych na analizie informacji z rejestrów, podobnie jak w klasycznych badaniach statystycznych, podstawowe jest ustalenie badanej populacji i możliwości dotarcia do niej. Oczywistym ograniczeniem wykorzystania informacji z rejestrów administracyjnych jako źródła danych do badań jest fakt, że nie dotyczą one wszystkich sfer życia. Bywa, że jeśli nie istnieje rejestr będący listą obiektów tworzących interesującą nas populację, to warunkiem możliwości utworzenia bazy danych, w której jednostkami obserwacji są elementy badanej zbiorowości statystycznej, jest istnienie zmiennej identyfikującej te obiekty w rejestrze obiektów innego typu. Na przykład, jeśli przedmiotem badania jest populacja powiatów, a istnieje wyłącznie rejestr osób zawierający informację o ich powiecie zamieszkania, to możliwe jest utworzenie zbioru danych o powiatach. Ponadto badane zjawiska można

analizować na różnych poziomach agregacji oraz z różnych perspektyw. Przykładowo, możemy rozpatrywać zjawisko bezrobocia w populacji osób, gospodarstw domowych, powiatów lub miesięcy. Wskazanie populacji konkretyzuje hipotezy i pytania badawcze – jeśli na ich podstawie nie da się wskazać populacji, to są one postawione zbyt ogólnie, by móc na nie sensownie odpowiedzieć.

Poza znacznym obniżeniem kosztów badań dużych i zróżnicowanych zbiorowości, które w przypadku badań sondażowych byłyby niezwykle kosztowne, z różnych względów, np. anonimizacji danych, możliwe jest prowadzenie badań na próbach losowych (w takiej sytuacji rejestr spełnia jednocześnie funkcję operatu).

Niezbędnym elementem konceptualizacji jest sprawdzenie kompletności danych. Ustalenie zakresu, skali oraz źródeł braków danych pozwoli na podjęcie działań, które umożliwią redukcję wykrytych obciążeń. Najlepszym, choć nie zawsze możliwym rozwiązaniem, jest uzupełnienie braków danych z innych źródeł. Wykorzystać można w tym celu zarówno dane z innych rejestrów, dopełniające brakujące informacje, jak i odwołać się do logicznych związków informacji zawartych w pojedynczym rejestrze (np. informację o fakcie uzyskania absolutorium przez studenta można uzyskać albo bezpośrednio na podstawie rejestracji tego zdarzenia, albo, w razie braku tego wpisu, na podstawie ewentualnej informacji o uzyskaniu dyplomu). Drugą możliwością to ograniczenie zakresu pytań badawczych, a w konsekwencji analiz. W przypadku, gdy dane spływają do rejestru z opóźnieniem lub nie obejmują niektórych faktów, może okazać się konieczne ograniczenie okresu lub zestawu zagadnień objętego badaniem. Zaobserwowane luki w danych mogą ponadto skłonić badacza do zrewidowania definicji niektórych pojęć (np. wobec nieobecności w rejestrach ZUS informacji o wysokości wynagrodzeń z tytułu samozatrudnienia i umów o dzieło, przez „łączne wynagrodzenia” określone na podstawie tego rejestru rozumieć można sumę wynagrodzeń z tytułu różnych umów o pracę i różnych umów-zleceń) lub w ostateczności zmiany planu badania (np. w oczekiwaniu na spłynięcie danych do rejestru).

Kolejnym etapem w procesie badawczym jest operacjonalizacja. W badaniach sondażowych tworzy się kwestionariusz, jednocześnie ustalając zestaw zmiennych statystycznych, jakie będą wykorzystywane w analizach. Konstruowanie zmiennych statystycznych na podstawie pytań kwestionariuszowych w badaniu sondażowym jest elementem pomiaru (por. Lissowski i in. 2011: 25). Badacze mają na tym etapie dość dużą swobodę. Muszą jednak pamiętać o ograniczeniach stosowanego narzędzia i samych badanych. Przeszkodą w uzyskaniu pożądaných informacji może być między innymi niewiedza i niepamięć badanych, zbytnia złożoność pytań lub ich drażliwość, czy też skłonność do udzielania nieprawdziwych odpowiedzi. W przypadku badań

opartych na rejestrach swoboda badacza ograniczona jest przede wszystkim przez zawartość rejestrów. Pomiar w przypadku tych badań jest procesem dwustopniowym. Za pierwszy etap odpowiadają twórcy rejestrów administracyjnych, którzy ustalają definicje pojęć oraz sposób przypisywania wartości obiektom należącym do rejestru. Drugim etapem jest tworzenie w zbiorze danych statystycznych nowych zmiennych na podstawie informacji pochodzących z jednego lub kilku rejestrów. Dopiero na tym etapie badacz może ustalać definicje zmiennych statystycznych wykorzystywanych w dalszych analizach. Przyjmowane algorytmy konstruowania zmiennych muszą być zakorzenione w teorii badanego zjawiska.

Oparcie badań na rejestrach nie pozwala na pełną dowolność w definiowaniu pojęć. Jednakże podejście to daje znaczną elastyczność w określaniu zakresu badań. Na podstawie tego samego zestawu rejestrów można skonstruować odmienne zbiory danych odnoszące się do zupełnie różnych zjawisk zachodzących na różnych poziomach. Przykładowo, na podstawie informacji zawartych w rejestrach ZUS możliwe jest badanie zarówno poziomu wynagrodzeń, jak i ruchliwości zawodowej czy też migracji. Konstruowane zmienne mogą dotyczyć zarówno pewnych momentów, okresów czy też osób lub grup. Dzięki temu, że badane elementy są w rejestrach identyfikowalne, możliwa jest synteza informacji o danym obiekcie/zjawisku z różnych rejestrów.

Ponieważ w badaniach opartych na rejestrach, w przeciwieństwie do badań sondażowych, badacz z zasady nie ma kontroli nad procesem gromadzenia danych, niezwykle ważne jest utrzymywanie stałego kontaktu z twórcami i administratorami rejestrów, w szczególności na etapie konstruowania zmiennych statystycznych. Badacz musi bowiem poznać dokładnie ograniczenia i wartość interpretacyjną informacji uzyskanych z rejestrów. Chodzi o znajomość zarówno stosowanych definicji, jak i zasady gromadzenia danych. Pozwala to między innymi zidentyfikować ewentualne luki w kompletności danych. Kontakt z administratorami rejestrów pozwala także śledzić ewentualne zmiany stosowanych definicji oraz sposobów gromadzenia danych. Przykładowo, zamiast korzystać z deklaracji badanych, rejestr może odwoływać się do informacji z innych rejestrów (np. uczelnie obecnie korzystają podczas rekrutacji z danych z rejestru Centralnej Komisji Egzaminacyjnej, choć w przeszłości bazowały na deklaracjach kandydatów na studia), co z zasady poprawia jakość danych, ale może też prowadzić do powielania błędów systematycznych rejestru źródłowego. Efektem współpracy z twórcami rejestrów może też być dodanie nowych zmiennych lub modyfikacja definicji tak, aby odpowiadały one w większym stopniu celom badania. W tym miejscu należy jednak zastrzec, że informacje gromadzone w rejestrach, które nie są wykorzystywane przez instytucje, z zasady są narażone na większe skrzywienia i luki. W wypadku tych zmiennych instytucja nie ma bowiem wystarczającej

motywacji, by zadbać o należyte zbieranie danych. Dobrym przykładem takiego zjawiska jest rubryka „stanowisko”, nadal obecna w formularzach ZUS, choć niewykorzystywana przez tę instytucję w żadnym celu. Informacje o stanowisku wprowadzane są do tych formularzy bez kontroli poprawności. Dlatego dane te nie nadają się do analiz statystycznych.

Od rejestru administracyjnego do bazy danych

W poprzednim podrozdziale przedstawiliśmy logikę prowadzenia badań opartych na rejestrach. Poniżej znajduje się opis technicznych kroków, które należy wykonać, aby na podstawie informacji pochodzących z rejestrów administracyjnych wytworzyć zbiór danych statystycznych zdalnych do analiz.

- Identyfikacja rejestrów i stworzenie możliwości eksportu informacji – wymaga to kontaktu z instytucjami zarządzającymi rejestrem oraz ustalenie zasad organizacyjnych, administracyjnych i prawnych eksportu informacji. Na tym etapie ustalane są między innymi format eksportu danych, zakres informacji i zasady bezpieczeństwa danych.
- Eksport informacji z rejestru.
- Kontrola jakości wyeksportowanych informacji – na tym etapie badane są rozkłady zmiennych źródłowych, identyfikowane są braki danych oraz ustalany ich charakter oraz przyczyny występowania, a także możliwości ich ewentualnego uzupełnienia. W tym celu często bywa niezbędny ponowny kontakt z administratorem rejestru. Możliwe, że wnioski płynące z analizy jakości danych skłonią do ponownego eksportu danych według zmodyfikowanych zasad.
- Selekcja jednostek obserwacji i wybór okresu objętego badaniem – jak już pisaliśmy, na podstawie ustalonego zestawu rejestrów możliwe jest prowadzenie różnych badań. Przed każdym z nich należy ustalić przedmiot badania – populację (osób, instytucji, jednostek administracyjnych, okresów, obiektów, zdarzeń itp.).
- Formułowanie algorytmów – jest to podstawowy etap badania statystycznego. W tym stadium konstruowane są zmienne statystyczne na bazie teorii badanego zjawiska i interpretacji dostępnych danych. Na tym etapie ustalane są zasady syntezy informacji z różnych źródeł: rejestrów, statystyk publicznych i innych badań. Jest realizacją opisanego w poprzednim podrozdziale etapu operacjonalizacji zmiennych, jednak konieczny może okazać się powrót do etapu konceptualizacji pojęć. Dostępne dane mogą bowiem nie pozwalać na opracowanie algorytmów, które stworzyłyby zmienne statystyczne zgodnie z przyjętymi pierwotnie definicjami. Formułowanie algorytmów wymusza precyzyjne ustalenie zawartości

informacyjnej tworzonych zmiennych. Na tym etapie również może być konieczny ponowny kontakt z administratorami rejestru, a nawet ponowny eksport danych według zmodyfikowanych zasad.

- Konstrukcja zbiorów danych na podstawie ustalonych algorytmów.

Potencjał badań opartych na rejestrach

Przedstawiona metodologia jest wynikiem doświadczeń zespołu PEJK UW zgromadzonych w ciągu kilku lat prowadzenia na Uniwersytecie Warszawskim analiz procesów rekrutacyjnych oraz badań losów absolwentów. Nasze wnioski metodologiczne zostały wzbogacone o rozwiązania zawarte w opracowaniach badaczy skandynawskich (Korpi 1989; Knudsen 1993; Harala 1995; Hodgins 1998; Longva i in. 1998; Wallgren i Wallgren 2007) – najbardziej chyba doświadczonych w prowadzeniu badań społecznych opartych na informacjach zawartych w rejestrach administracyjnych (por. UNECE 2007: 5).

Poniżej opisujemy w skrócie rejestry dostępne w Szwecji oraz wybrane typy badań prowadzonych w tym kraju na bazie informacji z tych rejestrów oraz przyjęte tam zasady wykorzystania tego typu danych do prowadzenia badań. W dalszej części tego podrozdziału przedstawimy metodologiczne możliwości prowadzenia analogicznych badań w Polsce. Rozważania zilustrujemy przykładowymi wynikami uzyskanymi przez nasz Zespół.

Badania oparte na rejestrach szwedzkich

Coroczne opracowania dotyczące zarobków i podatków odprowadzanych w Szwecji przygotowywane są na podstawie szeregu rejestrów. Na podstawie deklaracji podatkowych obywateli zawierających informacje o odprowadzonych podatkach, uzyskanych dochodach i rozmaitych świadczeniach uzupełniany jest rejestr Urzędu Skarbowego Szwecji (szw. *Skatteverket*, ang. *National Tax Board*). Na tej podstawie konstruowana jest baza danych o około 9 milionach obywateli zawierająca rokrocznie około 300 zmiennych. Baza ta dostarczana jest do Urzędu Statystycznego Szwecji (szw. *Statistiska centralbyrån*, ang. *Statistics Sweden*), tworząc podstawę rejestru dochodów i podatków. Do tych danych dołączane są informacje z dziesięciu rejestrów – zarówno z wewnętrznych rejestrów Urzędu Statystycznego Szwecji (zawierających między innymi informacje o związkach pokrewieństwa między obywatelami, miejscu zamieszkania, wykształceniu, branży zatrudnienia), jak i z zewnętrznych źródeł (między innymi centralne i samorządowe instytucje emerytalne, instytucje

stypendialne, ośrodki pomocy społecznej). Procedura łączenia wykorzystuje numery identyfikacyjne obywateli (szw. *personnummer*, ang. *PIN, Personal Identification Number*), szwedzkie odpowiedniki polskich numerów PESEL⁴.

Na podstawie połączonych danych tworzone są bazy danych opisujące różne populacje – zarówno osób, jak i gospodarstw domowych. Informacje te pozwalają na prowadzenie systematycznych badań społecznych prowadzonych przez Urząd Statystyczny Szwecji. Są również udostępniane do prowadzenia badań akademickich. Dostęp szwedzkiego środowiska naukowego do tych danych statystycznych zaowocował rozkwitem badań społecznych opartych na informacjach z rejestrów administracyjnych w tym kraju. Zakres tych badań jest niezwykle szeroki. Obejmuje między innymi zagadnienia dystrybucji dochodów, geograficznych nierówności dochodowych, wzdlużne analizy wynagrodzeń, badania edukacyjne itp. Przykładem jest monitoring losów absolwentów szkół różnego poziomu prowadzony z powodzeniem w tym kraju. Pozwolił on na dogłębne zdiagnozowanie mechanizmów kryzysu rynku pracy, który Szwecja przechodziła we wczesnych latach dziewięćdziesiątych XX wieku. Kryzys jest z zasady najbardziej dolegliwy dla osób wchodzących na rynek pracy. Okazało się, że jego skutki były znacznie mniej odczuwalne dla absolwentów uniwersytetów niż dla osób kończących szkoły średnie (szw. *Gymnasieskola*). Wyniki studium, niezwykle cenne dla kształtowania polityki społecznej, uzyskano przy bardzo niskich kosztach. Roczny koszt systematycznie prowadzonych badań aktywności zawodowej i edukacyjnej wszystkich obywateli Szwecji wynosi około 50 000 dolarów, z czego połowa to koszty druku (Wallgren i Wallgren 2007: 8–9). Badania losów absolwentów szkół wyższych stanowią jedynie niewielką część tych studiów.

Możliwość łączenia informacji pochodzących z różnych rejestrów pozwala na prowadzenie badań interdyscyplinarnych o zakresie znacznie szerszym, niż mogłoby to mieć miejsce w razie osobnego korzystania z tych informacji. Ciekawym przykładem są badania związków chorób psychicznych i zachowań przestępczych prowadzone w Szwecji, Danii i Finlandii na podstawie danych z rejestrów psychiatrycznych i kryminalnych tych krajów (zob. np. Hodgins 1992, 1998).

⁴Opracowana przez zespół PEJK UW procedura bezpiecznego łączenia informacji zapewniająca anonimowość badanym przebiega w kilku etapach. Numery identyfikacyjne jednostek (np. PESEL) są pseudonimizowane na poziomie pojedynczych rejestrów przy użyciu szeregu kluczy szyfrujących. Na podstawie spseudonimizowanych numerów identyfikacyjnych odbywa się łączenie informacji. Po połączeniu informacji spseudonimizowane numery identyfikacyjne są dodatkowo szyfrowane osobnym kluczem, a wszystkie klucze szyfrowania są następnie usuwane, a zbiory danych są sortowane tak, by uniemożliwić identyfikację jednostek na podstawie ich kolejności w bazie. Dane mogą być dodatkowo losowo „zaszumiane”, co uniemożliwia identyfikację jednostek nawet na podstawie znajomości unikatowej kombinacji wartości zmiennych.

Rozwój szwedzkich badań opartych na rejestrach administracyjnych był możliwy dzięki jasnym i precyzyjnym regułom prawnym i administracyjnym zapewniającym bezpieczeństwo prywatności obywateli podczas prowadzenia badań statystycznych. Dane powszechnie udostępniane do badań naukowych są wcześniej anonimizowane.

Anders i Britt Wallgren (2007: 2) przedstawiają cztery zasady wykorzystywania danych administracyjnych do prowadzenia badań statystycznych.

1. Urzędy statystyczne powinny mieć dostęp do rejestrów administracyjnych prowadzonych przez podmioty publiczne. Podstawą tego uprawnienia powinny być przepisy prawa gwarantujące ochronę prywatności.

2. Dane ze wspomnianych rejestrów powinny być przekształcane w rejestry statystyczne, łączące informacje z wielu źródeł.

3. Rejestry statystyczne powinny tworzyć zintegrowany system rejestrowy. System ten powinien zapewniać możliwość efektywnego wykorzystania wszystkich danych.

4. Uzgodnienie definicji pojęć i spójność danych z rejestrów dotyczących różnych populacji jest warunkiem koniecznym zgodności wyników uzyskiwanych na bazie różnych rejestrów.

Warunkiem umożliwienia środowiskom naukowym dostępu do informacji pochodzących z rejestrów administracyjnych jest zintegrowanie działań badawczych. Próba obsłużenia odrębnych wniosków o dostęp do informacji składanych przez wielu badaczy zdezorganizowałaby pracę administratorów rejestrów. W Szwecji problem rozwiązano poprzez działania Urzędu Statystycznego Szwecji, który integruje dane i udostępnia je do badań. W większości krajów Europy nie ma takiego rozwiązania. Problemy administracyjne i prawne stające na przeszkodzie środowiskom naukowym w Polsce opisujemy w końcowej części artykułu. Do czasu systemowego rozwiązania wskazanych problemów warunkiem minimalnym nawiązania współpracy z instytucjami zarządzającymi rejestrami jest koordynacja projektów badawczych w celu redukcji obciążenia administratorów rejestrów dodatkowymi obowiązkami.

Rejestry w Polsce

Potencjał badawczy polskich rejestrów administracyjnych nie wydaje się mniejszy od potencjału rejestrów szwedzkich. Zawartość informacyjna naszych rejestrów jest niezwykle bogata. Okolicznością sprzyjającą wykorzystaniu danych z rejestrów jako źródła informacji do badań statystycznych jest istnienie powszechnie wykorzystywanego systemu PESEL (Powszechny Elektroniczny System Ewidencji Ludności). Dzięki temu w większości

rejestrów administracyjnych obecny jest identyfikator (numer PESEL), pozwalający na integrację informacji zawartych w rozmaitych rejestrach.

Szczegółowe charakterystyki zawartości rejestrów administracyjnych w Polsce są przedstawione w dokumentach regulujących funkcjonowanie instytucji zarządzających tymi rejestrami i na ogół dostępne są na stronach internetowych tych instytucji. Poniżej omawiamy, w których rejestrach znajdują się wybrane informacje, mogące być podstawą prowadzenia badań społecznych.

Informacje o zarobkach i zatrudnieniu znaleźć można zarówno w rejestrze Zakładu Ubezpieczeń Społecznych, Kasy Rolniczego Ubezpieczenia Społecznego, jak też w rejestrach Urzędów Skarbowych. W rejestrach ZUS znajduje się również szereg innych informacji charakteryzujących sytuację zawodową: zatrudnienie (szczegółowa forma zatrudnienia, miesiąc zatrudnienia), miejsce pracy (adres siedziby, branża, liczba pracowników), doświadczenie bezrobocia. W rejestrach Urzędów Skarbowych, na podstawie deklaracji podatkowych, umieszczane są informacje o formach zawartych umów klasyfikowane według innych kryteriów niż te obecne w rejestrach ZUS. Poza tym w rejestrach ZUS nieobecne są np. informacje o zarobkach pochodzących z umów o dzieło, które rejestrowane są przez Urzędy Skarbowe.

Informacje pozwalające scharakteryzować sytuację zdrowotną ludności znajdują się zarówno w rejestrach Narodowego Funduszu Zdrowia, jak i ZUS. Rejestry NFZ zawierają informacje o przebytych chorobach, zastosowanych terapiach i wydatkach poniesionych na leki (na receptę), natomiast rejestry ZUS i KRUS – dane o niepełnosprawności, wypłaconych rentach oraz świadczeniach chorobowych.

Informacje charakteryzujące historię miejsc zamieszkania oraz zameldowania znajdują się zarówno w rejestrze PESEL, jak w innych rejestrach administracyjnych. Związane jest to z koniecznością komunikowania się poszczególnych instytucji z obywatelem (płatnikiem składek ubezpieczeniowych, podatnikiem, pacjentem itp.). W rejestrze PESEL ponadto zawarte są między innymi informacje o związkach małżeńskich. Nie ma natomiast informacji jednoznacznie identyfikujących więzi pokrewieństwa. Więzy te można by określać, gdyby gromadzono dane nie tylko o imionach i nazwiskach rodziców, lecz również ich numery PESEL⁵.

Informacje o elementach historii edukacyjnych zawarte są zarówno w rejestrach centralnych (rejestry Centralnej Komisji Egzaminacyjnej, system informacji o szkolnictwie wyższym POL-on), jak i w rejestrach elektronicznych

⁵ Numery PESEL mogą z powodzeniem zastąpić pozostałe dane rodziców, gdyż dane te są odtwarzalne na podstawie numeru PESEL. Imiona i nazwiska, nawet uzupełnione o nazwiska rodowe i adresy, nie identyfikują osób jednoznacznie.

poszczególnych uczelni. Dane gromadzone przez CKE obejmują wyniki wszystkich egzaminów zewnętrznych prowadzonych w Polsce, czyli: sprawdziany w VI klasie zdawane na koniec szkoły podstawowej, egzamin gimnazjalny, egzamin maturalny oraz egzaminy zawodowe. Dostępne są zatem informacje na temat osiągnięć uczniów na różnych etapach nauki, nie tylko z egzaminów maturalnych. Stwarza to znaczne możliwości w badaniach ścieżek edukacyjnych. Wszystkie uczelnie gromadzą dane na temat osiągnięć swoich studentów – od etapu rekrutacji do uzyskania dyplomu. Znaczna część tych danych istnieje w formie rejestrów elektronicznych. Uczelnie w Polsce korzystają z różnych systemów zarządzania rejestrami. Największą popularność zdobył system USOS (Uniwersytecki System Obsługi Studiów) i serwisy stowarzyszone, w tym w szczególności IRK (Internetowa Rejestracja Kandydatów). USOS jest przykładem komputerowego systemu służącego do zarządzania tokiem studiów. Został stworzony ponad 10 lat temu na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego, a obecnie jest wykorzystywany przez kilkadziesiąt uczelni w kraju, w tym przez te największe – Uniwersytet Warszawski, Uniwersytet Jagielloński i Uniwersytet im. A. Mickiewicza w Poznaniu⁶.

Centralny system POL-on zawiera między innymi informacje na temat wszystkich studentów, w tym dane na temat kierunku studiów i dat uzyskania dyplomu. Z założenia są to informacje znacznie mniej dokładne niż dane posiadane przez uczelnie, jednakże ich zaletą jest kompletność. W systemie POL-on nie są gromadzone dokładne wyniki z egzaminów itd. Uczelnie przekazują jedynie dość podstawowe informacje na temat swoich studentów.

Nie są to oczywiście wszystkie rejestry administracyjne w Polsce. Nawet pobieżne omówienie ich zawartości mogłoby stać się tematem osobnego opracowania. Przedstawione powyżej pobieżne omówienie polskich rejestrów publicznych pokazuje olbrzymi potencjał badawczy informacji zawartych w tych rejestrach.

Warto przy okazji wspomnieć, że ważnym problemem, zarówno ze względów administracyjnych, jak i badawczych, jest brak przepływu informacji między zdeintegrowanymi rejestrami. Po pierwsze, prowadzi to do nadmiernego obciążania obywateli obowiązkiem wielokrotnego podawania tych samych informacji na użytek wielu rejestrów. Z jednej strony naraża to obywateli na ryzyko pominięcia któregoś rejestru, z drugiej – każda sytuacja ręcznego wprowadzania danych jest potencjalnym źródłem błędów. W konsekwencji zwiększa się ryzyko niespójności danych w różnych rejestrach. Po drugie, integracja rejestrów publicznych umożliwiłaby znaczne

⁶ Obszerny opis systemu USOS oraz powiązanych z nim serwisów znajdzie Czytelnik na stronach: <http://usosownia.uw.edu.pl/start> oraz <http://usos.edu.pl/inne-projekty>.

uproszczenie procedur administracyjnych, zmniejszając zarazem obciążenia państwa.

Integracja informacji zawartych w przedstawionych rejestrach stworzyła by znaczne możliwości prowadzenia badań statystycznych w Polsce. Przykładowo, możliwa byłaby analiza ścieżek edukacyjno-zawodowych uczniów i studentów, jak również uwarunkowań sukcesu zawodowego czy ruchliwości zawodowej w Polsce. Informacje wyeksportowane z rejestrów mogłyby stać się podstawą do badania mechanizmów wyborów i barier edukacyjnych różnych grup (np. niepełnosprawnych, przewlekle chorych itp.) na kolejnych etapach kształcenia. Dane adresowe zawarte w rejestrach są bardzo dobrym materiałem badawczym do prowadzenia złożonych studiów migracji. Na uwagę zasługuje możliwość ustalania skuteczności i faktycznych kosztów terapii, a także ryzyka ubezpieczeniowego ZUS na podstawie zintegrowanych informacji pochodzących z rejestrów NFZ i ZUS. Koszty związane z chorobą lub niepełnosprawnością to nie tylko koszty terapii, ale również wydatki związane ze świadczeniami ZUS (np. renty, zasiłki chorobowe, świadczenia rehabilitacyjne).

Wybrane wyniki badań

Przykładem badań prowadzonych z wykorzystaniem informacji z rejestrów są badania różnych etapów trajektorii studiowania oparte na USOS i IRK.

Konsekwencją stworzenia rejestrów elektronicznych, w których gromadzone są między innymi informacje dotyczące przebiegu studiów każdego ze studentów, jest stworzenie warunków do prowadzenia badań trajektorii studiowania – ścieżek edukacyjnych.

Jako pierwsze zostały wykorzystane dane pochodzące z systemu IRK. Od roku 2006 przygotowywane są na ich podstawie opracowania dotyczące przebiegu procesu rekrutacji (m.in. Batorski 2006; Abramczuk 2008; Zajac 2013). Początkowo opracowania dotyczyły zagadnień takich jak liczba kandydatów, liczba dokonywanych przez nich rejestracji na studia, współkandydowania na wiele kierunków studiów. W późniejszych latach raporty dotyczące przebiegu rekrutacji zostały rozbudowane o kolejne elementy, jak np. części poświęcone miarom popularności kierunków, wynikom maturalnym i rekrutacyjnym kandydatów czy analizie pochodzenia kandydatów na studia drugiego stopnia. Dodawanie nowych modułów w raportach dotyczących rekrutacji wymagało w coraz większym stopniu wykorzystywania danych pochodzących także z systemu USOS, np. danych dotyczących uzyskanych dyplomów.

Analizy prowadzone na potrzeby władz uczelni były punktem wyjścia do przygotowania pogłębionych studiów. Dane pochodzące z IRK posłużyły do badań nad sposobami rekonstrukcji preferencji kandydatów na zbiorze kierunków studiów opartych na podejmowanych przez nich decyzjach oraz badań nad zachodzącymi mechanizmami autoselekcji kandydatów w zbiorowości kandydatów na studia pierwszego stopnia (Zajac 2011). Innym przykładem wykorzystania danych pochodzących z rejestrów uczelni jest analiza mobilności studentów w czasie przejścia między studiami pierwszego stopnia oraz studiami drugiego stopnia na Uniwersytecie Warszawskim (Zajac 2012). Uzyskane wyniki pokazywały, że niewiele osób korzysta z szansy zmiany kierunku kształcenia po ukończeniu studiów pierwszego stopnia, jaką stworzyło wdrażanie procesu bolońskiego.

Z punktu widzenia uczelni najważniejsze spośród badań opartych wyłączenie na rejestrach IRK i USOS wydają się jednak badania dotyczące wartości prognostycznej kryteriów rekrutacji na studia, w szczególności wartości prognostycznej świadectw maturalnych. W badaniach tych analizowany jest związek pomiędzy wynikiem rekrutacyjnym kandydatów oraz sukcesem w studiowaniu (Zajac 2010). Uzyskane wyniki pokazały znaczne zróżnicowanie skuteczności przyjmowanych na poszczególnych kierunkach kryteriów rekrutacyjnych w przewidywaniu sukcesu na studiach (Zajac 2013).

Rozwinięciem badań prowadzonych na podstawie USOS i IRK jest projekt badawczy realizowany przez Pracownię Ewaluacji Jakości Kształcenia UW pt. „Monitorowanie losów absolwentów uczelni wyższych z wykorzystaniem danych administracyjnych Zakładu Ubezpieczeń Społecznych”⁷. W ramach tego projektu opracowana została metodologia badań losów absolwentów na bazie informacji z rejestrów administracyjnych. Opracowano szereg wskaźników, których wartość informacyjna i trafność wynika z jednej strony z oparcia ich na wiarygodnych danych rejestrowych gromadzonych według stabilnych i kontrolowanych reguł administracyjnych, z drugiej zaś – z zestawiania ich z publicznymi statystykami charakteryzującymi lokalne rynki pracy i sytuację gospodarczą.

Wśród pytań badawczych postawionych w opisywanym badaniu znalazło się między innymi pytanie o typowe i nietypowe ścieżki edukacyjne i zawodowe absolwentów Uniwersytetu Warszawskiego.

⁷ Projekt zrealizowany na zamówienie Instytutu Badań Edukacyjnych w ramach projektu „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego” współfinansowanego z Europejskiego Funduszu Społecznego, w ramach Programu Operacyjnego Kapitał Ludzki, Priorytet III-Wysoka jakość systemu oświaty: Działanie: 3.1-Modernizacja systemu nadzoru i zarządzania w oświacie, Poddziałanie: 3.1.1- Tworzenie warunków do monitorowania ewaluacji i badań systemu oświaty.

W trakcie badania ZUS udostępnił PEJK UW informacje o losowej, testowej próbie 2000 absolwentów UW⁸ z lat 2007–2008 w postaci uniemożliwiającej połączenie ich z danymi charakteryzującymi ich historię edukacyjną. Z tego powodu niemożliwe było zbadanie powiązań między sukcesem edukacyjnym i zawodowym oraz scharakteryzowanie historii zawodowych absolwentów poszczególnych kierunków studiów, jednak dane te okazały się wystarczające do przedstawienia wiarygodnych wyników opisujących sytuację zawodową absolwentów Uniwersytetu Warszawskiego w ciągu 60 miesięcy objętych badaniem.

W wylosowanej próbie absolwentów UW (z lat 2007–2008) 500 osób było absolwentami studiów pierwszego stopnia, a pozostałe 1500 to absolwenci studiów jednolitych magisterskich lub studiów drugiego stopnia.

Poniżej przedstawiamy wybrane analizy odnoszące się do trzech form zatrudnienia, w niniejszym tekście określanych jako:

- etat,
- zlecenie,
- samozatrudnienie.

Kategorie te zostały ustalone na podstawie kodów tytułu ubezpieczenia stosowanych przez Zakład Ubezpieczeń Społecznych.

Przez osobę wykonującą pracę etatową będziemy rozumieć pracownika wykonującego pracę najemną o stałym charakterze, np. na podstawie różnych typów umów o pracę, osobę przebywającą na urlopie wychowawczym itp.

Przez osobę pracującą na zlecenie będziemy rozumieć osobę wykonującą pracę o niestałym charakterze, np. na podstawie umowy-zlecenia, umowy agencyjnej itp. Do tej kategorii nie zaliczają się osoby pracujące na umowy o dzieło, gdyż od tych umów nie są odprowadzane składki do ZUS, a zatem takie osoby nie figurują w jego rejestrach.

Przez osobę samozatrudnioną będziemy rozumieć osobę odprowadzającą składki na podstawie własnej działalności gospodarczej.

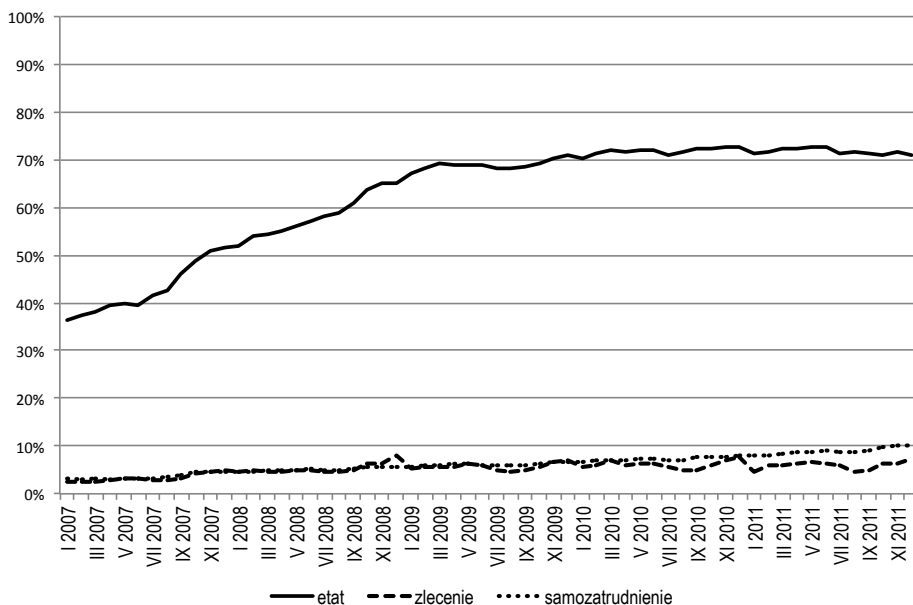
Badanie dotyczy losów zawodowych absolwentów z dwóch roczników – 2007 i 2008 – w okresie od stycznia 2007 do grudnia 2011 roku. Oznacza to, że przez pierwsze dwa lata nie wszyscy badani byli już absolwentami UW. Dlatego w analizach należy odróżniać dwa okresy: od stycznia 2007 do grudnia 2008 roku i od stycznia 2009 do grudnia 2011 roku. Dynamika zjawisk

⁸ Była to próba losowa bezzwrotna dyplomów na UW. Osoby, które uzyskały w latach 2007 i 2008 więcej niż jeden dyplom, mogły zostać wylosowane więcej niż jeden raz. W wylosowanej próbie były dwie osoby wylosowane dwukrotnie. Przekazana lista liczyła zatem 1998 różnych numerów PESEL. Zbiór danych uzyskany z ZUS obejmował 1913 osób. Pozostałe osoby nie figurowały w rejestrach ubezpieczonych ZUS, co oznacza, że do dnia eksportu danych (27 lipca 2012 roku) nie została za te osoby odprowadzona żadna składka do ZUS. Stanowią one niespełna 5% próby.

obserwowanych w pierwszym okresie jest powiązana również z dynamiką kończenia studiów przez badanych – rozpoczęciem funkcjonowania na rynku pracy jako absolwenci.

Przedstawione wyniki dotyczą absolwentów studiów drugiego stopnia i jednolitych magisterskich.

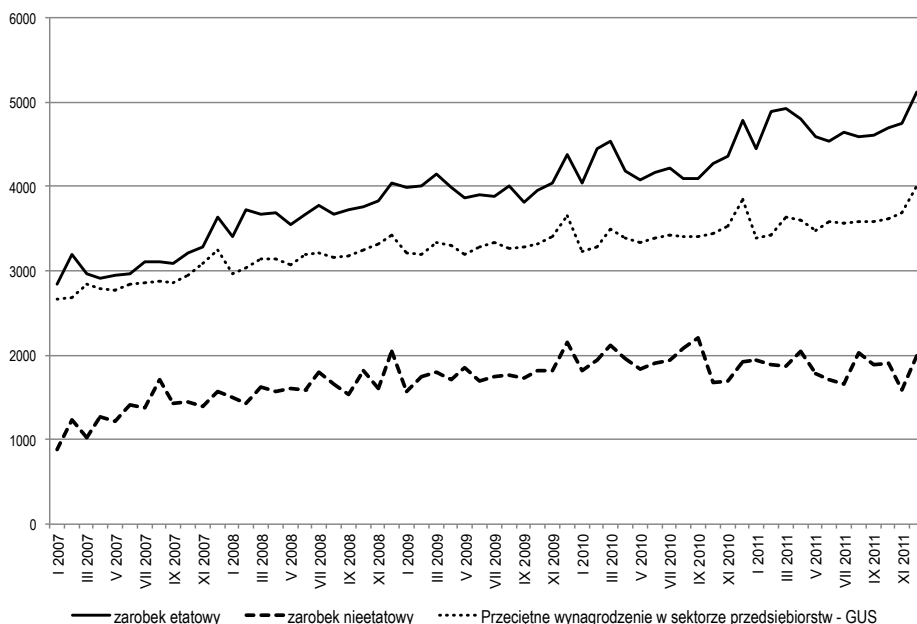
Rysunek 1. Liczba badanych z tytułem magistra zatrudnionych na etat, w ramach umów zleceń oraz samozatrudnionych (w %)



Na powyższym wykresie przedstawione są procenty badanych z tytułem magistra w różnych formach zatrudnienia w poszczególnych miesiącach. Przez cały okres objęty badaniem najbardziej typową formą zatrudnienia było zatrudnienie na etat, a procenty osób zatrudnionych na zlecenie i samozatrudnionych są zbliżone i wyraźnie mniejsze niż procenty zatrudnionych na etat. Średnia dokładność oszacowania procenta zatrudnionych na etat, zlecenie i samozatrudnionych wynosiła na poziomie ufności 95%, odpowiednio, 2,3%, 1,1% i 1,1%.

W pierwszej części okresu badania (tj. w 2007 i 2008 roku) procent osób zatrudnionych na etat wyraźnie wzrastał, stabilizując się w drugiej części okresu badania (I 2009 – XII 2011). Wynika to z faktu, że w pierwszym okresie badani uzyskiwali tytuł magistra, co nie pozostaje bez wpływu na ich sytuację na rynku pracy. Udział osób pracujących na zlecenie i samozatrudnionych nieznacznie wzrastał przez cały okres objęty badaniem.

Rysunek 2. Przeciętne zarobki* w kolejnych miesiącach badanych z tytułem magistra zatrudnionych na etat lub w ramach umów zleceń (w zł)



* wartości obliczone wyłącznie dla osób, które w danym miesiącu były zatrudnione

Powyższy wykres przedstawia średnie zarobki z różnych źródeł dla pracujących badanych z tytułem magistra. Średnia dokładność oszacowania zarobków z tytułu zatrudnienia na etat i na zlecenie wynosiła na poziomie ufności 95%, odpowiednio, 192 zł i 390 zł. Wysokim zarobkom etatowym absolwentów UW towarzyszyło znaczne zróżnicowanie – odchylenie standardowe zarobków etatowych w miesiącach objętych badaniem wynosiło od 2409 zł do 4880 zł, przy czym średnia wartość odchylenia standardowego w badanym okresie wynosiła 3149 zł. Dla porównania przedstawiono także przeciętne wynagrodzenie w sektorze przedsiębiorstw w danym miesiącu na podstawie danych Głównego Urzędu Statystycznego. Przez cały okres badania pracujący etatowo absolwenci studiów magisterskich UW uzyskiwali średnio zarobki powyżej średnich zarobków w sektorze przedsiębiorstw, a z czasem różnica ta zauważalnie rosła. Zwraca uwagę wyraźna zgodność dynamiki etatowych zarobków absolwentów UW i wynagrodzeń w gospodarce narodowej.

Przedstawione wyniki opierają się na informacjach z rejestrów ZUS niepołączonych z informacjami z rejestrów uczelni. Już te dane pozwoliły na szereg interesujących wniosków, sformułowanych w ramach realizacji omawianego

projektu⁹. Stworzenie możliwości łącznej analizy informacji o przebiegu studiów (w tym m.in. o kierunku, stopniu, formie, stypendiach, uzyskanych ocenach itp.) i kariery zawodowej umożliwiłoby systematyczne prowadzenie dynamicznych analiz pozwalających określić związki między wynikami osiąganymi w trakcie studiów a sukcesem na rynku pracy.

Wykorzystanie analiz wzdluznych pozwoliłoby monitorować procesy nauczania. Zestawienie informacji na temat powodzenia na rynku pracy, w tym np. czasu poszukiwania pracy czy też uzyskiwanych zarobków, z wynikami na studiach pozwoliłoby na określenie wymiernych efektów kształcenia. Umożliwiłoby to władzom uczelni ocenę dostosowania przygotowanych przez nie programów studiów do potrzeb rynku pracy, a następnie na wprowadzenie potrzebnych modyfikacji.

Dostęp do sprawdzonych i porównywalnych wyników badań ułatwiłby dokonywanie wyborów przez kandydatów na studia, studentów, absolwentów i pracodawców. Kandydaci na studia stoją bowiem przed trudną decyzją – wyborem kierunku i uczelni. Ze względu na brak rzetelnych badań, przy podejmowaniu decyzji brakuje im ważnych dla nich informacji na temat typowej trajektorii studiowania czy przeciętnej sytuacji absolwentów na rynku pracy. Luka informacyjna jest częściowo zapełniana przez rozmaite mniej lub bardziej rozpowszechnione opinie na temat poszczególnych kierunków i często stereotypowe przekonania o sytuacji ich absolwentów.

Wyniki analiz informacji pochodzących z rejestrów pozwoliłyby kandydatom na studia zweryfikować swoje intuicje. Kompleksowy charakter analiz, obejmujący zarówno rekrutację, jak i przebieg studiów oraz funkcjonowanie na rynku pracy po ukończeniu studiów, pozwoliłby kandydatom ocenić swoje szanse na bycie przyjętym na określony kierunek, prawdopodobieństwo pomyślnego ukończenia studiów oraz możliwe scenariusze kariery zawodowej po uzyskaniu dyplomu. Dzięki dostępowi do nowych, rzetelnych danych dotyczących poszczególnych kierunków studiów kandydaci mogliby podejmować decyzje o podjęciu studiów na podstawie szerszego zestawu informacji.

Ustalenie związków historii edukacyjnych z sukcesem na rynku pracy jest bardzo ważne również dla studentów. Podczas studiów podejmowane są liczne decyzje edukacyjne i zawodowe (zmiana kierunku studiów, studia równoległe, podjęcie pracy w czasie studiów, staże itp.). Dostęp do opisanych wyników umożliwi oszacowanie konsekwencji rozmaitych strategii i dzięki temu ułatwi dokonywanie wyborów.

Pracodawcy na podstawie sprawdzonych informacji o mechanizmach selekcji podczas studiów, a także o zatrudnieniu absolwentów (w szczególności

⁹ Szerszy opis wyników uzyskanych w czasie analiz danych ZUS, USOS oraz IRK znajduje się w końcowym raporcie z realizacji projektu.

np. o jego trwałości), mogliby wyrobić sobie pogląd na temat jakości wykształcenia osób uzyskujących dyplomy.

Jak widać, już samo prowadzenie regularnych badań losów absolwentów z wykorzystaniem informacji pochodzących z rejestrów stworzyłoby świetne podstawy do dialogu między uczestnikami rynku pracy i edukacyjnego, który byłby oparty na rzetelnej wiedzy, sprzyjając sprawniejszemu funkcjonowaniu tych rynków. Rozpowszechnienie tego typu badań statystycznych w innych dziedzinach nauk społecznych służyłoby dobru publicznemu (por. Dz.U. 1997 Nr 133 poz. 883: art. 23 ust. 1 pkt 4) i pozwoliłoby na optymalizację wydatków państwa. Przeszkodą w rozwoju badań opartych na rejestrach w naszym kraju są jednak szkodliwe praktyki administracyjne. Praktyki te uniemożliwiły jak dotąd pełną realizację planów badawczych założonych we wspomnianym projekcie.

Przedstawione wyniki oparte zostały na informacjach, które miały służyć wyłącznie celom testowym (ustaleniu zawartości informacyjnej rejestrów ZUS) w ramach przygotowań do prowadzenia badań na populacji absolwentów UW. Ograniczenie analiz wyłącznie do próbki testowej zostało wymuszone w związku z zablokowaniem głównej części badania przez Generalnego Inspektora Ochrony Danych Osobowych. Stało się tak pomimo braku jednoznacznych przeszkód prawnych do prowadzenia takich analiz. Instytut Badań Edukacyjnych przy wsparciu UW i ZUS przygotował szczegółową ekspertyzę prawną, która wykazała pełną zgodność z prawem działań badawczych założonych w projekcie (zob. m.in. Dz.U. 1997 Nr 133 poz. 883: art. 2 ust 3, art. 25 ust. 2 pkt 3 i art. 27 ust. 2 pkt 9, Dz.U. 1998 Nr 137 poz. 887: art. 68 ust. 1 pkt 4 i art. 74 ust. 3 pkt 10). Na uwagę zasługuje, że zanik współpracy zespołu badawczego PEJK UW i służb statystycznych ZUS został spowodowany jedynie słabo uargumentowanym prawnie „wyrażeniem wątpliwości” przez urzędnika GODO wobec planowanych działań¹⁰. Szkoda wywołana opisaną obstrukcją administracyjną jest tym większa, że uniemożliwiona działaniami GODO kompletna analiza miała stanowić ekspertyzę dla przygotowywanych w Sejmie zapisów nowelizacyjnych ustawy Prawo o szkolnictwie wyższym.

Generalnie rzecz biorąc, na przeszkodzie prowadzeniu w Polsce systematycznych badań statystycznych z wykorzystaniem informacji pochodzących z rejestrów administracyjnych nie stoją przepisy prawne (zob. np. Dz.U. 1995 Nr 88 poz. 439: art. 13 ust. 3, Dz.U. 1997 Nr 133 poz. 883: art. 2 ust 3, art. 25 ust. 2 pkt 3, art. 27 ust. 2 pkt 9 i art. 43 ust. 1 pkt 10). Realną barierą jest w tym wypadku praktyka będąca konsekwencją szkodliwych działań administracyjnych

¹⁰ Analiza tego przypadku jako przykładu destabilizacji zgodnych z prawem działań instytucji państwowych i badań akademickich poprzez niewiążące prawnie wątpliwości urzędników to skądinąd ciekawy temat studium z dziedziny socjologii organizacji.

jak opisane powyżej. Wydaje się, że do usunięcia tej przeszkody wystarczyłoby ustalenie jasnych procedur administracyjnych bezpiecznego korzystania z połączonych informacji pochodzących z różnych rejestrów do celów badań naukowych – na poziomie regulaminów i rozporządzeń. Być może remedium na opisany problem byłoby wpisanie badań opartych na rejestrach administracyjnych do programu badań statystycznych statystyki publicznej (zob. Dz.U. 1995 Nr 88 poz. 439: art. 13 ust. 3, art. 15 ust. 3 pkt 1 i art. 18).

Podsumowanie

W artykule przedstawiliśmy szereg zalet badań opartych na danych pochodzących z rejestrów administracyjnych. Za wykorzystaniem takich danych w badaniach społecznych przemawiają nie tylko niskie w porównaniu z badaniami sondażowymi koszty, ale także względy metodologiczne. Rejestry obejmują całe populacje obiektów i zapewniają dostęp do szczegółowo zapisanych sekwencji zdarzeń. Nie dotyczą ich problemy związane z pamięcią badanych lub niechęcią do udzielania odpowiedzi, z którymi muszą borykać się badacze korzystający z sondaży.

Od wielu lat rejestry administracyjne są wykorzystywane z powodzeniem w badaniach przez naukowców z krajów skandynawskich. Instytucje państwowe odpowiadają za gromadzenie danych, agregowanie zbiorów i udostępnianie ich badaczom. Efektem są np. wspomniane wcześniej szwedzkie populacyjne badania aktywności edukacyjnej i ekonomicznej ludności, której realizacja przy wykorzystaniu sondażu nie byłaby możliwa.

Także w Polsce gromadzone są znaczne ilości informacji w rejestrach administracyjnych (państwowych, samorządowych, uczelnianych itd.). W przyszłości można oczekiwać dalszego rozwoju tych zbiorów informacji. Dane te są jednak wykorzystywane niemal wyłącznie do bieżącej obsługi klientów poszczególnych instytucji.

Dostęp do informacji jest warunkiem koniecznym dokonywania racjonalnych wyborów, zarówno przez jednostki, jak i instytucje. Niewykorzystywanie istniejących, sprawdzonych informacji jest błędem. Wdrożenie w Polsce badań opartych na danych pochodzących z rejestrów publicznych pozwoliłoby na zapewnienie takiego dostępu różnym instytucjom kształtującym politykę państwową. Nie ma sensownego powodu, aby wyniki analiz statystycznych nie były również dostępne społeczeństwu. Dostęp ten byłby użyteczny zarówno przy podejmowaniu decyzji osobistych (wybór kierunku studiów, miejsca pracy, miejsca zamieszkania, zgoda na terapię itp.), jak również decyzji obywatelskich (poparcie programów społecznych itp.). Realizuje to zarazem prawo obywateli do informacji o stanie spraw publicznych.

Użycie danych w badaniach jest skutecznie utrudniane przez brak odpowiednich (zrozumiałych i jednoznacznych) przepisów prawnych i procedur administracyjnych, które zapewniłyby obywatelom i administracji dostęp do rzetelnej informacji statystycznej o stanie społeczeństwa, gwarantując jednocześnie pełną ochronę prywatności osób, których te informacje dotyczą. Rozwiązania takie są znane i z powodzeniem stosowane.

Literatura

- Abramczuk, Katarzyna. 2008. *Rekrutacja 2008*. Warszawa: Uniwersytet Warszawski.
- Batorski, Dominik. 2006. *Rekrutacja 2006/2007 na studia na Uniwersytecie Warszawskim. Analiza danych rekrutacyjnych*. Warszawa: Uniwersytet Warszawski.
- Bożykowski, Marek, Marcin Dwórznik, Ewa Giermanowska, Albert Izdebski, Mikołaj Jasiński, Joanna Konieczna-Sałamatin, Marek Styczeń i Tomasz Zajęc. 2014. *Monitorowanie losów absolwentów uczelni wyższych z wykorzystaniem danych administracyjnych Zakładu Ubezpieczeń Społecznych. Raport końcowy*. Warszawa: Instytut Badań Edukacyjnych.
- Durkheim, Emile. 2006. *Samobójstwo: Studium z socjologii*. Warszawa: Oficyna Naukowa.
- Dz.U. 1995 Nr 88 poz. 439. Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej.
- Dz.U. 1997 Nr 133 poz. 883. Ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych.
- Dz.U. 1998 Nr 137 poz. 887. Ustawa z dnia 13 października 1998 r. o systemie ubezpieczeń społecznych.
- Freedman, David. A. 2001. *Ecological Inference and the Ecological Fallacy*. W: N. J. Smelser i P. B. Baltes (red.). *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier.
- Goyder, John i Jean McKenzie Leiper. 1985. *The Decline in Survey Response: A Social Values Interpretation*. „Sociology” (19): 55–71.
- Harala, Riitta. 1995. *Evaluation of the Results of the Register Based Population and Housing Census 1990 in Finland*. „Statistical Journal of the United Nations Economic Commission for Europe” 12(1): 63–72.
- Hodgins, Sheilagh. 1992. *Mental Disorder, Intellectual Deficiency and Crime: Evidence from a Birth Cohort*. „Archives of General Psychiatry” 49(6): 476–483.
- Hodgins, Sheilagh. 1998. *Epidemiological Investigations of the Associations Between Major Mental Disorders and Crime: Methodological Limitations and Validity of the Conclusions*. „Social Psychiatry and Psychiatric Epidemiology” 33: 29–37.
- Jahoda, Marie, Paul Felix Lazarsfeld i Hans Zeisel. 1974. *Marienthal: The Sociography of an Unemployed Community*. London: Tavistock.
- Jasiński, Mikołaj. 2004. *Problem jednostek niedostępnych w badaniach statystycznych*. W: T. Dziedzic (red.). *Metodologia badań turystów zagranicznych*. Warszawa: Instytut Turystyki.

- Jasiński, Mikołaj, Tomasz Zajac, Marek Styczeń i Albert Izdebski. 2012. *Wykorzystanie danych administracyjnych Zakładu Ubezpieczeń Społecznych i uczelni do monitorowania losów absolwentów uczelni wyższych*. „Zarządzanie Publiczne” 3(19): 49–60.
- Knudsen, Lisbeth. 1993. *Fertility Trends in Denmark in the 1980s. A Register Based Socio-demographic Analysis of Fertility Trends*. Copenhagen: Danmarks Statistik.
- Korpi, Helena. 1989. *Main Type of Activity and Occupational Status in the 1985 Census: Register-based Parallel Data*. Studies 152. Helsinki: Central Statistical Office of Finland.
- Lissowski, Grzegorz. 1971. *Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych*. W: K. Szaniawski (red.). *Metody matematyczne a socjologia*. Wrocław: Ossolineum.
- Lissowski, Grzegorz, Jacek Haman i Mikołaj Jasiński. 2011. *Podstawy statystyki dla socjologów t. 1. Opis statystyczny*. Warszawa: Wydawnictwo Naukowe Scholar.
- Longva, Svein, Ib Thomsen i Paul Inge Severeide. 1998. *Reducing Costs of Censuses in Norway through Use of Administrative Registers*. Reprint from „International Statistical Review” 66(2): 223–234.
- Pohoski, Michał i Marek Styczeń. 2000. *Dynamika zatrudnienia i zarobków w okresie transformacji ustrojowej w Polsce*. Warszawa: Instytut Socjologii UW (<http://stat.is.uw.edu.pl/wp-content/uploads/2010/08/mmbad3.pdf>).
- Smith, George, Michael Noble, Chelsie Anttila, Leicester Gill, Asghar Zaidi, Gemma Wright, Chris Dibben i Helen Barnes. 2004. *The Value of Linked Administrative Records for Longitudinal Analysis*. Social Disadvantage Research Centre, Department of Social Policy and Social Work, University of Oxford.
- Starr, Paul. 1983. *The Sociology of Official Statistics*. W: W. Alonso i P. Starr (red.). *The Politics of Numbers*. New York: Russel Sage Foundation, s. 7–58.
- Sulek, Antoni. 2002. *Socjolog wobec danych urzędowych. Przyczynek do metodologii badań ankietowych*. W: A. Sulek. *Ogród metodologii socjologicznej*. Warszawa. Wydawnictwo Naukowe Scholar.
- UNECE (United Nations Economic Commission for Europe). 2007. *Register-based statistics in the Nordic countries*. New York, Geneva: United Nations.
- USOS: <http://usos.edu.pl/inne-projekty>
- USOSownia: <http://usosownia.uw.edu.pl/start>
- Wallgren, Anders i Britt Wallgren. 2007. *Register-based Statistics. Administrative Data for Statistical Purposes*. West Sussex: John Wiley & Sons.
- Zajac, Tomasz. 2010. *Wykorzystanie danych z Uniwersyteckiego Systemu Obsługi Studiów oraz systemu Internetowej Rejestracji Kandydatów w procesie ewaluacji jakości kształcenia*. W: W. Przybylski i in. (red.). *Ewaluacja jakości dydaktyki w szkolnictwie wyższym. Metody. Narzędzia. Dobre praktyki*. Kraków: Wyższa Szkoła Europejska im. ks. Józefa Tischnera.
- Zajac, Tomasz. 2011. *Jak kandydaci starają się dostać na studia? Analiza strategii kandydatów na wybrane kierunki studiów na Uniwersytecie Warszawskim*. „Decyzje” (16): 73–104.

Zając, Tomasz. 2012. *(Nie)mobilność pionowa studentów UW. Analiza efektów wdrażania Procesu Bolońskiego*. „Zarządzanie Publiczne” 3 (19): 73–87.

Zając, Tomasz. 2013. *Wartość prognostyczna wyników rekrutacyjnych z 2012 roku. Studia pierwszego stopnia i jednolite magisterskie*. Warszawa: Uniwersytet Warszawski.

Accurate, Reliable and Cheap. Register-based Analysis as an Opportunity for Social Research in Poland

Summary

A social research project concerning various social issues (including education, labour market, demography, migration, epidemiology) can be conducted with register-based analysis. The knowledge about the potential and limits of this approach allows to reconcile the seemingly contradictory features of the study of complex social phenomena: effectiveness and low cost. Register-based analysis can solve a number of methodological problems, such as a decreasing response rate and a low reliability of data.

The paper presents the methodology of register-based analysis. The experiences gathered during the research project conducted by the authors allowed to improve the above mentioned methodology. The methodological reflections are complemented with the results of research on educational and professional careers of the University of Warsaw graduates, based on the registers of UW and Social Insurance Institution. In addition, the authors present selected technological, administrative, and organisational problems that need to be answered in order to conduct register-based research. The methodological conclusions of the paper develop the work of Scandinavian researchers who are the most experienced scientists in the field of register-based analysis.

Keywords: statistical methodology; research on graduates; labour market; administrative register; Social Insurance Institution.