PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 2 – 2012

WOJCIECH ROSZKA

SYSTEM STATYSTYKI PUBLICZNEJ OPARTY NA ZINTEGROWANYCH ŹRÓDŁACH DANYCH

1. WSTĘP

Wspieranie rozwijającej się gospodarki aktualną, rzetelną i wielowymiarową informacją jest rosnącym wyzwaniem dla służb statystycznych. Postulaty udostępniania wiarygodnych szacunków na niskim poziomie agregacji przestrzennej oznaczają, w dotychczasowym podejściu do badań statystycznych, zwiększanie liczebności próby, co podnosi koszty i czas ich przeprowadzenia. Te trudności, a także zwiększone obciążenie respondentów mogą uniemożliwiać spełnienie żądań odbiorców informacji statystycznych.

Rozwiązaniem problemów kosztowności i czasochłonności badań wydaje się być wykorzystane rejestrów administracyjnych w systemie statystyki publicznej. Zawierają one informacje o bardzo dużej liczbie jednostek, opisują bardzo wiele sfer życia i aktywności ludności, a organy statystyki publicznej mogą je pozyskać od dysponujących nimi instytucji i urzędów. Dodatkowo, dynamicznie rozwijająca się informatyzacja urzędów i wzrost mocy obliczeniowej komputerów powodują, że przesyłanie i przetwarzanie danych z rejestrów jest nie tylko możliwe, ale również stosunkowo szybkie i tanie. Ponadto podmioty administracji państwowej gromadzące dane do rejestrów są gwarantem rzetelności.

W Narodowym Spisie Powszechnym 2011 (a także Państwowym Spisie Rolnym 2010) wykorzystano informacje z dużej liczby źródeł administracyjnych[1]. Poprzez efektywne wykorzystanie rejestrów, zredukowano koszty spisów, obciążenia społeczne związane z pozyskiwaniem danych, poprawiono bezpieczeństwo przechowywanych informacji jednostkowych, a także zwiększono ich wiarygodność i spójność. Dane pochodzące z wielu źródeł umożliwiają oprócz tego wysokie pokrycie informacyjne dotyczące badanych zagadnień spisowych. Rejestry jako bezpośrednie źródła danych zostały wykorzystane w Polsce po raz pierwszy, tworząc zalążek systemu statystycznego opartego na rejestrach (Dygaszewicz 2012; Janczur-Knapek 2012).

Wieloźródłowość baz danych niesie ze sobą jednak pewne trudności, takie jak brak zmiennych mogących służyć jako klucz połączeniowy (lub błędy w takich zmiennych), różne definicje zmiennych zawartych w różnych rejestrach (nawet jeżeli zmienna

[1] W PSR 2010 i NSP 2011 wykorzystano informacje pochodzące z 27 rejestrów należących do 15 gestorów, jak również wykorzystano 3 zbiory danych poza administracyjnych (Janczur-Knapek, 2012).

nazywa się tak samo), różne sposoby wypełniania tych rejestrów, jak i istnienie zduplikowanych rekordów. Stanowi to duże wyzwanie dla badaczy.

Artykuł opisuje nowatorskie podejście do statystyki publicznej opartej na zintegrowanych administracyjnych źródłach danych. Przedstawia zalety połączonych źródeł informacji w stosunku do tradycyjnych systemów statystycznych opartych o badania reprezentacyjne oraz wady związane z różnym pochodzeniem zbiorów. Przedstawiono także koncepcję systemu statystycznego opartego na zintegrowanych źródłach danych oraz omówiono metody stochastyczne stosowane w procesie integracji.

2. REJESTRY ADMINISTRACYJNE W SŁUŻBIE STATYSTYKI PUBLICZNEJ

Informacja w systemie statystyki publicznej pozyskiwana jest drogą badań statystycznych, zarówno próbkowych, jak i pełnych. W „klasycznym" podejściu do badania zjawisk społeczno-ekonomicznych punktem wyjścia są potrzeby informacyjne odbiorców. Dążąc do ich zaspokojenia, służby statystyczne projektują badania obejmujące różne zagadnienia, wśród których można wymienić: aktywność ekonomiczną ludności, budżety gospodarstw domowych, czy działalność przedsiębiorstw. W kolejnym etapie następuje utworzenie narzędzia badawczego (kwestionariusza), sporządzany jest operat losowania[2] i na jego podstawie losuje się próbę. Następnie prowadzona jest obserwacja statystyczna. Zebrany materiał statystyczny poddany zostaje kontroli i po odpowiednim przetworzeniu (imputacji braków danych, przeważeniu, edycji danych itp.) jest podstawą szacunków, których wyniki są publikowane w formie tabel i wykresów statystycznych. Publikacje te służą wielu różnym celom, wśród których można wymienić wspomaganie organów rządowych i samorządowych w formułowaniu strategii rozwoju oraz prowadzenia polityki społecznej i gospodarczej.

Problemem w „klasycznym" podejściu jest fakt, że dla różnych celów przeprowadzane są oddzielne badania, przez co opracowania statystyczne obejmują pojedyncze tematy. Uniemożliwia to wielowymiarowe szacunki obejmujące różne zagadnienia społeczno-gospodarcze. Dodatkowo, ograniczenia budżetowe[3] powodują, że liczebność próby w badaniu jest zwykle zbyt niska, by szacunki mogły być dokonywane dla małych jednostek terytorialnych. Może to powodować brak zaspokojenia potrzeb informacyjnych samorządów (np. powiatów) dotyczących szczegółowej informacji o kształtowaniu się zjawisk na ich terytorium. Jednocześnie niewielka próba powoduje trudności w wykryciu i badaniu zjawisk rzadkich w skali kraju, które w skali małej domeny mogą być dużym problemem.

Uwzględnienie potrzeb wynikających z globalizacji gospodarki wymaga połączenia informacji z różnych dziedzin. Badania statystyczne obejmujące szeroki zakres

[2] Do tworzenia operatów wykorzystywane są również rejestry administracyjne.

[3] Badanie Aktywności Ekonomicznej Ludności w 2012 roku obejmujące około 54,7 tys. gospodarstw domowych kosztowało około 41 mln złotych, zaś Badanie Budżetów Gospodarstw Domowych w tym samym okresie aż 58 mln złotych (Program Badań Statystycznych Statystyki Publicznej na 2012 rok, 2011).

merytoryczny analizowanych zagadnień są jednak bardzo kosztowne. Ich przeprowadzenie wiąże się jednocześnie z bardzo dużym obciążeniem respondentów (van der Laan, 2000) i wynikającym z tego wzrostem liczby braków odpowiedzi i odmów wypełnienia kwestionariusza[4], nawet przy zastosowaniu nowoczesnych, obciążających respondentów w mniejszym stopniu, metod zbierania informacji (CATI[5], CAWI[6] itp.). Koszty zbierania informacji z długich kwestionariuszy mogłyby również prowadzić do zmniejszenia próby dodatkowo utrudniając wnioskowanie dla małych domen.

Informacja na niskim poziomie agregacji przestrzennej dostępna jest z badań pełnych. W badaniu takim pomiarem objęte są wszystkie jednostki należące do populacji docelowej. Ze względu jednak na zasięg takiego badania, jest ono dużym wyzwaniem zarówno finansowym, jak i organizacyjnym. W przypadku spisu powszechnego, nawet bogate kraje nie mogą sobie pozwolić na powtarzanie go częściej niż raz na kilka lat. W okresach międzyspisowych powstaje luka informacyjna, której badania reprezentacyjne nie są w stanie wypełnić. Zapewnienie precyzyjnej informacji dla małych domen jest wyzwaniem dla organów statystyki publicznej pod względem metodologicznym (np. związane z zastosowaniem metod statystyki małych obszarów).

Należy wyraźnie podkreślić, że wykorzystanie rejestrów administracyjnych może zapewnić uogólnianie wyników na niskim poziomie agregacji z dużą częstotliwością. W art. 13 ust. 1 ustawy z dnia 29 czerwca 1995 roku o statystyce publicznej ustawodawca nakazuje organom administracji rządowej i samorządowej przekazywanie danych administracyjnych służbom statystyki publicznej w terminach i formie każdorazowo wymienianej w programie badań statystycznych statystyki publicznej. Zbiory te opisują jednak pojedyncze zagadnienia, takie jak: bezrobocie rejestrowane, ruch naturalny i wędrówkowy ludności, czy działalność podmiotów gospodarczych, nie dając możliwości dokonywania wielowymiarowych szacunków obejmujących różnorodne relacje i zależności w funkcjonowaniu społeczeństwa, gospodarki i państwa jako całości. Dodatkowo definicje cech zawartych w rejestrach mogą się różnić od przyjętych w systemie statystyki publicznej. Rejestry administracyjne, z definicji, stworzone są do wypełniania zadań publicznych (Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne, Dz.U. Nr 64, poz.565, z późn. zm.), nie zaś bezpośrednio do celów statystycznych.

Bezpośrednie wykorzystanie rejestrów w statystyce publicznej nie może nastąpić w sposób automatyczny. Wynika to z odrębności systemów statystycznych i administracyjnych (por. tabela 1).

Informacje zawarte w rejestrach służą jako podstawa w podejmowaniu decyzji administracyjnych, które wpływają na funkcjonowanie jednostek. Natomiast informacje

[4] Przytoczyć tu można stale zmniejszający się poziom realizacji próby w badaniu Polski Generalny Sondaż Społeczny. W pierwszej edycji badania, w 1992 roku, wynosił 82,4%. W następnych latach ulegał stałemu spadkowi, by w 2008 roku wynosić już zaledwie 51,8% (Cichomski *et al.* 2009).

[5] *Computer-Assisted Telephone Interview* – wywiad wspomagany telefonicznie.

[6] *Computer-Assisted Web Interview* – wspomagany komputerowo wywiad przy pomocy strony www.

Tabela 1.

System administracyjny a system statystyczny

Charakterystyka		System administracyjny	System statystyczny
1. Cel powstania		Podejmowanie decyzji administracyjnych	Dokonywanie szacunków i analiz
2. Zbiorowość		Wszystkie podmioty (jednostki) prawnie podległe danemu gestorowi	Wszystkie podmioty (jednostki) objęte badaniem statystycznym
3. Jednostka		Pojedynczy element zbiorowości; pobierane dane niezbędne są do podejmowania decyzji administracyjnych	Pojedynczy element zbiorowości; pozyskiwane informacje są podstawą do szacowania statystyk dotyczących populacji lub jej podgrup
4. Cecha	**Definicja**	Wynika z aktów prawnych – mogą być odrębne dla różnych rejestrów	Wynika z ustaleń organizacji międzynarodowych – porównywalne dla wszystkich badań.
	Warianty	Nie muszą być zestandaryzowane	Zharmonizowane i porównywalne.
5. Błędy		Błędy nielosowe. Brak kontroli statystycznej	Błędy losowe i nielosowe. Kontrola statystyczna
6. Użyteczność		Dobre źródło statystyk małych obszarów	Jakość i możliwości szczegółowej analizy ograniczone wielkością próby
7. Terminowość i punktualność		Zróżnicowane w zależności od źródła. Niektóre bardzo aktualne, inne mniej terminowej niż badania statystyczne	Często mają charakter retrospektywny
8. Dostępność i przejrzystość		Wpływ uregulowań prawnych. Możliwe bariery techniczne i instytucjonalne	Bezpośrednia kontrola urzędu statystycznego
9. Porównywalność		Zależy od zmian w czasie definicji administracyjnych	Bezpośrednia kontrola urzędu statystycznego

Źródło: opracowanie własne na podstawie (Penneck, 2007)

zawarte w systemach statystycznych służą do analiz, na podstawie których formułowane są wnioski o całej populacji.

W systemach administracyjnych jednostka jest przedmiotem decyzji i działań poszczególnych organów wykonawczych, a baza danych służy do pobrania informacji na temat określonego podmiotu. W systemach statystycznych jednostka jest traktowana raczej jako część zbiorowości, dla której tworzone są pewne informacje agregatowe – jednostka nie leży więc w centrum zainteresowania (wyłączając badania monograficzne).

Definicja cechy w systemie administracyjnym wynika z aktów prawnych i może być różna dla różnych rejestrów. Warianty cechy nie muszą być spójne, ponieważ zwykle system administracyjny nie jest podstawą tworzenia zestawień statystycznych. W systemie statystycznym definicje cech, podobnie jak warianty, są spójne dla wszystkich badań (często wynika to z przyjętych ustaleń organizacji międzynarodowych).

Z punktu widzenia jakości danych, wszystkie dane w rejestrach muszą być zgodne i nie mogą zawierać błędów, jednak nie ma konieczności by zapis danej kategorii był taki sam w każdym rekordzie (np. kod pocztowy pisany z myślnikiem lub bez, pełny zapis nazw ulic lub skrócony itp.). Występujące nieścisłości (np. brak numeru PESEL, czy NIP) nie mają charakteru losowego i zazwyczaj wynikają z awarii systemów kon-

troli (które często występują w konkretnej jednostce administracyjnej). W badaniach statystycznych poszczególne warianty cech muszą być ujednolicone, by możliwe było tworzenie spójnych komunikatów. Prowadzony również w badaniach statystycznych rachunek błędów powoduje, że nieścisłości w wynikach są kontrolowane i podejmowane są działania w celu ich redukcji. Różnice w podejściu do gromadzenia danych administracyjnych i statystycznych mogą powodować rozbieżności w publikowanych komunikatach.

Wysokie pokrycie rejestrów administracyjnych może stanowić podstawę do tworzenia statystyk dla małych obszarów (również jako źródło pomocnicze dla estymacji pośredniej). W przypadku systemów statystycznych, możliwości agregacji szacunków są w dużej mierze ograniczone przez wielkość próby (pokrycie rzadko przekracza 1% populacji). Również terminowość rejestrów może być większa od badań statystycznych (mają zwykle charakter retrospektywny). Duża część zbiorów będących w posiadaniu organów publicznych jest uzupełniana na bieżąco i komunikaty tworzone na ich podstawie mogłyby być publikowane z częstotliwością miesięczną (a nawet i większą).

Dostępność i porównywalność danych publikowanych przez organy statystyki publicznej jest ściśle kontrolowana, co wynika w dużej mierze z uregulowań organizacji międzynarodowych. W przypadku systemów administracyjnych, duży wpływ mają regulacje prawne i zmiany definicji administracyjnych w czasie. Przystosowanie poszczególnych rejestrów do konkretnych celów publicznych może tworzyć również problemy techniczne i instytucjonalne – architektura baz i hurtowni danych może być bardzo odmienna, a zapisy prawne, zwłaszcza dotyczące ochrony danych osobowych, mogą zniechęcać instytucje do udostępniania swoich repozytoriów.

Należy również zaznaczyć, że wielu gestorów rejestrów posiada własne służby statystyczne (np. Zakład Ubezpieczeń Społecznych, Narodowy Bank Polski) tworzące sprawozdania na podstawie danych administracyjnych. Publikowane na ich podstawie informacje są przeznaczone jednak na potrzeby tych instytucji i nie uwzględniają potrzeb innych odbiorców.

Natomiast system statystyki oparty na zintegrowanych źródłach tworzony jest poprzez łączenie baz danych w taki sposób, by możliwa była łączna obserwacja zmiennych ze wszystkich integrowanych baz danych dla wszystkich jednostek.

Wallgren i Wallgren (2007) jako podstawowe zalety systemu statystyki opartego na zintegrowanych źródła wymieniają przede wszystkim: niskie koszty i szybkość pozyskania danych, wysoki stopień pokrycia, jak i bogatą zawartość informacyjną rejestrów (por. tabela 2). Przeciwstawiają im jednak pewne wady, wśród których wyszczególniają: niższą niż w tradycyjnych badaniach jakość danych, opóźnienia w przekazywaniu danych od gestorów wynikające często z ograniczeń prawnych), brak kontroli nad poprawnością statystyczną na etapie zbierania informacji, a także, a wręcz przede wszystkim, niebezpieczeństwo ujawnienia informacji osobowych, co mogłoby doprowadzić do utraty zaufania społecznego dla wykorzystania rejestrów w sprawozdawczości statystycznej.

Tabela 2.

Zalety i wady wykorzystania rejestrów administracyjnych w statystyce publicznej

Zalety	Wady
Niskie koszty pozyskania danych	Jakość wyników jest gorsza niż w badaniach tradycyjnych
Szybkość pozyskiwania danych	Opóźnienia w przekazywaniu rejestrów od gestorów do organów statystyki publicznej
Wysokie pokrycie informacyjne	Różne definicje zmiennych, problemy z kompatybilnością danych, brak kontroli nad poprawnością statystyczną
Pokrycie całej populacji	Niebezpieczeństwo ujawnienia wrażliwych danych osobowych

Źródło: opracowanie własne na podstawie (Wallgren, Wallgren, 2007).

3. ISTOTA INTEGRACJI ADMINISTRACYJNYCH BAZ DANYCH NA POTRZEBY SPRAWOZDAWCZOŚCI STATYSTYCZNEJ

Pionierem w wykorzystaniu źródeł administracyjnych w systemie statystyki publicznej były Finlandia (Statistics Finland, 2004) oraz Norwegia (Tonder, 2008). Już w spisie w 1970 wykorzystano w tych krajach rejestr ludności, stopniowo wprowadzając w kolejnych falach spisowych kolejne rejestry. W 1981 roku do tych krajów dołączyła Dania (Borchsenius, 2000), która od razu wprowadziła wszystkie dostępne rejestry w system spisowy (Finlandia wprowadziła spis w pełni oparty o rejestry dopiero w 1990, a Norwegia w 2011). Innymi krajami wykorzystującymi administracyjne źródła danych w systemie spisowym są m.in. Austria (Statistics Austria, 2008), Australia (Ralphs, Tutton, 2011), Holandia (Nordholt, 2004), Izrael (Kamen, 2005), Kanada (Ballano, 2009), Nowa Zelandia (Bycroft, 2011), Stany Zjednoczone (Prevost, Leggieri, 1999), Szwajcaria (Swiss Federal Statistical Office, 2008) oraz Szwecja (Bruhn, 2001; Wallgren, Wallgren, 2007).

Należy zauważyć, że we wszystkich tych krajach, rejestry administracyjne zostały wykorzystane wyłącznie w spisach powszechnych. Problemy z bieżącym dostępem do zbiorów (głównie prawne oraz techniczne) powodują, że możliwe jest tylko ich okresowe wykorzystanie w sprawozdawczości.

Rejestry przeznaczone do integracji mogą się od siebie różnić, co wynika m.in. z odmienności przechowywania danych (różne architektury baz danych, por. Dygaszewicz, 2010), różnych celów, dla których poszczególne rejestry są tworzone, niezgodności momentów referencyjnych (dnia, na który rejestr jest aktualny), czy odmienności w definicjach zmiennych. Dodatkowo, populacje, definicje, czy warianty poszczególnych cech mogą odbiegać od przyjętych w systemie statystyki publicznej. Istnieje zatem potrzeba przetworzenia rejestrów administracyjnych w taki sposób, by odpowiadały potrzebom badań statystycznych i mogły zostać włączone w system badań statystycznych.

Gill (2001) oszacował, że czas potrzebny na przeprowadzenie całego procesu integracji można podzielić w następujący sposób:

– 75% czasu pracy – przygotowanie zbiorów,
– 5% czasu pracy – przeprowadzenie łączenia rekordów w bazach,
– 20% czasu pracy – sprawdzanie poprawności wyników.

Przygotowanie zbiorów do integracji oraz weryfikacja poprawności połączenia to praca w dużej mierze manualna, wymagająca odpowiedniej wiedzy dotyczącej zarówno algorytmu integracji, jak również dziedzin, z których integrowane bazy zawierają informacje (np. definicje zmiennych, specyfikę populacji itp.). Przeprowadzanie łączenia zwykle wykonywane jest za pomocą odpowiednich programów komputerowych. Stąd duża dysproporcja między czasem potrzebnym na przeprowadzenie poszczególnych etapów łączenia.

Pierwszym krokiem integracji jest zebranie informacji na temat danych źródłowych (por. rysunek 2). Może to zostać osiągnięte poprzez zapoznanie się z dostępną dokumentacją, a także poprzez kontakt z gestorami poszczególnych zbiorów danych, niekoniecznie mając dostęp do plików. Informacje zdobyte w ten sposób umożliwiają oszacowanie na ile przeznaczone do łączenia zbiory spełniają wymogi formalne, merytoryczne i jakościowe procesu integracji.

Nowozelandzki Urząd Statystyczny (Data Integration Manual, 2006), korzystając z wieloletniego doświadczenia w integracji, sporządził listę trzech punktów składających się na wstępne zbieranie informacji o zbiorach:

1. Zestawienie istniejącej wiedzy wewnętrznej

W przypadku, gdy te same lub podobne zbiory były używane wcześniej, wiedza i doświadczenie, w postaci dokumentów lub procesów, pracowników biorących udział przy ich łączeniu może być bardzo pomocna. Wraz ze wzrostem liczby integrowanych zbiorów, wypracowanie wewnętrznego systemu gromadzenia wiedzy i dzielenia się nią prowadzi do zwiększenia efektywności pracy nad kolejnymi projektami.

2. Przegląd ogólnie dostępnych informacji

Bardzo często potrzebne i użyteczne informacje na temat integrowanych zbiorów znajdują się na stronach internetowych poszczególnych gestorów. W wielu przypadkach również dostępne są też bardziej szczegółowe informacje w postaci tzw. „często zadawanych pytań" (*Frequently Asked Questions, FAQ*) lub udostępnionych słowników danych[7].

3. Spotkanie z dostawcami danych

Spotkanie z dostawcami danych umożliwia efektywny transfer wiedzy od ludzi, którzy z określonymi repozytoriami pracują na codzień. Podczas takich spotkań zadawane pytania oraz przekazywane dokumenty bardzo często w szybki sposób rozwiewają wątpliwości, które mogły się wyłonić podczas pracy nad danymi.

[7] Centralny element systemu zarządzania bazą danych, w którym przechowuje się m.in. opisy relacji i perspektyw, deklaracje kluczy głównych, grup użytkowników i uprawnień, informacje o indeksach, plikach i ich strukturach (Adamczewski, 2005).

Identyfikacja populacji docelowej, a także należących do niej jednostek statystycznych, w obu zbiorach, jest ważnym punktem zbierania informacji o danych źródłowych, gdyż może się zdarzyć, że pokrycie populacji jest różne w każdej z baz. Każde repozytorium danych odnosi się do pewnej populacji docelowej, będącej populacją teoretyczną, o której informacje baza powinna zawierać. Zaś poszczególne rekordy danych zawierają informacje o populacji rzeczywistej. Celem integracji jest stworzenie nowego, bogatszego, zbioru danych, który również będzie odnosił się do pewnej populacji docelowej, którą również należy zdefiniować.

Należy również zweryfikować zgodność typów jednostek w obu zbiorach (np. czy bazy zawierają dane o obywatelach, gospodarstwach domowych, czy przedsiębiorstwach) oraz wykonanie wszelkich potrzebnych transformacji danych. Jednostki z baz źródłowych mogą się różnić od jednostek w zintegrowanym zbiorze, jak również różnice mogą wystąpić pomiędzy jednostkami w obu łączonych bazach. Np. w zmiennej „stan cywilny" w jednej bazie może być kategoria „żonaty/zamężna", zaś w drugiej obie te warianty mogą występować oddzielnie. Przy dużej liczbie zmiennych, przekodowanie wszystkich tak, by zawierały dokładnie te same może okazać się czasochłonnym, ale koniecznym procesem.

Ostatnią częścią etapu zbierania informacji o danych źródłowych jest sporządzenie formularza zawierającego „dane o danych" – tzw. metadanych („informacji o danych") (Data Integration Manual, 2006). Jest on ważnym elementem, zapewniającym odpowiednią jakość przyszłych wyników, przy czym jakość rozumiana jest sześciowymiarowo, poprzez: przydatność, dokładność, terminowość, dostępność, interpretowalność oraz spójność. Wymienione wymiary jakości mogą być użyte jako kryterium oceny wszystkich szacunków uzyskanych za pomocą zintegrowanych baz danych, odnosząc się zarówno do baz zintegrowanych, jak i źródłowych.

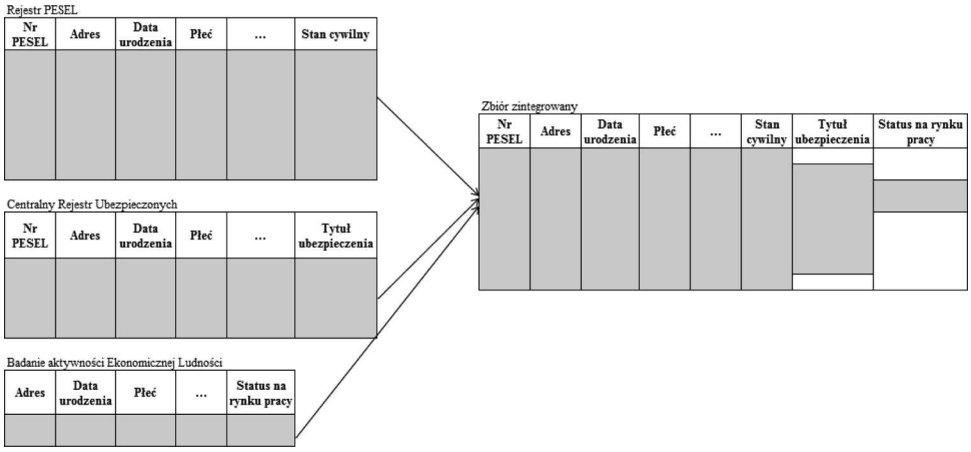
Ważnym zagadnieniem jest również określenie regulacji prawnych umożliwiających wykorzystanie rejestrów w statystyce publicznej. Art. 13 pkt. 4 ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej nakłada na gestorów rejestrów administracyjnych obowiązek nieodpłatnego przekazywania danych, w tym całych zbiorów, w zakresie, formie i terminach określanych w programie badań statystycznych statystyki publicznej. Dane te jednak służą do tworzenia komunikatów „jednowymiarowych", nie wykorzystujących zalet integracji. Dopiero ustawa o narodowym spisie powszechnym ludności i mieszkań w 2011 roku (art. 8) nakazuje utworzenie tzw. Bazy Danych NSP 2011 złożonej ze zintegrowanych administracyjnych źródeł informacji. Wykorzystanie połączonych rejestrów w okresach międzyspisowych wymaga dodatkowych aktów prawnych.

Efektem etapu badania infrastruktury rejestrów jest wiedza umożliwiająca dokonanie integracji rejestrów, edycji danych i w końcowym etapie tworzenie spójnych i rzetelnych komunikatów statystycznych.

Następnym etapem jest łączenie rejestrów na podstawie unikatowych identyfikatorów jednostek. W źródłach administracyjnych występują zwykle zmienne jak, np. numer PESEL dla osób lub REGON dla podmiotów gospodarczych. W badaniach

reprezentacyjnych jednostkę można zidentyfikować na podstawie zestawu zmiennych, takich jak wiek, płeć i adres zamieszkania[8]. Są to tzw. unikalne klucze połączeniowe. Na ich podstawie możliwe jest jednoznaczne (deterministyczne) wskazanie, które rekordy w łączonych bazach odnoszą się do tej samej jednostki.

Jak pokazano na rysunku 1, do rejestru PESEL, zawierającego m.in. dane demograficzne dodano, stosując numer PESEL jako unikalny klucz, informacje z Centralnego Rejestru Ubezpieczonych (CRU). W ten sposób uzyskano łączną informację na temat stanu cywilnego (zmienna obserwowana w zbiorze PESEL) i tytułu ubezpieczenia społecznego (zmienna obserwowana w CRU). Następnie, wykorzystując kombinację wartości zmiennych płeć, wiek oraz adres zamieszkania jako klucz, dodano informację o statusie na rynku pracy (obserwowana w BAEL). Zbiory CRU i BAEL charakteryzują się mniejszym pokryciem niż PESEL[9], stąd łączna obserwacja cech (kolor szary) możliwa jest tylko dla tych jednostek, które występowały w każdym ze zbiorów. Dla pozostałych występują braki danych (kolor biały). Zintegrowana w ten sposób baza może być bogatym, tanim (nie ma potrzeby pomiaru statystycznego – dane już istnieją) źródłem informacji społeczno-gospodarczych.



Rysunek 1. Integracja zbiorów danych pochodzących z różnych źródeł

Źródło: opracowanie własne

Efektem zastosowania łączenia deterministycznego jest surowa baza danych zawierająca wszystkie zmienne z łączonych baz. Są to zmienne niezharmonizowane, tj. o definicjach, wariantach, momentach referencyjnych, itp. pochodzących z baz wejściowych.

[8] Adres zamieszkania nie jest przedmiotem pomiaru w badaniach próbkowych, jednak jest on znany z operatu losowania.

[9] Np. do badania BAEL wylosowanych jest około 200 tys. jednostek z około 30-milionowej populacji, co daje pokrycie rzędu ok. 0,7%.

W przypadku, gdy zmienne kluczowe o unikatowych wartościach nie są dostępne, lub zawierają wspomniane wyżej błędy, niemożliwe jest wykorzystanie metod deterministycznych. W takich przypadkach możliwe jest stosowanie metod probabilistycznego łączenia rekordów (*probabilistic record linkage*). Idea tej metody sprowadza się do wyboru kilku zmiennych (nazywanych zmiennymi parującymi), które zawarte są w obu zbiorach i wykorzystanie ich zgodności, częściowej zgodności lub niezgodności do oszacowania prawdopodobieństwa, że poszczególne rekordy należą (lub nie) do tej samej jednostki.

W integrowanych bazach, ze względu na brak kompatybilności, często te same wartości zapisywane są w niejednolity sposób (np. adresy, imiona i nazwiska, nazwy własne itp.), które mogą wynikać zarówno z przyjętych przez gestorów baz różnych standardów zapisu lub wynikających z różnorodnego rodzajów błędów zapisu (np. ortograficznych, typograficznych, wynikających z niedoskonałości sprzętu i oprogramowania skanującego itp.). Za pomocą metodyki probabilistycznego łączenia rekordów można również zintegrować takie jednostki porównując pewne wartości w zmiennych występujących w obu bazach, które choć różnią się sposobem zapisu należą do tej samej obserwacji.

Najczęściej, w literaturze przedstawia się probabilistyczne łączenie rekordów jako proces kilkustopniowy. Pierwszym krokiem jest wybór zmiennych, na podstawie których przeprowadzone zostanie łączenie (tzw. zmienne parujące). W kolejnym kroku przygotowuje się bazy do procesu integracji poprzez usunięcie duplikatów oraz standaryzację wariantów cech parujących. Następnie dokonuje się operacji grupowania (nazywanej również blokowaniem) mającej na celu podziałintegrowanych baz na podzbiory, w których znajdą się jednostki w jakiś sposób do siebie podobne (np. mieszkańcy jednego powiatu lub przedstawiciele jednej gałęzi przemysłu). Grupowania dokonuje się z w celu optymalizacji algorytmu poprzez zredukowanie liczby połączeń. Następnie na podstawie pewnych algorytmów wykonuje się łączenie baz oraz sprawdzenie jego efektywności.

Efektem zastosowania probabilistycznego łączenia rekordów jest surowa baza danych uzupełniona o rekordy nieprzyłączone na etapie deterministycznego łączenia rekordów.

Połączenie zbiorów jest dopiero pierwszym krokiem do pełnej integracji. W celu uzyskania zbioru danych odpowiadającego celom statystycznym należy ujednolicić zawarte w nich informacje. Ten etap w literaturze nosi nazwę przetwarzania danych zintegrowanych (*micro-integration processing*). Wallgren, Wallgren (2007) oraz Linder (2004) wymieniają następujące etapy przetwarzania zintegrowanych rejestrów:

– kodowanie zmiennych – ujednolicanie wariantów cech,
– edycja braków danych – imputacja,
– wyrównywanie momentów lub okresów referencyjnych rejestrów – w celu zapewnienia możliwości porównań wyników (przykładem rozwiązania tego problemu na potrzeby NSP jest praca Bijaka, 2009),

– tworzenie jednostek pochodnych (np. gospodarstw domowych złożonych z osób mieszkających pod jednym adresem),
– tworzenie zmiennych pochodnych (np. utworzenie zmiennej „dochody całkowite" będącej sumą dochodów z różnych źródeł),
– porównywanie zmiennych z różnych źródeł w celu korekty błędów.

Etap przetwarzania danych zintegrowanych umożliwia utworzenie operacyjnej bazy mikrodanych[10], w której znajdują się cechy o ujednoliconym charakterze, zdefiniowane według norm przyjętych w statystyce publicznej oraz o zadowalającej jakości. Przetwarzanie danych zintegrowanych jest stałym procesem, który zapewnia spójność i wysoką jakość informacji statystycznej.

Bakker (2010) wymienia cztery podstawowe zalety przetworzonej, operacyjnej bazy danych:
– rzetelność i wiarygodność komunikatów statystycznych sporządzonych na podstawie zintegrowanych źródeł jest poprawiona (w porównaniu do bazy „surowej"),
– możliwa jest publikacja oszacowań na niskim, niedostępnym dla badań reprezentacyjnych, poziomie agregacji przestrzennej i merytorycznej,
– zmienne z różnych źródeł są połączone i możliwa jest ich łączna obserwacja,
– możliwe jest przeprowadzenie badań panelowych.

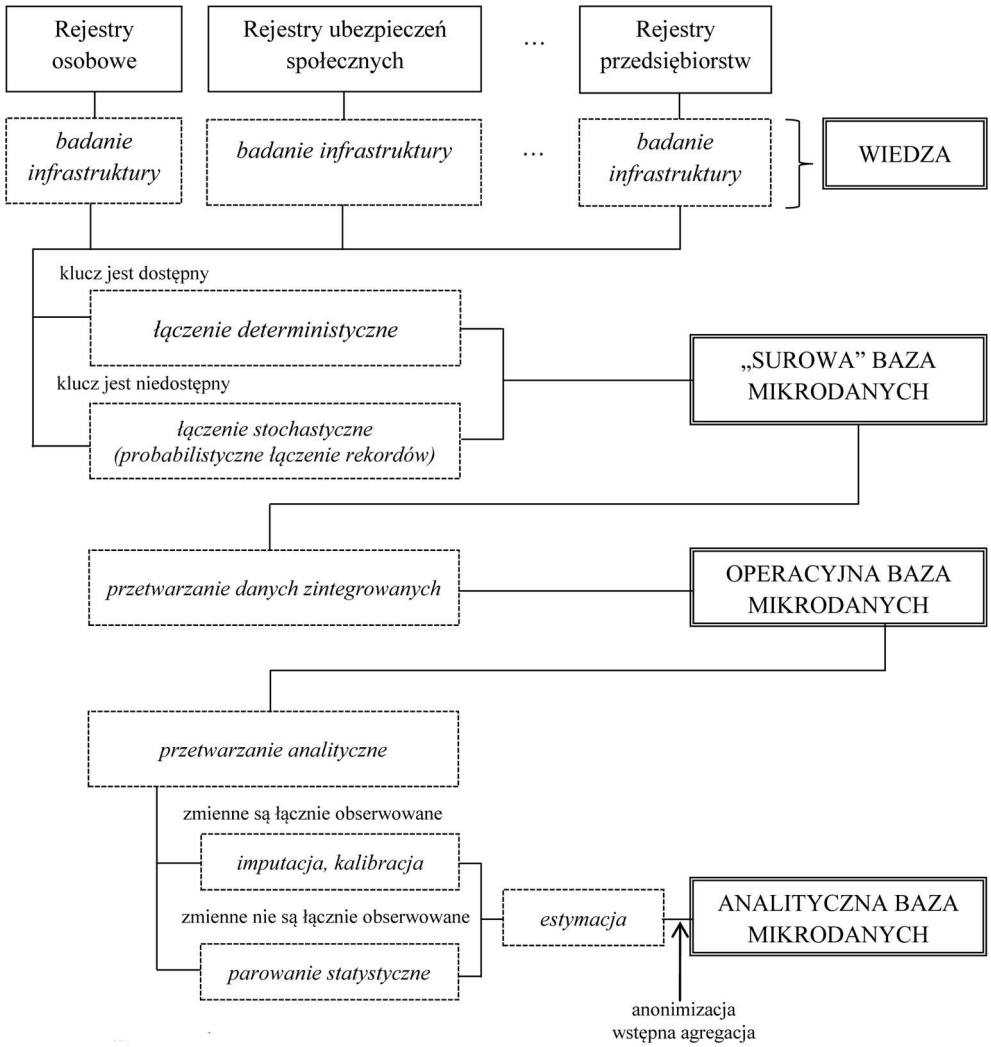
Integrowane zbiory danych charakteryzują się różnym stopniem pokrycia. Rejestry administracyjne zawierają informacje o bardzo dużej liczbie jednostek, natomiast badania reprezentacyjne charakteryzują się niewielkim pokryciem. Stąd też zintegrowana, operacyjna baza danych będzie zawierać pełną informację wyłącznie dla jednostek, które wystąpiły w każdym z integrowanych źródeł. Rekordy, które wystąpiły tylko w pojedynczych źródłach, dla cech dołączonych z innych baz będą charakteryzować się brakami danych.

Ostatnim etapem integracji danych jest proces przetwarzania analitycznego mikrodanych. Polega on na imputacji braków danych, kalibracji wag analitycznych oraz estymacji finalnych komunikatów statystycznych. Celem procesu jest zachowanie spójności numerycznej danych w sensie uzyskania takich samych wyników dla wszystkich oszacowań bez względu na źródło pochodzenia zmiennych w zintegrowanym zbiorze.

W zależności od źródła pochodzenia zmiennej, szacunki mogą charakteryzować rozbieżnościami nawet w skali całego kraju. Konsekwencje dywersyfikacji źródeł informacji mogą zostać zniwelowane poprzez różne metody analityczne, wśród których najczęściej wymienia się metody ważenia (wielokrotnego ważenia) oraz masowej imputacji (Kroese et al., 2000).

W metodzie wielokrotnego ważenia, w pierwszym etapie następuje wyodrębnienie podzbioru bazy operacyjnej bazy mikrodanych pochodzących z różnych źródeł. Następnie przyporządkowane są wagi początkowe. Później dostosowuje się wagi kalibracyjne uzyskane dla wszystkich podzbiorów danych w taki sposób, by uzyskane rezultaty były zgodne dla cech znajdujących w każdym z nich. Odbywa się to poprzez rekalibrację

[10] Noszącej również nazwę rejestru statystycznego (Wallgren, Wallgren, 2007).



Rysunek 2. Integracja administracyjnych źródeł danych

Źródło: opracowanie własne

wag w odniesieniu do rozkładów brzegowych zmiennych występujących w każdym podzbiorze. Zapewnia się w ten sposób spójność oszacowań przy jednoczesnym zachowaniu rzetelności oszacowań w sensie ich nieobciążoności (Zhang, 2012).

Alternatywą dla ważenia danych jest masowa imputacja. Zastosowanie metod imputacji w celu uzupełnia braków danych wartościami syntetycznymi zapewnia łączną obserwację wszystkich zmiennych w zintegrowanej bazie danych jednostkowych, jak również numeryczną spójność szacunków wszystkich cech (wszystkie wartości będą się sumować do liczebności ujednoliconej populacji generalnej). Wadą takiego podejścia jest fakt, że podstawione wartości są w dużej mierze sztuczne, nawet jeżeli

wynikają z dobrze dobranego modelu uzupełniania braków. W wyniku imputacji powstają jednostki nierzeczywiste (o nierzeczywistych wartościach cech), co z jednej strony powoduje, że maleje niebezpieczeństwo identyfikacji prawdziwych jednostek, ale z drugiej strony może prowadzić do obniżenia zgodności danych (np. 15-letni emeryt lub 40-letni przedszkolak). Dodatkowo, nawet jeżeli do imputacji stosuje się model skonstruowany w oparciu o wartości empiryczne (np. imputacja regresyjna), bardzo często na podstawie stosunkowo niewielkiej liczby wartości empirycznych, podstawia się wielką liczbę wartości teoretycznych (np. imputując aktywność ekonomiczną z BAEL – ok. 100 tys. rekordów – do rejestru PESEL – ok. 30 mln rekordów).

Może się zdarzyć, że nie wszystkie cechy są łącznie obserwowane lub łączna obserwacja cech zachodzi tylko dla niewielkiego podzbioru obserwacji uniemożliwiającego dokonywanie szacunków dla odpowiedniego przekroju danych. Proces estymacji może zostać w takim przypadku wsparty metodami parowania statystycznego (*statistical matching*). Parowanie statystyczne to grupa metod służących do integracji dwóch (lub więcej) źródeł danych (zwykle pochodzących z badań próbkowych) odnoszących się do tej samej populacji generalnej. Ponieważ prawdopodobieństwo wylosowania tej samej jednostki do dwóch różnych badań reprezentacyjnych jest bardzo małe (zbliżone do zera), zakłada się, że integrowane zbiory są rozłączne w sensie pokrycia. W każdym zbiorze (oznaczonymi jako A i B) znajduje się zwykle pewien wspólny wektor zmiennych (np. w badaniach dotyczących osób mogą być to zmienne demograficzne) o tych samych lub zbliżonych definicjach i wariantach. Nazywa się je zmiennymi wspólnymi (oznaczonymi jako X). Zbiór A zawiera wektor zmiennych obserwowanych wyłącznie w nim, oznaczy jako Y, natomiast zbiór B zawiera analogiczny wektor – Z (por. tabela 3). Celem parowania statystycznego jest analiza związków pomiędzy zmiennymi nieobserwowanymi łącznie w pojedynczym źródle.

Produktem integracji baz danych metodą parowania statystycznego są jednostki syntetyczne. Oznacza to, że w zintegrowanym zbiorze obserwacji podlegają podmioty niewystępujące w rzeczywistości. U źródeł koncepcji tworzenia zbiorowości utworzonych w sposób sztuczny leży założenie, że jednostki, które są do siebie podobne pod względem określonych cech (np. demograficznych, takich jak: wiek, płeć, miejsce zamieszkania, czy wykształcenie lub ekonomicznych, takich jak: aktywność ekonomiczna, źródło utrzymania, czy dochody) będą również podobne pod względem innych cech będących przedmiotem parowania.

Dzięki zastosowaniu metody parowania statystycznego możliwa jest łączna obserwacja cech nieobserwowanych wspólnie w żadnym ze źródeł. Umożliwia to dokonywanie analiz wielowymiarowych, jak np. badanie współzależności (wyznaczenie współczynników korelacji, czy też stworzenie tabel kontyngencji).

Metody parowania statystycznego stosuje się w sposób podobny do metody masowej imputacji, jednak kładzie się w nich nacisk, by, o ile to możliwe, dołączane

Tabela 3.

Dane wejściowe w parowaniu statystycznym

Zbiór A

$\mathbf{Y_1}$	...	$\mathbf{Y_Q}$	$\mathbf{X_1}$	...	$\mathbf{X_P}$
y^A_{11}	...	y^A_{1Q}	x^A_{11}	...	x^A_{1P}
...	...	...	...	...	...
y^A_{a1}	...	y^A_{aQ}	x^A_{a1}	...	x^A_{aP}
...	...	...	...	...	...
$y^A_{n_A1}$	...	$y^A_{n_AQ}$	$x^A_{n_A1}$	...	$x^A_{n_AP}$

Zbiór B

$\mathbf{X_1}$	...	$\mathbf{X_P}$	$\mathbf{Z_1}$	...	$\mathbf{Z_R}$
x^B_{11}	...	x^B_{1P}	z^B_{11}	...	z^B_{1R}
...	...	...	...	...	...
x^B_{b1}	...	x^B_{bP}	z^B_{b1}	...	z^B_{bR}
...	...	...	...	...	...
$x^B_{n_B1}$	...	$x^B_{n_BP}$	$z^B_{n_B1}$	...	$z^B_{n_BR}$

Źródło: opracowanie własne

wartości były wartościami empirycznymi[11]. Dołącza się również jedną zmienną jednocześnie dokładając starań, by model integracji zapewniał zgodność wielowymiarowych rozkładów w integrowanych zbiorach (Raessler, 2002).

Zintegrowane, w sposób deterministyczny lub (i) stochastyczny, dane z różnych źródeł, zapewniające wysokie pokrycie i łączną obserwację cech z różnych obszarów funkcjonowania społeczeństwa, gospodarki i państwa są również punktem wyjścia do tworzenia kompleksowych systemów statystycznych zapewniających wsparcie w tworzeniu symulacji i prognoz skuteczności działań organizacji państwowych (polityki podatkowej, opieki społecznej etc.) i prywatnych (inwestowanie w określone rejony, lokowanie produktów, zdobywanie klientów poprzez precyzyjnie przeprowadzane kampanie marketingowe). Takie wykorzystanie zintegrowanych źródeł nosi nazwę mikrosymulacji (*microsimulation*) i jest wykorzystywane m.in. w Europie (Atkinson *et al.* 1999), Kanadzie (Morrison 1998) oraz Australii (Kelly, 2003; Hardling et al., 2009).

[11] Np. poszukiwanie tzw. „statystycznych bliźniąt" (Bacher, 2002), czyli wyszukiwanie w jednym zbiorze rekordów najbardziej podobnych do tych w drugim zbiorze i ich łączenie (Raessler, 2002; Di Zio et al., 2006). Inną metodą jest tzw. wielokrotna imputacja (Raessler, 2002), gdzie dla jednej dołączanej wartości wyznacza się kilka wynikających z modelu lub najbardziej podobnych w innym zbiorze i albo dołącza się rekord wylosowany z tego zbioru „bliźniąt", albo np. imputuje się wartość średnią.

4. ZAKOŃCZENIE

Integracja danych administracyjnych i badań reprezentacyjnych, przy zachowaniu odpowiednich procedur i metod prowadzi do utworzenia zbioru o szerokim spektrum informacyjnym i wysokim pokryciu. Koszt pozyskania informacji ze zbiorów zintegrowanych jest dużo niższy niż w przypadku klasycznych badań statystycznych, nie występuje obciążenie respondentów, a zbiory charakteryzują się wysoką jakością.

Dokonywanie szacunków na podstawie zintegrowanych źródeł wiąże się z zastosowaniem nowatorskich metod statystycznych, przekonstruowania systemu statystycznego oraz zachowania tajemnicy statystycznej. Zachowanie określonych procedur i norm integracji zapewnia możliwość tworzenia wielowymiarowych analiz na niskim poziomie agregacji przestrzennej, co może przyczyniać się do podejmowania właściwych decyzji dotyczących rozwoju społeczno-gospodarczego małych jednostek terytorialnych poprzez precyzyjne i rzetelne zasilanie informacyjne.

Zastosowanie metod statystycznej integracji danych wspomaga system statystyczny oparty o administracyjne źródła danych poprzez łączenie rekordów pozbawionych klucza identyfikacyjnego oraz umożliwia łączną obserwację cech nieobserwowanych zarówno w żadnym pojedynczym źródle, jak i w zbiorze zintegrowanym.

Łączna obserwacja cech opisujących różne zagadnienia społeczno-gospodarcze umożliwia zastosowanie mikrosymulacji wspomagającej organy administracyjne w procesach decyzyjnych dotyczących prowadzenia określonych polityk w ujęciu przestrzennym.

Uniwersytet Ekonomiczny w Poznaniu

LITERATURA

[1] Adamczewski P., (2005), *Słownik informatyczny*, Wydawnictwo Helion, Gliwice.

[2] Atkinson A.B., Bourguignon F., O'Donoghue C., Sutherland H., Utili F., (1999), *Microsimulation and the formulation of policy: a case study of targeting in the European Union*, EUROMOD, Working Papers Series, Working Paper No. EM2/99.

[3] Bacher J., (2002), *Statistisches Matching - Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS*, ZA-Informationen, 51. Jg.

[4] Bakker B., (2010), *Micro-Integration: State of the art* (w:) *Draft Report of WP1. State of the art on statistical methodologies for data integration*, ESSnet on Data Integration, WP1/D1.32/2010JUN

[5] Ballano C., (2009), *A Census of Population Based on an Administrative Register*, Proceedings of Statistics Canada Symposium 2008, Data Collection: Challenges, Achievements and New Directions.

[6] Borchsenius L., (2000), *From a conventional to a register-based census of population*, INSEE-Eurostat seminar on the censuses after 2001.

[7] Bruhn A., (2001), *The next Population and Housing Census in Sweden is planned for 2005 – it will be totally register-based*, Symposium on Global Review of 200 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, Statistic Division, Department of Economic and Social Affairs, United Nations, New York.

[8] Bycroft Ch., (2011), *A register-based census: what is the potential for New Zealand*, Statistics New Zealand, Tatauranga Aotearoa, Wellington, New Zealand.

[9] Cichomski B. (kierownik programu), Jerzyński T., Zieliński M., (2009), *Polskie Generalne Sondaże Społeczne: skumulowany komputerowy zbiór danych 1992-2008*, Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa.
[10] *Data Integration Manual*, (2006), praca zbiorowa Statistics New Zealand, Wellington.
[11] D'Orazio M., Di Zio M., Scanu M., (2006), *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England.
[12] Dygaszewicz J., (2010), *Integracja rejestrów publicznych*, Główny Urząd Statystyczny, Warszawa.
[13] Dygaszewicz J., (2012), *Spisy powszechne jako źródło danych do analiz geoprzestrzennych*, Archiwum Fotogrametrii, Kartografii i Teledetekcji, Vol. 23.
[14] Gill L., (2001), *Methods for Automatic Record Matching and Linkage and their use in National Statistics*, National Statistics Methodological Series No 25, National Statistics, United Kingdom.
[15] Hardling A., Kelly S., Percival R., Keegan M., (2009), *Population Ageing and Government Age Pension Outlays*, ESRI International Collaboration Project, NATSEM, University of Canberras.
[16] Janczur-Knapek M., (2012), *Spisy powszechne PSR 2010 i NSP 2011 oraz systemy informacji geograficznej w statystyce publicznej*, referat wygłoszony na Kongresie Statystyki Polskiej, Poznań.
[17] Kamen C.S., (2005), *The 2008 Israel Integrated Census of Population and Housing, Basic conception and procedure*, State of Israel, Central Bureau of Statistics.
[18] Kelly S., (2003), *Australia's Microsimulation Model – Dynamod*.
[19] Kroese B., Renssen R.H., Trijssenaar M., (2000), *Weighting or imputation: constructing a consistent set of estimates based on data from different sources*, „Netherlands Official Statistics", vol. 15, Summer 2000, Special issue: *Integrating administrative registers and household surveys*, Statistics Netherlands, Voorburg/Heerlen.
[20] Linder F., (2004), *The use of administrative registers and sample surveys in the Dutch Census* of 2001 (w:) *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen
[21] Morrison R., (1998), *Overview of DYNACAN: a full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes*, http://www.actuaries.org/CTTEES_SOCSEC/Documents/dynacan.pdf
[22] Nordholdt E.S., (2004), *Introduction to the Dutch Virtual Census of 2001* (w:) *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen.
[23] Penneck S., (2007), *Using administrative data for statistical purposes*, Economic & Labour Market Review.
[24] Prevost R., Leggieri Ch., (1999), *Expansion of Administrative Records Uses at the Census Bureau: A Long-Range Research Plan*, U.S. Bureau of the Census, Washington D.C.
[25] Program badań statystycznych statystyki publicznej na 2012 rok, załącznik do rozporządzenia Rady Ministrów z dnia 22.07.2011 r. w sprawie programu badań statystycznych statystyki publicznej na rok 2012 (Dz. U. Nr 173, poz. 1030).
[26] Raessler S., (2002), *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA.
[27] Raessler S., (2004), *Data fusion: identification problems, validity, and multiple imputation*, Austrian Journal of Statistics 33(1-2).
[28] Ralphs M., Tutton P., (2011), *Beyond 2011: International models for census taking: current processes and future developments*, Beyond 2011 Project, Office for National Statistics.
[29] Statistics Austria, (2008), *Register-based census 2010 and census test 2006*, Joint UNECE/Eurostat meeting on population and housing censuses, Geneva.
[30] Statistics Finland, (2004), *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*, Tilastokeskus, Statistikcentralen, Statistics Finland, Helsinki.
[31] Swiss Federal Statistical Office, (2008), *The Swiss Census 2010: Moving towards a comprehensive system of household and person statistics*, Federal Statistical Office.

[32] Tonder J-K., (2008), *The register-based statistical system – preconditions and processes*, IAOS Conference, Shanghai.
[33] Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne (Dz.U. Nr 64, poz.565, z późn. zm.).
[34] Ustawa z dnia 4 marca 2010 roku o narodowym spisie powszechnym ludności i mieszkań w 2011 r. (Dz.U. z 2010 nr 47 poz. 277).
[35] Ustawa z dnia 29 czerwca 1995 roku o statystyce publicznej, tekst jednolity (Dz. U. 2012.591).
[36] van der Laan P., (2000), *Integrating administrative registers and household surveys*, „Netherlands Official Statistics", vol. 15, Summer 2000, Special issue*: Integrating administrative registers and household surveys*, Statistics Netherlands, Voorburg/Heerlen.
[37] Wallgren A., Wallgren B., (2007), *Register-based Statistics. Administrative Data for Statistical Purposes*, John Wiley and Sons Ltd.
[38] Zhang L-C., (2012), *Micro calibration for data integration*, referat wygłoszony na Kongresie Statystyki Polskiej, Poznań.

SYSTEM STATYSTYKI PUBLICZNEJ OPARTY NA ZINTEGROWANYCH ŹRÓDŁACH DANYCH

S t r e s z c z e n i e

Zwiększające się zapotrzebowanie na rzetelną i aktualną informację na możliwie niskim poziomie agregacji jest rosnącym wyzwaniem dla polskiej statystyki publicznej. Zastosowanie integracji danych z różnych, w tym administracyjnych, źródełumożliwia wykorzystanie informacji pełnej w sensie pokrycia oraz bogatej merytorycznie. Łączna obserwacja cech obserwowanych w oddzielnych zbiorach generuje efekt synergii zwiększający zasób wiedzy pochodzący z badań społeczno-ekonomicznych.

Słowa kluczowe: integracja danych, rejestry administracyjne, spisy powszechne

THE SYSTEM OF PUBLIC STATISTICS BASED ON INTEGRATED DATA SOURCES

A b s t r a c t

Increasing demand for reliable and current information at the lowest possible level of aggregation is a growing challenge for the Polish public statistics. Application of data integration of different, including administrative, sources enables the use of information in terms of full coverage and rich in substance. The joint observation of variables observed in separate data collections generate a synergy effect increasing amount of knowledge derived from socio-economic research.

Key words: data integration, administrative registers, censuses