POLSKA AKADEMIA NAUK
KOMITET STATYSTYKI I EKONOMETRII

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

NUMER SPECJALNY
1

2012

WARSZAWA 2012

WYDAWCA

Komitet Statystyki i Ekonometrii Polskiej Akademii Nauk

Wydanie publikacji dofinansowane przez Ministra Nauki i Szkolnictwa Wyższego

Strona WWW „Przegląd Statystyczny”:
http://www.przegladstatystyczny.pan.pl

Nakład: 250 egz.

PAN Warszawska Drukarnia Naukowa
00-056 Warszawa, ul. Śniadeckich 8
Tel./fax:+48 22 628 76 14

Z okazji jubileuszu 100-lecia
Polskiego Towarzystwa Statystycznego

Honorowy Patronat Prezydenta Rzeczypospolitej Polskiej
Bronisława Komorowskiego

Publikacja dofinansowana ze środków
Narodowego Banku Polskiego

NBP

Narodowy Bank Polski

OD REDAKCJI

Szanowni Państwo,

Oddajemy w Państwa ręce numer specjalny czasopisma *Przegląd Statystyczny*, poświęcony w całości publikacjom, których prezentacja miała miejsce podczas Kongresu Statystyki Polskiej, zorganizowanego w dniach 18-20 kwietnia 2012 r. w Poznaniu z okazji 100-lecia Polskiego Towarzystwa Statystycznego. Były to główne obchody trwającego przez cały rok szacownego Jubileuszu. Honorowy Patronat nad tym wyjątkowym wydarzeniem objął Prezydent Rzeczypospolitej Polskiej Bronisław Komorowski.

Numer specjalny składa się z dwóch zeszytów, w których zamieszczono łącznie 20 artykułów naukowych. Tematyka tych opracowań jest zróżnicowana. Część z nich ma charakter przeglądu historycznego, próby podsumowania pewnego etapu badań, czy nawet osobistych wspomnień związanych z pracami nad rozwojem statystyki w Polsce. Artykuł Walentego Ostasiewicza dotyczy rozwoju myśli statystycznej w Polsce w XIX w., zaś w artykule Krzysztofa Jajugi omówiony został rozwój polskiej myśli statystycznej w naukach ekonomicznych. Z kolei Tadeusz Caliński omawia rozwój i osiągnięcia biometrii polskiej. Podobnie, przeglądowy charakter posiada opracowanie Jana Kordosa na temat współzależności między rozwojem teorii a praktyki badań reprezentacyjnych w Polsce. Jan Paradysz przedstawia interesującą syntezę statystyki regionalnej, natomiast opracowania Wojciecha Roszki oraz Jacka Kowalewskiego i Moniki Natkowskiej dotyczą systemu statystyki publicznej i jej zastosowania w badaniach przedsiębiorstw. W pozostałych artykułach przedstawione są rozwiązania wybranych problemów szczegółowych. Daniel Kosiorowski prezentuje artykuł, w którym omawia wybrane zastosowania procedur indukowanych przez głębię położenia-rozrzutu Mizery i Müller w odpornej analizie strumienia danych ekonomicznych, Jan Owsiński zajmuje się problemem optymalnego podziału rozkładu empirycznego, natomiast Wioletta Grzenda analizuje determinanty pozostawania bez pracy osób młodych z wykorzystaniem semiparametrycznego modelu Coxa. Z kolei Maria Szmuksta-Zawadzka i Jan Zawadzki omawiają wybrane metody prognozowania na podstawie niekompletnych szeregów czasowych z wahaniami okresowymi. Artykuł Tomasza Klimanka dotyczy wykorzystania estymacji pośredniej z autokorelacją przestrzenną w badaniach opinii społecznej w Polsce, natomiast Andrzej Czyżewski i Aleksander Grzelak analizują możliwości wykorzystania statystyki bilansów przepływów międzygałęziowych dla makroekonomicznych ocen w gospodarce. Tematem pracy Magdaleny Okupniak jest zastosowanie analizy blokowej na przykładzie badania zróżnicowania dochodów gospodarstw domowych, zaś Łukasz Wawrowski bada zjawisko ubóstwa w przekroju powiatów w województwie wielkopolskim, z wykorzystaniem metod statystyki małych obszarów. Przedmiotem opracowania Romy Ryś-Jurek jest zastosowanie hierarchicznej klasyfikacji aglomeracyjnej do grupowania krajów Unii Europejskiej ze względu na strukturę i skalę produkcji gospodarstw rolnych, natomiast Hanna Gruchociak przedstawia problematykę delimitacji lokalnych rynków pracy w Polsce. Z kolei Aleksandra Łuczak i Feliks Wysocki omawiają zagadnienie zastosowania uogólnionej miary odległości GDM oraz metody TOPSIS do oceny poziomu rozwoju społeczno-gospodarczego powiatów województwa wielkopolskiego, zaś Iwona Bąk stosuje wybrane metody statystyczno-ekonometryczne do badania aktywności turystycznej seniorów w Polsce. Wreszcie Marzena Piotrowska-Trybull oraz Stanisław Sirko analizują wpływ jednostki wojskowej na rozwój gminy.

W zeszycie pierwszym przedstawione zostały ponadto: zarys historii Polskiego Towarzystwa Statystycznego opracowany przez Kazimierza Kruszkę, sprawozdanie z Kongresu, przygotowane przez Elżbietę Gołatę, sylwetki laureatów medalu im. Jerzego Spławy Neymana, przedstawione przez Czesława Domańskiego, a także Uchwała Kongresu Statystyki Polskiej z dnia 20 kwietnia 2012 r.

Wszystkie zamieszczone w numerze specjalnym artykuły naukowe były recenzowane, zgodnie z procedurą przyjętą przez Redakcję *Przeglądu Statystycznego* (zamieszczoną na stronie internetowej czasopisma).

Zachęcamy Czytelników do zapoznania się z tym wyjątkowym i unikatowym materiałem.

Redakcja

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

KAZIMIERZ KRUSZKA

POLSKIE TOWARZYSTWO STATYSTYCZNE (1912-2012)

Polskie Towarzystwo Statystyczne (PTS) należy do najstarszych organizacji tego rodzaju na świecie[1]. Powstało ono na początku 1912 r. i prowadziło swoją działalność statutową do 1915 r., ale o formalną jego likwidację nie wystąpiono. Można przyjąć, że kontynuatorem idei PTS od 1917 r. było Towarzystwo Ekonomistów i Statystyków Polskich, zwłaszcza w ramach istniejącej w nim Sekcji Statystyki. Powrót do organizacji pod nazwą Polskie Towarzystwo Statystyczne nastąpił w końcu 1937 r. Od tego czasu (z przerwą spowodowaną przez wojnę w latach 1939-1946) PTS działało faktycznie do 1953 r., a formalnie do 1955 r. W marcu 1953 r. powołana została Sekcja Statystyki w Polskim Towarzystwie Ekonomicznym, którą można traktować jako *sui generis* łącznik w ciągłości historycznej PTS. Reaktywowanie Polskiego Towarzystwa Statystycznego nastąpiło w kwietniu 1981 r. i odtąd jego działalność rozwijała się nieprzerwanie.

W miarę obszerną charakterystykę rozwoju organizacyjnego i działalności statutowej PTS zawiera monografia wydana w związku z Kongresem Statystyki Polskiej[2]. W poniższej syntezie uwagę skupiono tylko na wybranych faktach z historii PTS, podkreślając ich znaczenie dla rozwoju statystyki, a także wspominając tych członków Towarzystwa, którzy wnieśli szczególny wkład do skarbnicy wiedzy statystycznej. Biogramy zdecydowanej większości tych osób zawierają leksykony wydane z okazji 80. rocznicy istnienia GUS[3] oraz 100. rocznicy powstania PTS[4].

Inicjatywa zorganizowania stowarzyszenia statystyków polskich pojawiła się w środowisku krakowskim. Głównym celem działalności takiej organizacji miało być przygotowanie publikacji statystycznych obejmujących swym zasięgiem terytorialnym wszystkie trzy zabory. Do realizacji tego zamierzenia przyczynił się zwłaszcza kierownik

[1] Najwcześniej (1834 r.) powstało Brytyjskie Towarzystwo Statystyczne (Royal Statistical Society, RSS). Nieco później (1839 r.) założone zostało Amerykańskie Towarzystwo Statystyczne (American Statistical Association, ASA). Przyjmuje się, że Francuskie Towarzystwo Statystyczne (Société Française de Statistique, SFdS) swoją historią sięga do roku 1860, a Niemieckie Towarzystwo Statystyczne (Deutsche Statistische Gesellschaft, DSG) – do roku 1911. Później niż PTS powstały m. in. następujące stowarzyszenia statystyczne: japońskie (1931), belgijskie (1937), włoskie (1939), austriackie (1951), australijskie (1962), słowackie (1970), kanadyjskie (1972), szwajcarskie (1988), czeskie (1990), estońskie (1992).

[2] *Polskie Towarzystwo Statystyczne 1912-2012,* (2012), red. K. Kruszka, Polskie Towarzystwo Statystyczne, Rada Główna, Warszawa.

[3] *Słownik biograficzny statystyków polskich,* (1988), GUS i PTS, Warszawa.

[4] *Statystycy polscy*, (2012), GUS i PTS, Warszawa.

krakowskiego Miejskiego Biura Statystycznego doc. dr hab. Kazimierz Władysław Kumaniecki (późniejszy profesor UJ) przy poparciu prezydenta m. Krakowa, którym wtedy był prof. dr Juliusz Leo (również profesor UJ). W wyniku podjętych starań doszło 9 kwietnia 1912 r. do formalnej rejestracji stowarzyszenia pod nazwą *Polskie Towarzystwo Statystyczne* z siedzibą w Krakowie. Jego prezesem został prof. J. Leo, a sekretarzem był prof. K. W. Kumaniecki.

Za największe osiągnięcie „krakowskiego" PTS uznaje się przygotowanie i wydanie (w 1915 r.) publikacji pt. „Statystyka Polski". Znajdują się w nim zestawienia liczbowe (łącznie 315 tabel) dla obszaru Polski w granicach przedrozbiorowych z retrospekcją sięgającą nawet do 1815 r. Dzieło to zostało opracowane przez komitet redakcyjny pod przewodnictwem prof. Franciszka Bujaka, w którego skład wchodzili: dr Edward Grabowski, prof. Adam Krzyżanowski, prof. Kazimierz W. Kumaniecki i prof. Stefan Surzycki, a początkowo również dr Władysław Studnicki. Współtwórcami byli też m. in. dr Marcin Nadobnik z Lwowa i Michał Römer z Wilna, którzy nadesłali opracowane przez siebie materiały statystyczne. Radą zaś i – jak to określono w przedmowie tego wydawnictwa – „życzliwym poparciem" służyli: prof. Władysław Leopold Jaworski, prof. Stanisław Kutrzeba, prof. Michał Rostworowski oraz dr Franciszek Stefczyk. Przygotowaniem do druku zajmował się K. W. Kumaniecki razem z A. Krzyżanowskim, stąd też te dwie osoby wymienione są na karcie tytułowej jako autorzy całego opracowania.

Bez obawy o przesadę można twierdzić, że „Statystyka Polski" była dziełem epokowym, o wielkiej wartości poznawczej wówczas i obecnie, które również wielce przysłużyło się do wyznaczenia granic II Rzeczypospolitej. Jego opublikowanie należy ocenić jako poważne osiągnięcie naukowe, biorąc pod uwagę próbę integracji danych z trzech zaborów, zastosowane metody szacunków, a następnie kompilację, prezentację oraz interpretację otrzymanych wyników. Po wydaniu „Statystyki Polski" działalność krakowskiego PTS zamarła.

Pod koniec I wojny światowej centrum życia społeczno-politycznego tworzyło się w Warszawie. Tutaj w gronie współpracowników redakcji kwartalnika „Ekonomista" powstała wówczas myśl powołania do życia specjalistycznej organizacji naukowej pod nazwą Towarzystwo Ekonomistów i Statystyków Polskich (TEiSP). Celem tego stowarzyszenia miało być podnoszenie w Polsce poziomu wiedzy ekonomiczno-społecznej, zarówno teoretycznej jak i praktycznej. Metody statystyczne w dziedzinach niezwiązanych z państwem i prawem oraz gospodarką stosowano wówczas sporadycznie, toteż przy powołaniu TEiSP ekonomia była dyscypliną główną.

Pierwsze organizacyjne zebranie TEiSP odbyło się 3 grudnia 1917 r. Do Rady TEiSP zostali wybrani: prof. Jan Dmochowski, prof. Franciszek Doleżal, red. Stefan Dziewulski, prof. Kazimierz Kasperski, Stanisław A. Kempner, prof. Antoni Kostanecki, prof. Ludwik Krzywicki, prof. Zdzisław Ludkiewicz, prof. Jerzy Michalski, prof. Edward Strasburger, prof. Włodzimierz Wakar i prof. Władysław Zawadzki. Rada wyłoniła spośród swoich członków Zarząd, który ukonstytuował się następująco: przewod-

niczący – prof. Antoni Kostanecki, zastępcy przewodniczącego – Stefan Dziewulski i prof. Ludwik Krzywicki, skarbnik – prof. Władysław Zawadzki.

Na zebraniu organizacyjnym TEiSP w grudniu 1917 r. uchwalono utworzenie pięciu sekcji: teorii ekonomii, skarbowości, statystyki, polityki ekonomicznej, polityki społecznej. Pierwsze zebranie Sekcji Statystyki odbyło się 14 stycznia 1918 r. Jej przewodniczącym został prof. Ludwik Krzywicki, zastępcą przewodniczącego – prof. Edward Grabowski, a sekretarzem – prof. Stefan Szulc.

TEiSP rozpoczęło swą działalność bardzo energicznie. W ciągu sześciu początkowych miesięcy 1918 r. odbyły się trzy posiedzenia ogólne, w których brali udział przedstawiciele rządu polskiego, senatów Uniwersytetu Warszawskiego i Politechniki Warszawskiej oraz instytucji naukowych i społecznych. Prowadzono też intensywną działalność odczytową. W drugiej połowie 1918 r. TEiSP przejawiało już znacznie mniejszą aktywność. W 1919 r. zorganizowano wyłącznie zebrania ogólne TEiSP, które liczyło wtedy ok. 180 członków, a wśród nich było 15 osób spoza Warszawy. Działalność TEiSP zanikła w drugiej połowie 1920 r., by ożywić się wyraźnie w 1921 r. Liczba członków stale się zwiększała; w końcu 1924 r. było ich 218, w 1925 r. – 236, a w 1926 r. – 247. Na Walnym Zgromadzeniu 27 października 1924 r. do Rady TEiSP zostali m. in. wybrani: prof. Ludwik Krzywicki, Wacław Fabierkiewicz, Tadeusz Szturm de Sztrem i prof. Stefan Szulc. W 1925 r. na łamach „Ekonomisty" ukazał się artykuł J. Spławy-Neymana pt. *Uwagi o istocie badań statystycznych.*

W sierpniu 1929 r. miało miejsce wydarzenie wyjątkowo ważne dla środowiska statystyków polskich. W Warszawie odbyła się XVIII Sesja Międzynarodowego Instytutu Statystycznego (MIS). Przewodniczącym jej Komitetu Organizacyjnego był prof. Józef Buzek, a z Polski wzięły w niej udział 54 osoby, wśród których było 4 członków MIS. Warto tu dodać, że cały MIS liczył wtedy 190 członków, a w sesji warszawskiej brało udział ogółem (razem z gośćmi) 170 osób.

Lata wielkiego kryzysu gospodarczego, zwłaszcza zaś okres 1930-1932, były ciężkim czasem dla całego TEiSP. Liczba jego członków (według stanu w październiku 1932 r.) zmniejszyła się do 231. Dopiero w 1933 r. nastąpiło wyraźne ożywienie w środowisku statystyków, co było związane ze zmianami organizacyjnymi w TEiSP, a zwłaszcza z reaktywowaniem Sekcji Statystyki. Należało do niej początkowo 14 członków. Byli to: Jan Derengowski, Michał Kalecki, Ignacy Kräutler, Ludwik Landau, Zygmunt Limanowski, Stefan Moszczeński, przebywający za granicą Jerzy Spława-Neyman, Jan Piekałkiewicz, Franciszek Piltz, Edward Strzelecki, Edward Szturm de Sztrem, Tadeusz Szturm de Sztrem, Stefan Szulc i Jan Wiśniewski. Przewodniczącym Sekcji z ramienia Towarzystwa Ekonomistów i Statystyków Polskich został prof. Zygmunt Limanowski, wiceprzewodniczącym z wyboru - prof. Stefan Szulc, a sekretarzem - Jan Derengowski. Zarząd w tym samym składzie był wybierany przez aklamację jeszcze dwukrotnie (w 1936 i 1937 r.).

Liczba członków Sekcji Statystyki w TEiEP podwoiła się w ciągu 1934 r. W grudniu 1935 r. należało do niej 29 osób, a w końcu 1936 r. było ich 32 i tak pozostało do 16 grudnia 1937 r. Członkowie Sekcji Statystyki rekrutowali się głównie spośród

pracowników GUS (ok. 35%), Instytutu Badania Koniunktur Gospodarczych i Cen (ok. 17%), Wydziału Statystycznego Zarządu m. st. Warszawy (ok. 10%), a resztę stanowili pracownicy innych instytucji warszawskich. Łącznie w latach 1934-1937 odbyło się 30 zebrań naukowych Sekcji Statystyki TEiSP, na których 17 osób przedstawiło 33 referaty.

Wprawdzie statystycy mieli formalne możliwości działania w ramach TEiSP, jednak dostrzegali w tej organizacji faktyczną dominację ekonomistów. To odczucie oraz rosnące aspiracje zawodowe i naukowe, zwłaszcza w kontekście rozwijającej się działalności Głównego Urzędu Statystycznego i specjalistycznego kształcenia na wyższych uczelniach, a także rosnące w praktyce gospodarczej zapotrzebowanie na fachowców posługujących się metodami statystycznymi – sprawiały, że coraz częściej i wyraźniej podnoszony był postulat utworzenia własnej organizacji statystyków o zasięgu ogólnopolskim. Tej sprawie poświęcone zostało specjalne zebranie Sekcji Statystyki TEiSP, które odbyło się 16 grudnia 1936 r. Podjęto na nim uchwałę, że „Sekcja Statystyczna Towarzystwa Ekonomistów i Statystyków Polskich uznaje potrzebę zwołania zjazdu statystyków polskich o charakterze organizacyjnym i naukowym oraz potrzebę stworzenia ogólnopolskiego towarzystwa statystycznego". Dało to podstawy do rozpoczęcia przygotowań mających na celu ponowne powołanie Polskiego Towarzystwa Statystycznego.

Zebranie organizacyjne PTS odbyło się 17 stycznia 1937 r., a Walne Zgromadzenie Konstytucyjne obradowało 31 października i 1 listopada 1937 r. Na przewodniczącego Zgromadzenia wybrano prof. Jana Czekanowskiego, a do prezydium – dra Rajmunda Buławskiego, prof. Antoniego Łomnickiego prof. Marcina Nadobnika i prof. Jana Piekałkiewicza. W trakcie obrad omawiano plany pracy, działalność wydawniczą i organizację zebrań naukowych. Akcentowano też potrzebę promocji i popularyzacji statystyki. Ustalono powołanie czterech sekcji tematycznych jako komórek naukowych Towarzystwa, tj. Sekcji Statystyki Matematycznej, Sekcji Statystyki Ludności, Sekcji Statystyki Gospodarczej i Społecznej oraz Sekcji Statystyki w Przedsiębiorstwie. Do Zarządu, wyłonionego spośród Rady PTS, weszli: prezes – Edward Szturm de Sztrem, wiceprezes – prof. Jan Czekanowski, sekretarz – Jan Derengowski, skarbnik – dr Jan Wiśniewski, członkowie – dr Rajmund Buławski, Zbigniew Łomnicki i Edward Strzelecki, zastępcy członków – dr Bronisław Biegeleisen, Karol Czernicki i Edward Rosset. Członkami Rady – poza członkami Zarządu – zostali: dr Stanisław Antoniewski, prof. Franciszek Bujak, dr Marcin Kacprzak, prof. Adam Krzyżanowski, prof. Kazimierz W. Kumaniecki, prof. Stanisław Lencewicz, prof. Zygmunt Limanowski, prof. Antoni Łomnicki, prof. Stefan Mazurkiewicz, prof. Marcin Nadobnik, prof. Jerzy Spława-Neyman, prof. Jan Piekałkiewicz, dr Stanisław Pszczółkowski, prof. Stefan Szulc i Zygmunt Zaleski.

Powstanie Polskiego Towarzystwa Statystycznego było przychylnie zauważone na arenie międzynarodowej, o czym świadczyły m. in. liczne pisma z życzeniami pomyślnego rozwoju, które władze PTS otrzymały od pokrewnych towarzystw statystycznych

za granicą. Podobne życzenia złożył także prezydent Międzynarodowego Instytutu Statystycznego, prof. Armand Julin.

Na Zwyczajnym Walnym Zgromadzeniu PTS, które odbyło się 2 kwietnia 1939 r., ponownie prezesem PTS został Edward Szturm de Sztrem, a wiceprezesem – prof. Jan Czekanowski. Ponadto do Zarządu PTS jako członkowie weszli: Jan Derengowski, dr Stanisław Kołodziejczyk, prof. Jan Piekałkiewicz, Stanisław Rutkowski i dr Jan Wiśniewski, a zastępcami członków byli Zbigniew Łomnicki, Wacław Skrzywan i Edward Strzelecki. Do Rady PTS z urzędu należeli członkowie Zarządu, a ponadto jej skład tworzyli: dr Stanisław Antoniewski, dr Jan Blaton, prof. Franciszek Bujak, prof. Marcin Kacprzak, prof. Adam Krzyżanowski, prof. Kazimierz W. Kumaniecki, prof. Zygmunt Limanowski, prof. Edward Lipiński, dr Józef Poniatowski, dr Stanisław Pszczółkowski, prof. Jerzy Spława-Neyman, prof. Stefan Szulc i Zygmunt Zaleski.

Liczba członków PTS szybko się zwiększała. W połowie marca 1938 r. było w nim 197 członków zwyczajnych i 11 wspierających, a na liście sporządzonej według stanu w dniu 15 czerwca 1939 r. znajdowało się 291 członków zwyczajnych (w tym 20 kobiet) i 30 wspierających. Największa grupa członków zwyczajnych mieszkała w Warszawie lub w jej pobliżu (63%). Poza Warszawą najliczniej reprezentowany był Lwów (12%), a dalsze miejsca zajmowały: Kraków (5%), Katowice z pobliskimi miastami (prawie 5%), Poznań i Wilno (po ok. 4%). Poza granicami ówczesnej Polski mieszkało 6 członków (2%), a pozostali (5%) byli rozproszeni w różnych miastach kraju (m.in. byli to mieszkańcy Bydgoszczy, Gdyni, Lublina i Łodzi).

Jeszcze w 1937 r. powstały i rozpoczęły działalność trzy sekcje PTS, tj. Sekcja Statystyki Matematycznej, Sekcja Statystyki w Przedsiębiorstwie oraz Sekcja Statystyki Gospodarczej i Społecznej. Przewodniczącymi zarządu tych sekcji byli, odpowiednio: prof. Antoni Łomnicki, prof. Jan Piekałkiewicz i prof. Władysław Zawadzki. W marcu 1938 r. powstała Sekcja Statystyki Ludności, której przewodniczącym zarządu został prof. Stefan Szulc.

Działalność PTS w latach 1938-1939 koncentrowała się w Warszawie, ale nie utworzono tu wyodrębnionej terytorialnie jednostki organizacyjnej. Powołane zostały natomiast cztery oddziały terenowe PTS: Oddział Śląsko-Dąbrowski w Katowicach, Oddział Wileński, Oddział Poznański i Oddział Lwowski. Przewodniczącymi zarządu w tych oddziałach byli, odpowiednio: dr Rajmund Buławski, Teodor Nagurski, prof. Marcin Nadobnik i prof. Antoni Łomnicki.

PTS już w listopadzie 1937 r. postawiło sobie jako jedno z naczelnych zadań wydawanie czasopisma poświęconego teorii i praktyce statystycznej. Stał się nim „Przegląd Statystyczny", uznany za oficjalny organ PTS. Jego komitet redakcyjny tworzyli: Zygmunt Limanowski (przewodniczący), Stefan Szulc (zastępca przewodniczącego), Jan Wiśniewski (sekretarz), Antoni Łomnicki i Jan Piekałkiewicz (członkowie). Później w skład tego komitetu wszedł jeszcze Stanisław Pszczółkowski. Czasopismo to miało ukazywać się cztery razy w roku. Postanowiono publikować w nim oryginalne prace teoretyczne autorów polskich i obcych, oryginalne i ważne ze względów metodologicznych prace z zastosowań statystyki, referaty przedstawiające stan lub postępy

poszczególnych gałęzi statystyki, recenzje i krytyki dochodzeń statystycznych, kronikę informującą o życiu organizacyjnymi i naukowym PTS oraz pokrewnych instytucji zagranicznych i międzynarodowych, a także polską bibliografię statystyczną. Program redakcyjny był konsekwentnie realizowany w kolejnych numerach „Przeglądu Statystycznego", z których pierwszy ukazał się już na początku 1938 r.

Drugim periodykiem przedwojennego PTS była „Statystyka w Przedsiębiorstwie", wydawana jako biuletyn Sekcji Statystyki w Przedsiębiorstwie. Zakładano, że będzie to miesięcznik, ale w praktyce okazało się, że większość edycji stanowiły tzw. numery łączone, obejmujące najczęściej okresy dwumiesięczne. Znajdowały się w nich sprawozdania i kronika Sekcji, streszczenia wygłoszonych referatów oraz dyskusje nad tymi referatami, a także bibliografia prac z dziedziny statystyki w przedsiębiorstwie.

Wybuch II wojny światowej 1 września 1939 r. zniweczył plany kontynuacji dobrze rozwijającej się pracy Polskiego Towarzystwa Statystycznego. Wielu spośród członków PTS uczestniczących w kampanii wrześniowej poległo lub zmarło w wyniku odniesionych ran, niektórzy dostali się do niewoli i zostali zamordowani lub przebywali w obozach jenieckich, a o kilkunastu wiadomo tylko, że zaginęli w zawierusze wojennej. Pewna część żyjących znalazła się poza granicami Polski, ale większość musiała zmierzyć się z nową ponurą rzeczywistością w kraju. Nie wszyscy z nich doczekali wyzwolenia. Dotychczas ustalono, że w czasie II wojny światowej i tuż po jej zakończeniu w różnych okolicznościach (także podczas powstania warszawskiego) straciło życie 54 członków PTS (w tym 18 profesorów i 6 doktorów), czyli ponad 18% ogólnej ich liczby z czerwca 1939 r. Prawie połowa z nich należała do władz Towarzystwa i pełniła w nich ważne funkcje.

W warunkach okupacji hitlerowskiej, a także na terenach okupowanych przez ówczesny ZSRR niemożliwa była jakakolwiek działalność statutowa PTS. O niektórych członkach PTS wiadomo, że uczestniczyli w ruchu oporu (w różnych formach i organizacjach), prowadzili tajne nauczanie, badali i dokumentowali warunki życia w okupowanej Polsce, organizowali i realizowali inne badania statystyczne (najczęściej w konspiracji), chronili przedwojenne materiały statystyczne i księgozbiór biblioteczny, pracowali za granicą na rzecz kraju, który musieli opuścić.

Ratowanie przedwojennych zasobów informacyjnych GUS i innych instytucji, prowadzenie badań statystycznych w okupowanej Polsce, dokumentowanie ówczesnej sytuacji w wielu dziedzinach i aspektach, samokształcenie i działalność edukacyjna – to szczególne wyróżniki pracy członków PTS w warunkach wojennych, której wyniki tworzyły cenny depozyt przekazany do dyspozycji i pomnażania w następnych latach.

Działalność PTS po wojennej przerwie została wznowiona w marcu 1947 r., a Walne Zgromadzenie Członków (zebranie konstytucyjne) w pierwszej powojennej kadencji PTS odbyło się 15 czerwca 1947 r. Wybrano na nim władze Towarzystwa oraz podjęto uchwały w sprawie organizacji i podstawowych kierunków działania. Skład wybranego Zarządu PTS był następujący: prof. Stefan Szulc – przewodniczący, prof. Tadeusz Banachiewicz – wiceprzewodniczący, Zygmunt Zaremba – sekretarz, dr Egon Vielrose – skarbnik oraz prof. Marcin Kacprzak, Franciszek Król i dr Kazimierz Romaniuk.

Zastępcami członków Zarządu zostali: Kazimierz Kochański, Zygmunt Padowicz i dr Stanisław Waszak. Do Rady PTS, poza członkami i zastępcami członków Zarządu, weszli: dr Stanisława Adamowiczowa, dr Stanisław Antoniewski, prof. Franciszek Bujak, dr Rajmund Buławski, prof. Jan Czekanowski, prof. Adam Krzyżanowski, prof. Edward Lipiński, Juliusz Miller, Wiktor Morawski, prof. Marcin Nadobnik, Stanisław Róg, prof. Wacław Skrzywan, prof. Hugo Steinhaus, prof. Edward Strzelecki i prof. Edward Szturm de Sztrem. Komisję Rewizyjną tworzyli: dr Konstanty Czerniewski (przewodniczący), prof. Henryk Greniewski, dr Halina Grużewska, Stanisław Stęplewski i Józef Wojtyniak.

Pierwsze posiedzenie nowo wybranej Rady PTS odbyło się 6 lipca 1947 r. Postanowiono na nim wznowić działalność czterech sekcji z okresu przedwojennego i utworzyć nową Sekcję Statystyki Miejskiej, a także zorganizować pracę w oddziałach Towarzystwa. Podjęto też decyzję o wydawaniu „Przeglądu Statystycznego" i powołano Komitet Redakcyjny w składzie: prof. Stefan Szulc (przewodniczący), prof. Tadeusz Banachiewicz (wiceprzewodniczący), Maria Namysłowska (sekretarz) oraz prof. Marcin Kacprzak i dr Kazimierz Romaniuk (członkowie).

Pierwszy zaczął działać Oddział Morski (Gdański) PTS, któremu status formalny nadano w styczniu 1948 r. Warszawski Oddział PTS został zorganizowany 17 lutego 1948 r. W początkowym stadium organizacji były cztery dalsze oddziały: krakowski, łódzki, poznański i wrocławski. Według stanu na 1 lipca 1949 r. PTS miało 154 członków zwyczajnych (w tym 37 kobiet) i 11 członków wspierających. Działalność sekcji i oddziałów PTS w latach 1947-1949, choć nie była tak intensywna jak w okresie przedwojennym, miała istotne znaczenie dla dalszego kształtowania polskiej myśli i praktyki statystycznej. Wpływały na to zwłaszcza zebrania naukowe poświęcone różnym zagadnieniom teoretycznym i bieżącym problemom badań statystycznych. W następnych latach aktywność PTS stopniowo gasła. Złożyły się na to głównie narastające trudności natury politycznej i finansowej, a także ograniczenia organizacyjne.

Przytoczone tu okoliczności oraz szybko słabnąca wśród członków PTS wola kontynuacji pracy w tym Towarzystwie sprawiły, że w 1951 r. jego działalność faktycznie zamarła. W zaistniałej sytuacji Zarząd PTS zebrał się ostatni raz 19 listopada 1953 r., podejmując uchwałę o zwołaniu Walnego Zgromadzenia w celu likwidacji Towarzystwa. Odbyło się ono 12 grudnia 1953 r., a uczestniczyło w nim tylko 17 członków PTS (formalnie do PTS należało wówczas 178 osób). Zebrani udzielili absolutorium Zarządowi PTS i podjęli uchwałę o likwidacji Polskiego Towarzystwa Statystycznego. Obowiązki komisji likwidacyjnej powierzono ostatniemu Zarządowi PTS. Jej posiedzenie odbyło się dopiero 4 kwietnia 1955 r. Sporządzono na nim protokół przekazujący Polskiemu Towarzystwu Ekonomicznemu akta, księgozbiory i wydawnictwa PTS[5], a także protokół dla Prezydium Rady Narodowej m. Warszawy o likwidacji PTS. Te dokumenty formalnie rozpoczęły trwającą 26 lat nieobecność Polskiego Towarzystwa Statystycznego wśród stowarzyszeń naukowych.

[5] Ostatni numer „Przeglądu Statystycznego" jako organu PTS ukazał się na początku 1950 r.

Realizując wytyczne I Kongresu Nauki Polskiej i postulaty Ogólnopolskiej Konferencji Katedr Statystyki, Zarząd Główny Polskiego Towarzystwa Ekonomicznego podjął w marcu 1953 r. uchwałę o powołaniu Sekcji Statystyki w ramach PTE. Sekcje statystyczne w ramach oddziałów PTE zostały zorganizowane w Katowicach, Krakowie, Łodzi, Poznaniu, Szczecinie, Warszawie i we Wrocławiu, a rolę koordynatora wyznaczono sekcji warszawskiej. Przewidywano też tworzenie sekcji terenowych w innych oddziałach PTE, co później realizowano. Dostępne publikacje wskazują, że aktywność statystyków zrzeszonych w PTE przejawiała się głównie w akcji odczytowej, organizowaniu i prowadzeniu szkoleń oraz we współpracy ze szkołami i szkolnymi kołami zainteresowań (zwłaszcza w szkołach ekonomicznych i rolniczych). Na przeszkodzie w rozwinięciu szerszej działalności stawały braki finansowe, małe zainteresowanie ze strony Zarządu Głównego PTE i podobne okoliczności, ale główna przyczyna słabego rozwoju Sekcji Statystyki tkwiła chyba w małym zaangażowaniu i nikłej inicjatywie jej członków, którą krępowały ówczesne realia gospodarcze i polityczne[6]. Jednak dzięki temu forum, jakim była Sekcja Statystyki PTE, nie doszło do pełnej dezintegracji środowiska statystyków polskich i ich zagubienia w ówczesnej rzeczywistości. Jej członkowie oraz inni statystycy coraz bardziej uświadamiali sobie potrzebę ponownego zorganizowania się w odrębnym stowarzyszeniu, do czego doprowadzono na początku 1981 r.

Idea reaktywowania PTS trafiła na podatny grunt i 16 kwietnia 1981 r. w Warszawie obradowało Zgromadzenie Założycielskie tego stowarzyszenia. Wybrana na nim Tymczasowa Rada Główna PTS, której przewodniczył prof. Mikołaj Latuch, działała 19 miesięcy. W tym czasie został zarejestrowany statut PTS, utworzono 9 oddziałów terenowych, przyjmowano nowych członków i przygotowano program dalszych prac. Pierwsze po reaktywowaniu PTS Walne Zgromadzenie Delegatów odbyło się 29 listopada 1982 r. Prezesem został prof. M. Latuch, a na wiceprezesów wybrano prof. Andrzeja Barczaka i prof. Ryszarda Zasępę. Kadencja nowych władz PTS trwała nieco ponad 3 lata. Funkcjonowało 9 oddziałów terenowych, a liczba członków zwyczajnych zwiększyła się do prawie 400. Towarzystwo borykało się z dużymi trudnościami finansowymi, co zasadniczo ograniczało jego aktywność.

Kolejny etap działalności PTS rozpoczęło Walne Zgromadzenie Delegatów obradujące 14 grudnia 1985 r. w Warszawie. Nowym prezesem PTS został prof. Jan Kordos, a funkcje wiceprezesów objęli profesorowie Czesław Domański i R. Zasępa. Od grudnia 1985 r. do marca 1990 r. powołano w PTS dwie sekcje, pięć komisji i dwa zespoły problemowe; gremia te prowadziły działalność o zasięgu ogólnopolskim. Przy Oddziale PTS w Krakowie zlokalizowana została Ogólnopolska Sekcja Taksonomiczna, a przy Oddziale Gdańskim – Ogólnopolska Sekcja Metodologii i Organizacji Badań. W październiku 1986 r. podjęto decyzję o wydawaniu pisma pt. „Biuletyn Informacyjny

[6] Miało to również odzwierciedlenie na łamach „Przeglądu Statystycznego", który w latach 1954-1973 ukazywał się jako organ Sekcji Statystyki PTE. Od 1974 r. „Przegląd Statystyczny" stał się organem Komitetu Statystyki i Ekonometrii PAN.

RG PTS". Pismo to spełniało bardzo ważną rolę nie tylko jako przekaźnik informacji i wiedzy, ale również – co trzeba silnie podkreślić jako wielką wartość – w istotny sposób przyczyniało się do integracji wszystkich ogniw i członków stowarzyszania oraz kształtowało wizerunek i tożsamość PTS[7]. W marcu 1987 r. utworzono Biuro Badań i Analiz Statystycznych przy Radzie Głównej PTS (BBiAS), którego kierownikiem został Wiesław Łagodziński. W 1987 r. mijało 75 lat od chwili założenia PTS w Krakowie. Centralnym wydarzeniem w obchodach tego jubileuszu była konferencja naukowa zorganizowana 26 października 1987 r. w Warszawie. W sierpniu 1989 r. miesięcznik „Wiadomości Statystyczne" stał się wspólnym organem Głównego Urzędu Statystycznego i Polskiego Towarzystwa Statystycznego. Zwiększyło się oddziaływanie PTS i wzrosła jego atrakcyjność w środowisku statystyków. W styczniu 1990 r. do PTS należało ok. 950 osób zgrupowanych w jego 26 zarejestrowanych oddziałach. Można zatem powiedzieć, że Polskie Towarzystwo Statystyczne przekształciło się w organizację masową o zasięgu ogólnopolskim.

Dynamiczny rozwój PTS notuje się również w następnym okresie, tj. od marca 1990 r. do września 1994 r. Funkcję prezesa w tym czasie ponownie pełnił prof. Jan Kordos, a wiceprezesami byli prof. Andrzej Balicki i prof. Ryszard Zasępa. W 1992 r. PTS obchodziło 80-lecie swojego założenia, a uroczystości jubileuszowe odbyły się 19 maja w Mogilanach k. Krakowa i były połączone z konferencją naukową. W styczniu 1992 r. Sekcja Taksonomiczna została przekształcona w Sekcję Klasyfikacji i Analizy Danych (SKAD) i pod taką nazwą nadal prowadzi swoją działalność, będąc od 1994 r. członkiem Międzynarodowej Federacji Towarzystw Klasyfikacyjnych (IFCS). W kwietniu 1994 r. PTS stało się członkiem zbiorowym Międzynarodowego Instytutu Statystycznego (ISI). Ważnym wydarzeniem – tak dla PTS, jak i całej polskiej statystyki – stało się powołanie czasopisma „Statistics in Transition". Ten organ PTS, o charakterze międzynarodowym, powstał z inicjatywy prof. Jana Kordosa, który też został jego redaktorem naczelnym. Pierwszy numer nowego pisma, które jest wydawane w języku angielskim, ukazał się w lipcu 1993 r. Zyskało ono wysokie uznanie statystyków z różnych krajów, a wielu z nich ściśle współpracuje przy jego tworzeniu. We wrześniu 1994 r. funkcjonowało 25 oddziałów, zrzeszających około 1100 członków PTS. Nadal bardzo efektywnie działało Biuro Badań i Analiz Statystycznych, zapewniając środki finansowe dla rozwijającego się Towarzystwa. Można ogólnie stwierdzić, że PTS w latach 1990-1994 umocniło się organizacyjnie, rozbudowało swoje struktury ogólnopolskie i terenowe, weszło szerszym frontem na areną międzynarodową. Rozwijały się jego kontakty z innymi stowarzyszeniami, dobrze układała się współpraca z Głównym Urzędem Statystycznym i wojewódzkimi urzędami statystycznymi, wzrastała aktywność PTS na terenie wyższych uczelni i instytutów naukowych oraz w szkolnictwie średnim.

[7] Na okładce tego pisma po raz pierwszy pojawił się emblemat PTS, który umieszczano na wszystkich jego wydaniach. Był to zarazem pierwowzór dzisiejszego logo PTS. Znalazł się on również na znaczku organizacyjnym PTS.

Sumaryczne ujęcie dorobku PTS w latach 1990-1994 przedstawiono na Walnym Zgromadzeniu Delegatów w październiku 1994 r. Prezesem wybrano na nim prof. Czesława Domańskiego, a wiceprezesami zostali: dr Kazimierz Kruszka, Wiesław Łagodziński i prof. Mirosław Szreder. W kwietniu 1997 r. minęło 85 lat od założenia PTS w Krakowie. Z tej okazji ukazał się numer specjalny „Biuletynu Informacyjnego RG PTS", gdzie zamieszczono kronikę działalności Polskiego Towarzystwa Statystycznego w latach 1912-1997. Zorganizowano również konferencję naukową i wystawę w Łodzi (27 i 28 listopada 1997 r.), którą poświęcono prof. Edwardowi Rossetowi w 100-lecie jego urodzin. W grudniu 1998 r. ukazał się ostatni (44) numer „Biuletynu Informacyjnego RG PTS", który – zgodnie z uchwałą Rady Głównej PTS – został zastąpiony przez nowe pismo pt. „Kwartalnik Statystyczny".

W kadencji 2000-2005 ponownie prezesem PTS był prof. Czesław Domański, a funkcje wiceprezesów pełnili: dr Bogusława M. E. Bulska, Wiesław Łagodziński i prof. Aleksander Zeliaś. W uznaniu wybitnych zasług dla statystyki członkami honorowymi PTS w 2000 r. zostali: dr Maria Czarnowska, dr Richard Platek, dr Stanisław Róg i prof. Kazimierz Zając. W 2002 r. PTS obchodziło 90-lecie swojej działalności. Z tej okazji 15 i 16 lipca w Modlnicy k. Krakowa odbyła się uroczysta konferencja połączona z wystawą obrazującą historię PTS. Wybity został również okolicznościowy medal, który otrzymali zasłużenie członkowie i sympatycy PTS. W listopadzie 2005 r. Członkami Honorowymi PTS zostali profesorowie: Jan Kordos, Tadeusz Walczak i Stanisław Wierzchosławski. W następnych latach działalność PTS znacznie osłabła. Po licznych zmianach organizacyjnych do grudnia 2005 r. funkcjonowało 18 oddziałów regionalnych, skupiając łącznie około 750 członków.

W następnej kadencji (2005-2010) prezesem PTS był dr Kazimierz Kruszka, a na wiceprezesów wybrano: W.W. Łagodzińskiego, prof. Mirosława Szredera i prof. Aleksandra Zeliasia. Po śmierci prof. Aleksandra Zeliasia wiceprezesem został prof. Walenty Ostasiewicz (od 5 V 2006). W 2007 r. mijało 95 lat od założenia PTS. Z tej okazji od 10 do 12 października we Wrocławiu obradował I Ogólnopolski Zjazd Statystyków połączony z konferencją naukową pt. „Statystyka wczoraj, dziś i jutro". Wzięło w nim udział ponad 100 osób reprezentujących instytucje naukowe, służby statystyki publicznej i resortowe komórki statystyczne W listopadzie 2009 r. powołane zostały kolejne, obok SKAD, sekcje PTS, tj. Sekcja Statystyki Matematycznej i Sekcja Historyczna. Założona została profesjonalna strona www.stat.gov.pl/pts (rozwinięta jej wersja redagowana jest w języku polskim, a skrócona – w języku angielskim). Ukazują się na niej m.in. komunikaty o bieżących wydarzeniach, o konferencjach i seminariach naukowych, zamieszczane są kolejne zeszyty „Wiadomości Statystycznych" oraz „Statistics in Transition-ns". Własną stronę internetową, również w wersji polskiej i angielskiej, prowadzi SKAD.

Ważnym wydarzeniem dla członków PTS i całego środowiska statystycznego było ustanowienie Dnia Statystyki Polskiej. Z inicjatywą organizowania takiego święta wystąpiła Rada Główna PTS, co znalazło poparcie również w innych gremiach. Formalnym tego wyrazem stała się uchwała podjęta 2 grudnia 2008 r. przez Radę Główną

PTS, Komitet Statystyki i Ekonometrii PAN oraz Główny Urząd Statystyczny, która spowodowała, że na datę tego corocznego święta obrano 9 marca. Po raz pierwszy obchodzono Dzień Statystyki Polskiej w 2009 r., a miejscem uroczystej inauguracji obchodów, połączonej z seminarium naukowym, był gmach Sejmu RP.

Ogólnie można stwierdzić, że w kadencji obejmującej okres od 15 XI 2005 r. do 10 II 2010 r. działalność Polskiego Towarzystwa Statystycznego cechowała dążność do konsolidacji wewnętrznej i zacieśnienia współpracy ze służbami statystyki publicznej, z wyższymi uczelniami, a także z innymi stowarzyszeniami w kraju i za granicą. Łączenie wysiłków i środków tych podmiotów dla realizacji wspólnych celów okazało się ze wszech miar korzystne, wpływając też znacząco na postrzeganie roli i znaczenia PTS. Dzięki zaś wysiłkom podejmowanym przez Biuro Badań i Analiz Statystycznych nastąpiło wyraźne polepszenie kondycji finansowej Towarzystwa, co procentowało również w następnej kadencji władz PTS.

W wyniku kolejnych wyborów prezesem PTS od 10 lutego 2010 r. został prof. Czesław Domański, a wiceprezesami są: Wiesław W. Łagodziński, dr Krzysztof Najman i prof. Grażyna Trzpiot. W 2011 r. do PTS należało 756 osób zgrupowanych w 17 oddziałach regionalnych. Od lutego 2011 r. jako najważniejszą sprawę uznano zorganizowanie Kongresu Statystyki Polskiej z okazji 100. rocznicy powstania PTS.

Polskie Towarzystwo Statystyczne było zawsze płaszczyzną sporu naukowego jako nieodzownego stymulatora postępu, ale jednocześnie umożliwiało współdziałanie wielu ludzi, zespołów, racji czy poglądów. Wynikało to ze statutu stowarzyszenia, a raczej należałoby powiedzieć, że zostało do statutu wpisane jako formalny wyraz dobrej praktyki. Jest wiele dowodów na to, iż PTS było inicjatorem licznych przedsięwzięć o dużym znaczeniu dla praktyki i teorii statystycznej. Pełniło też wielokrotnie rolę stymulatora i katalizatora rozwoju wielu dziedzin statystyki, choć bywało, że próbowano je zatrudnić jako „hamulcowego". Zasługą PTS jest także to, że jego forum mogło służyć do skracania dystansu między teorią i praktyką statystyczną, do zbliżenia postaw cechujących statystyków matematycznych, reprezentantów statystyki społecznej, ekonomicznej itp.

* * *

Przypominając w dużym skrócie historię Polskiego Towarzystwa Statystycznego (we wszystkich formach i nazwach, pod którymi występowało), chcemy podkreślić prawdę wydawałoby się oczywistą, że jego dorobek i wizerunek tworzyli ludzie, tzn. setki a nawet tysiące członków i sympatyków tego zasłużonego stowarzyszenia. Zarówno indywidualny, jak i zbiorowy trud tych osób przyczynił się do rozwoju statystyki w ogóle, a rodzimej w szczególności. Polskie Towarzystwo Statystyczne ma też niemałe zasługi w budowaniu „gmachu", który nazywa się Rzeczpospolita Polska.

W krakowskim okresie PTS (1912-1915) powstało dzieło nazwane *Statystyka Polski*, które – mimo zaborów – opisało Rzeczpospolitą jako całość i wydatnie przyczyniło się do określenia jej granic w okresie porozbiorowym. Niemałe są też zasługi działającego w okresie międzywojennym Towarzystwa Ekonomistów i Statystyków Polskich (1917-1937), a zwłaszcza Polskiego Towarzystwa Statystycznego (1937-1939).

W czasie II wojny światowej członkowie PTS pracowali dla Polski na miarę ówczesnych możliwości, nierzadko ponad tę miarę, a wielu oddało swoje życie w służbie dla Ojczyzny. Po wojnie, w ramach PTS (1947-1953) oraz w Sekcji Statystyki PTE (1953-1980), mimo panujących wtedy ograniczeń członkowie tych stowarzyszeń również dołożyli sporo „dobrych cegiełek" do konstrukcji metaforycznie traktowanego tu „gmachu". PTS reaktywowane w 1981 r. ma także niewątpliwy udział w przemianach społeczno-politycznych, jakie dokonały się w naszym kraju, w budowaniu społeczeństwa obywatelskiego i kształtowaniu pozycji Polski w Unii Europejskiej. Oddanie statystyków służbie Rzeczypospolitej upamiętnia okolicznościowy medal wydany z okazji jubileuszu 100-lecia PTS.

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

ELŻBIETA GOŁATA

KONGRES STATYSTYKI POLSKIEJ Z OKAZJI – 100-LECIA POLSKIEGO TOWARZYSTWA STATYSTYCZNEGO POZNAŃ, 18-20 KWIETNIA 2012 R.

W dniach 18-20 kwietnia 2012 roku obradował w Poznaniu Kongres Statystyki Polskiej będący centralnym wydarzeniem inaugurującym uroczyste obchody 100-lecia Polskiego Towarzystwa Statystycznego. Na trwające przez cały rok obchody składają się także wystawy, seminaria i konferencje organizowane przez oddziały regionalne PTS, między innymi we Wrocławiu, Lublinie, Toruniu i Bydgoszczy.

Kongres Statystyki Polskiej był wyjątkowym wydarzeniem o międzynarodowym wymiarze, nad którym Honorowy Patronat objął Prezydent Rzeczypospolitej Polskiej Bronisław Komorowski. W liście gratulacyjnym skierowanym do uczestników kongresu, Prezydent Rzeczypospolitej wyraził uznanie dla rangi tego wydarzenia oraz przypomniał główny cel, jaki przyświecał założycielom Towarzystwa. Podkreślając znaczenie opublikowanej w 1915 r. *Statystyki Polski,* wskazał, że była ona cennym źródłem wiedzy o ziemiach ojczyzny podzielonej miedzy trzech zaborców, którą wykorzystano miedzy innymi w procesie ustalania granic odrodzonej Rzeczypospolitej. Prezydent Komorowski zauważył, że dalsze dzieje PTS noszą ślady dramatycznej historii Polski, że *„. . . Członkowie Towarzystwa angażowali się w walkę o odzyskanie niepodległości i jej obronę, intensywnie działali na rzecz zachowania polskiej tożsamości narodowej oraz tworzenia i rozwoju społeczeństwa obywatelskiego."*

Godną oprawę wielkiego święta polskiej statystyki zapewnił udział znamienitych przedstawicieli instytucji i organizacji statystycznych z Austrii, Australii, Estonii, Francji, Holandii, Niemiec, Norwegii, Stanów Zjednoczonych i z Włoch. Obrady Kongresu Statystyki Polskiej zgromadziły około 600 uczestników. Wśród gości specjalnych uroczystej Sesji Jubileuszowej znaleźli się przedstawiciele najwyższych władz państwowych oraz samorządowych miasta i regionu: prof. Marek Ratajczak – Podsekretarz Stanu w MNiSW, prof. R. Słowiński członek korespondent PAN – Prezes Oddziału PAN w Poznaniu, prof. E. Gatnar – Członek Zarządu NBP, dr K. Łybacka – Poseł Sojuszu Lewicy Demokratycznej, P. Florek – Wojewoda Wielkopolski, T. Kayser – Zastępca Prezydenta Miasta Poznania.

Uroczystości jubileuszowe uświetnili także przedstawiciele międzynarodowych instytucji statystycznych: dr M. Bohata – Zastępca Dyrektora Generalnego Urzędu Statystycznego Krajów UE – Eurostatu, prof. I. Ritschelova – Prezes Czeskiego Urzędu Statystycznego, dr E. Laczka – Wiceprezes Węgierskiego Urzędu Statystycznego oraz przedstawiciele międzynarodowych towarzystw naukowych: prof. R. Chambers

– Przewodniczący Międzynarodowego Towarzystwa Statystyków Badań Reprezentacyjnych Międzynarodowego Instytutu Statystycznego, prof. R. Wasserstein – Dyrektor Wykonawczy Amerykańskiego Towarzystwa Statystycznego, prof. I. Traat – Prezes Estońskiego Towarzystwa Statystycznego, prof. W. Schmid – członek Rady Głównej Niemieckiego Towarzystwa Statystycznego. Licznie reprezentowane były wszystkie środowiska statystyków polskich: z uniwersytetów, instytutów, organizacji naukowych, Głównego Urzędu Statystycznego, urzędów regionalnych i innych instytucji statystyki publicznej, przedstawiciele środowiska gospodarczego, administracji i przedsiębiorców oraz młodzież akademicka.

O randze kongresu świadczy jego wymiar merytoryczny wynikający z treści referatów wygłoszonych podczas trzech sesji plenarnych, panelu dyskusyjnego, 35 równoległych sesji tematycznych (w tym 12 w języku angielskim) oraz około 20 opracowań przygotowanych na sesję plakatową. Podczas kongresu wygłoszono łącznie 155 referatów, które zostały przygotowane przez blisko dwustu naukowców. Program obejmował szereg sesji tematycznych, w tym sesje poświęcone metodologii badań statystycznych, statystyce regionalnej, statystyce ludności, statystyce społecznej i gospodarczej, problematyce danych statystycznych oraz statystyce zdrowia, sportu i turystyki.

Sesje pierwszego dnia kongresu zatytułowano *Rozwój i dorobek polskiej myśli statystycznej*. W pierwszej z nich wybitni reprezentanci poszczególnych dziedzin badań statystycznych, omówili największe polskie osiągnięcia. Rozwój myśli statystycznej w Polsce w ujęciu historycznym przedstawił W. Ostasiewicz. O wkładzie Polaków w rozwój statystyki matematycznej i aplikacyjnej mówił J. Mielniczuk. Z kolei T. Caliński ukazał osiągnięcia polskich statystyków w biometrii. Rozwój statystyki w naukach ekonomicznych zaprezentował K. Jajuga. Natomiast rozwój historyczny i aktualne wyzwania stojące przed statystyką publiczną przedstawił J. Witkowski. Drugą sesję poświęcono prezentacji sylwetek dwóch najwybitniejszych polskich statystyków. T. Ledwina przygotowała prezentację życia i osiągnięć Jerzego Neymana. Natomiast M. Krzyśko przybliżył życiorys i dorobek Jana Czekanowskiego. Nestorka polskiej statystyki, S. Bartosiewicz przedstawiła (wspólnie z E. Stańczyk) historię społeczno-ekonomiczną Polski w latach 1918-2008, a J. Kordos omówił współzależności między rozwojem teorii i praktyki badań reprezentacyjnych w Polsce.

Bardzo ważnym wydarzeniem pierwszego dnia kongresu był panel dyskusyjny poświęcony problemom przyszłości statystyki nad którego wymiarem merytorycznym czuwał T. Caliński. W dyskusji udział wzięli profesorowie: J. Jóźwiak w zakresie demografii, J. Witkowski oraz T. Panek omówili problemy statystyki społecznej, J. Wilkin przedstawił potrzeby statystyki gospodarczej, J. Paradysz zwrócił uwagę na zagadnienia statystyki regionalnej, J. A. Moczko wskazał problemy statystyki zdrowia, M. Szreder skomentował dylematy jakości danych statystycznych, a J. Koronacki przedstawił perspektywy rozwoju metodologii badań statystycznych.

Obrady drugiego i trzeciego dnia kongresu były bardzo intensywne. Program obejmował szereg sesji tematycznych. Ich organizacji podjęli się uznani i cenieni w kraju oraz na arenie międzynarodowej statystycy. O wygłoszenie wiodących referatów po-

proszono najwybitniejszych polskich i zagranicznych przedstawicieli poszczególnych dziedzin badań statystycznych. Niniejszy tom specjalny *Przeglądu Statystycznego* prezentuje wiele z przygotowanych na kongres artykułów. Prezentacji dorobku kongresu poświęcony jest także numer specjalny *Studiów Demograficznych* (nr 1/2012). Natomiast *Wiadomości Statystyczne,* począwszy od numeru 7/2012, w dziale zatytułowanym *Z Obrad Kongresu Statystyki Polskiej* zamieszczają merytoryczne omówienie poszczególnych sesji oraz wybrane opracowania. W kongresie aktywnie uczestniczyło wielu znamienitych statystyków polskich i zagranicznych. Jego dorobek będziemy długo analizować i do niego się odwoływać. W tym miejscu prezentujemy tematykę poszczególnych sesji oraz referatów proszonych, natomiast szczegółowy program kongresu dostępny jest w Internecie pod adresem:
http://www.stat.gov.pl/pts/kongres2012/dok/ProgramPolski_18-04-2012.pdf.

Sesje tematyczne Kongresu Statystyki Polskiej, ich organizatorzy oraz tematyka referatów proszonych:

1. Statystyka matematyczna – prof. Mirosław Krzysko, Uniwersytet im. A. Mickiewicza w Poznaniu
   - Tadeusz Bednarski (Uniwersytet Wrocławski), Statystyczna analiza przyczyn obciążoności próby w badaniach sondażowych rynku pracy
   - Jacek Koronacki (Instytut Podstaw Informatyki PAN w Warszawie), Analiza danych o dużym wymiarze w przypadku małej liczności próby
   - Jana Jureckova (Charles University in Prague), Regression quantiles and their two-step modifications
   - Zbigniew Szkutnik (Akademia Górniczo-Hutnicza w Krakowie), Algorytm EM i jego modyfikacje
2. Metoda reprezentacyjna i statystyka małych obszarów – prof. Janusz Wywiał, Uniwersytet Ekonomiczny w Katowicach
   - Malay Ghosh (University of Florida), Finite population sampling: a model-design synthesis
   - Ray Chambers, Gunky Kim (Wollongong University), Regression analysis using data obtained by probability linking of multiple data sources
   - Li-Chun Zhang (Statistics Norway), Micro calibration for data integration
   - Lorenzo Fattorini (University of Siena), Design-Based Inference on Ecological Diversity
   - Jean-Claude Deville, Daniel Bonnéry, Guillaume Chauvet (ENSAE), Neyman type optimality for marginal quota sampling
3. Statystyka ludności – prof. Janina Jóźwiak, Szkoła Główna Handlowa w Warszawie
   - Nico Keilman (University of Oslo), Challenges for statistics on households and families
   - Irena E. Kotowska (Warsaw School of Economics), Evolving population structures and their relevance for demographic and social change
   - Francesco Billari (Bocconi University, Milan), Challenges for new methods in demographic analysis (tentative

4. Statystyka społeczna – prof. Tomasz Panek, Szkoła Główna Handlowa w Warszawie
   - Achille Lemmi (Siena University, Dipartimento di Economia Politica) Dimensions of poverty. Theory, models and new perspectives
   - Walenty Ostasiewicz (Uniwersytet Ekonomiczny we Wrocławiu), Jakość życia jako przedmiot badań statystycznych
   - Anne Valia Goujon, Ramon Bauer, Samir K.C., Michaela Potančoková (Vienna Institute of Demography (VID) of the Austrian Academy of Sciences), The human capital puzzle: Why it is so hard to find good data on educational attainment?
   - Urszula Sztanderska (Uniwersytet Warszawski), Polska statystyka publiczna jako inspiracja i bariera rozwoju badań rynku pracy
5. Statystyka gospodarcza – prof. Jerzy Wilkin, Uniwersytet Warszawski
   - Włodzimierz Siwiński (Uniwersytet Warszawski, Akademia L. Koźmińskiego), Statystyczny obraz kryzysu gospodarczego 2008-2010
   - Barbara Liberda (Uniwersytet Warszawski), Rachunki generacyjne metodą pomiaru bogactwa kolejnych pokoleń
   - Elżbieta Mączyńska (Instytut Nauk Ekonomicznych PAN, Polskie Towarzystwo Ekonomiczne), Statystyka jako źródło danych w badaniach naukowych. Dylematy i niedostosowania
   - Andrzej P. Wiatrak (Uniwersytet Warszawski), Aktualne problemy statystyki rolniczej
6. Statystyka regionalna – prof. Tadeusz Borys, Uniwersytet Ekonomiczny we Wrocławiu
   - Jan Paradysz (Centrum Statystyki Regionalnej, Uniwersytet Ekonomiczny w Poznaniu), Statystyka regionalna: stan, problemy i kierunki rozwoju
   - Tadeusz Borys (Uniwersytet Ekonomiczny we Wrocławiu), Tomasz Potkański (Związek Miast Polskich), Monitorowanie rozwoju regionalnego i lokalnego
7. Analiza i klasyfikacja danych – prof. Krzysztof Jajuga, prof. Marek Walesiak, Uniwersytet Ekonomiczny we Wrocławiu
   - Reinhold Decker (Universität Bielefeld), Model-based analysis of online consumer reviews – methods and applications
   - Wojtek Krzanowski (University of Exeter), Classification: some old principles applied to new problems
   - Patrick Groenen (Erasmus University Rotterdam), The support vector machine as a powerful tool for binary classification
8. Dane statystyczne – prof. Mirosław Szreder, Uniwersytet Gdański
9. Statystyka zdrowia – prof. Jerzy Andrzej Moczko, Uniwersytet Medyczny w Poznaniu
10. Statystyka sportu i turystyki – prof. Bogdan Sojkin, Uniwersytet Ekonomiczny w Poznaniu

Z okazji kongresu przygotowano szereg publikacji okolicznościowych, wystaw oraz imprez i wydarzeń towarzyszących. Okazją do uhonorowania najbardziej zasłużonych członków oraz przybliżenia działań i aktywności podejmowanych w poszczególnych regionach, było uroczyste, otwarte posiedzenie Rady Głównej PTS. Odznakę honorową "Za zasługi dla statystyki RP" z rąk prof. J. Witkowskiego – Prezesa GUS otrzymali: M. Krzyśko, J. Pociecha, B. Podolec, A.S. Barczak, A. Młodak oraz T. Klimanek. Medal im. Jerzego Spławy-Neymana prof. C. Domański – Prezes PTS wręczył profesorom: T. Calińskiemu (Uniwersytet Przyrodniczy w Poznaniu), W. Krzanowskiemu (University of Exeter), J. Jurečkovej (Uniwersytet Karola w Pradze), M. Ghosh (University of Florida), A. Atkinsonowi (London School of Economics). Rada Główna PTS wyróżniła dra K. Kruszkę tytułem Honorowego Członka Polskiego Towarzystwa Statystycznego.

Z okazji kongresu nakładem ZWS GUS wydano dwie publikacje okolicznościowe. W książce „Statystycy polscy" przedstawiono biogramy 85 statystyków polskich, którzy odegrali znaczącą rolę w historii statystyki polskiej. Praca przygotowana została przez zespół w składzie: W. Adamczewski, J. Berger, K. Kruszka, M. Krzyśko (przewodniczący), B. Łazowska. Natomiast praca „Polskie Towarzystwo Statystyczne 1912-2012" zawiera szczegółowe informacje dotyczące rozwoju organizacyjnego Polskiego Towarzystwa Statystycznego oraz opis działalności statutowej Towarzystwa. Publikację przygotował zespół w składzie: J. Berger, J. Kordos, K. Kruszka (przewodniczący), W.W. Łagodziński, W. Ostasiewicz. Kongresowi towarzyszyły także dwie wystawy okolicznościowe: „Polskie Towarzystwo Statystyczne w latach 1912-2012" oraz „Statystyka w Wielkopolsce". Przybliżyły one kilkuwiekowy dorobek statystyki, prezentowały archiwalne wyniki badań i publikacje, które wniosły istotny wkład w rozwój myśli i praktyki statystycznej w Polsce oraz w Wielkopolsce. Przypomniano także sylwetki osób tworzących ten dorobek. Obie wystawy dostępne są w wersji elektronicznej. Przygotowana przez J. Bergera wystawa „Polskie Towarzystwo Statystyczne w latach 1912-2012" dostępna jest stronie Internetowej kongresu http://www.stat.gov.pl/pts/kongres2012/dok/Polskie_Towarzystwo_Statystyczne.pdf.

Przygotowana w Urzędzie Statystycznym w Poznaniu wystawa „Statystyka w Wielkopolsce", prezentowana była także w holu Urzędu Miasta Poznania i Wielkopolskiego Urzędu Wojewódzkiego oraz w gmachu Narodowego Banku Polskiego – Oddział Okręgowy w Poznaniu. Elektroniczna wersja wystawy służyć ma jeszcze szerszemu jej upowszechnieniu http://www.stat.gov.pl/pts/kongres2012/dok/statystyka_w_wielkopolsce.pdf.

W ramach imprez towarzyszących kongresowi, w dniach 16-17 kwietnia 2012 r. odbyły się dwa warsztaty, w których uczestniczyło łącznie około 60 osób. Tematyka warsztatów dotyczyła najważniejszych zagadnień metodologicznych: integracji danych statystycznych i estymacji dla małych obszarów (prowadzone przez specjalistów z Włoch) oraz z metod data miningu (przygotowane przez firmę Statsoft). Podczas obrad kongresu zorganizowano także uroczyste zakończenie wielkopolskiego konkursu „Statystyka mnie dotyka" z okazji Dnia Statystyki Polskiej w 2012 r. Konkurs wiedzy statystycznej dla młodzieży szkół ponadgimnazjalnych organizowany jest przez

Urząd Statystyczny w Poznaniu oraz Polskie Towarzystwo Statystyczne. W programie kulturalnym i turystycznym kongresu znalazły się miedzy innymi, spacer po Starym Mieście z koncertem w Kościele Farnym, zwiedzanie Centrum Sztuki i Biznesu – Stary Browar, Welcome Party w zabytkowej Słodowni, Koncert Jubileuszowy „Piękne głosy, piękne melodie" w wykonaniu studentów i absolwentów Akademii Muzycznej w Poznaniu oraz chóru kameralnego Uniwersytetu Ekonomicznego *Musica Viva* pod dyrekcją prof. M. Gandeckiego. Koncert przygotowała i prowadziła Pani prof. H. Lorkowska, Prorektor Akademii Muzycznej w Poznaniu.

Uroczyste obchody stulecia Polskiego Towarzystwa Statystycznego zorganizowane zostały w Poznaniu jako wspólne przedsięwzięcie całego środowiska statystyków reprezentowanego przez cztery instytucje: Polskie Towarzystwo Statystyczne, Główny Urząd Statystyczny, Urząd Statystyczny w Poznaniu oraz Uniwersytet Ekonomiczny w Poznaniu. W ten sposób, nadano kongresowi wyjątkowy charakter w dziejach polskiej statystyki, jako wydarzenia łączącego środowiska reprezentujące zarówno praktyczny jak i teoretyczny wymiar badań statystycznych, będącego okazją do spotkania oraz wymiany poglądów i doświadczeń przedstawicieli statystyki publicznej, ośrodków naukowych, władz administracyjnych, przedsiębiorców i innych partnerów uczestniczących w badaniu procesów gospodarczych, społecznych i demograficznych. Kongres stworzył niepowtarzalną platformę dla ukierunkowania poszukiwań metodologicznych i poznawczych podejmowanych przez środowiska statystyczne w nadchodzących latach, które zostały ujęte w Uchwale Kongresu. Już dzisiaj można powiedzieć, że Kongres był jednym z ważniejszych i bardzo znaczących wydarzeń naukowych w historii polskiej myśli i praktyki statystycznej.

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

CZESŁAW DOMAŃSKI

LAUREACI MEDALU IM. JERZEGO SPŁAWY-NEYMANA

1. WPROWADZENIE

W okresie 100-letniej działalności Polskiego Towarzystwa Statystycznego wielu przedstawicieli różnych dyscyplin naukowych, w swych badaniach na szerszą skalę wykorzystywali metody statystyczne, przyczyniając się jednocześnie do ich rozwoju. Wzmogły się w tym okresie wykłady ze statystyki na większości fakultetach w uczelniach wyższych, zwłaszcza wydziałach ekonomicznych i przyrodniczych. Przeważała tam tematyka związana z oceną i analizą społeczno-gospodarczą oraz biostatystyką. Równocześnie z tym kierunkiem wykładano również statystykę matematyczną z szeroko rozumianym wnioskowaniem statystycznym, której twórcą był Jerzy Spława-Neyman.

Polscy uczeni w każdym okresie, począwszy od XV wieku od Jana Długosza, wnosili istotny wkład w rozwój podstawowych pojęć i teorii współczesnej dyscypliny, którą nazywamy statystyką.

Niezwykle szeroki i różnorodny dorobek przedstawionych statystyków zawiera pionierskie prace z zakresu statystyki. Wiele prac nie ujrzało światła dziennego z wielką szkodą dla rozwoju myśli i teorii statystyki. Prezentowane osiągnięcia statystyków polskich pozwalają na jednoznaczne określenie rodowodu statystyki, a przede wszystkim wskazanie zaszczytnego miejsca polskich autorów i uczonych w rozwoju statystyki jako dyscypliny naukowej. Dziedzina badań teoretycznych i metodologicznych w Polsce charakteryzowała się wyjątkowo wysoką aktywnością zarówno w okresie rozbiorów, międzywojennym, jak i obecnie. Osiągnięcia statystyków polskich oraz Polskiego Towarzystwa Statystycznego prezentowane są w dwóch monografiach wydanych z okazji Kongresu Statystyki Polskiej związanego z obchodami 100-lecia Polskiego Towarzystwa Statystycznego, mianowicie: „Statystycy Polscy", Zakład Wydawnictw Statystycznych, Warszawa 2012 oraz „Polskie Towarzystwo Statystyczne 1912-2012" Zakład Wydawnictw Statystycznych, Warszawa 2012.

2. PREZENTACJA POSTACI UHONOROWANYCH MEDALEM IM. JERZEGO SPŁAWY-NEYMANA

Podczas Kongresu wręczono medal im Jerzego Spławy-Neymana przyznany przez Kapitułę Polskiego Towarzystwa Statystycznego. Medalem zostali uhonorowani profesorowie: Anthony Atkinson, Tadeusz Caliński, Malay Ghosh, Zdzisław Hellwig, Jana

Jureckova, Wojtek Krzanowski, Kazimierz Zając, Ryszard Zieliński oraz Witold Klonecki.

Przedstawiamy krótką prezentację laureatów medalu Jerzego Spławy-Neymana.

Sir Anthony Barnes Atkinson urodził się 4 września 1944 roku. Specjalizuje się w dystrybucji dochodów i nierówności społecznej. Miara nierówności została nazwana jego imieniem: indeks Atkinsona.

Studiował na Uniwersytecie w Cambridge, gdzie uzyskał tytuł magistra w 1966 roku. W latach 1967-1971 był członkiem Kolegium świętego Jana w Cambridge, następnie profesorem ekonomii na Uniwersytecie w Essen (1971-1976), potem profesorem ekonomii na University College London (1976-1979) oraz profesorem ekonomii w Cambridge i Fellow Churchill College(1976-1994). W okresie 1994-2005 Prezes (Warden), Nuffield College w Oksfordzie, gdzie jest aktualnie Senior Research Fellow. Pracował w charakterze doradczym w różnych komitetach zarówno Brytyjskich, Rządu francuskiego i Unii Europejskiej. W 2001 r. został mianowany Kawalerem Francuskiej Legii Honorowej.

Sir Anthony Barnes Atkinson jest autorem i redaktorem wielu monografii. Opublikował m.in. ponad 140 artykułów w renomowanych czasopismach ekonomicznych i statystycznych. Najważniejsze publikacje Sir Atkinsona: Poverty in Britain and the Reform of Social Security (1969), Unequal Shares – Wealth in Britain (1972), Economics of Inequality (1975, 1983), The Distribution of Personal Wealth in Britain (con A. J. Harrison 1978), Lectures on Public Economics (con J. E. Stiglitz 1980), Social Justice and Public Policy (1982), Parents and Children (con A. K. Maynard e C. G. Trinder 1983), Poverty and Social Security (1989), Empirical Studies of Earnings Mobility (con F. Bourguignon e C. Morrisson 1992), Economic Transformation in Eastern Europe and the Distribution of Income (con J. Micklewright 1992), Public Economics in Action (1995), Income Distribution in OECD Countries (con L. Rainwater e T. Smeeding 1995), Incomes and the Welfare State (1996), Poverty in Europe (1998), The Economic Consequences of Rolling Back the Welfare State (1999), Social Indicators (con B. Cantillon, E. Marlier e B. Nolan 2002).

Prof. Tadeusz Caliński urodził się 4 września 1928 roku w Poznaniu. Po zdaniu matury w 1948 r. w Liceum im. Karola Marcinkowskiego rozpoczął studia rolnicze, które ukończył w 1953 r. otrzymując tytuł inżyniera magistra rolnictwa na Wydziale Rolniczym Wyższej Szkoły Rolniczej w Poznaniu. Studia te uzupełnił trzyletnimi studiami matematycznymi na Wydziale Matematyki, Fizyki i Chemii Uniwersytetu im. Adama Mickiewicza w Poznaniu. W roku 1961 uzyskał stopień doktora nauk rolno-leśnych, nadany przez Radę Wydziału Rolniczego Wyższej Szkoły Rolniczej w Poznaniu, na podstawie pracy wykonanej pod kierunkiem profesora Stefana Barbackiego. Stopień naukowy doktora habilitowanego, w zakresie doświadczalnictwa rolniczego i biometrii, otrzymał od tejże Rady w roku 1966 na podstawie rozprawy opublikowanej w Journal of the Royal Statistical Society.

W latach 1953-1988 pracował w Katedrze Metod Matematycznych i Statystycznych Wyższej Szkoły Rolniczej w Poznaniu (od 1972 r. zmiana nazwy na Akademia

Rolnicza), będąc jej kierownikiem w okresie 1968-1984. W 1988 roku przeszedł na emeryturę i aktualnie jest emerytowanym profesorem Uniwersytetu Przyrodniczego w Poznaniu. W 1998 roku uzyskał tytuł doktora Honoris Causa Uniwersytetu Przyrodniczego w Poznaniu, a w 2007 roku tytuł doktora Honoris Causa Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie.

Do głównych osiągnięć prof. Tadeusza Calińskiego należy zaliczyć: stworzenie szkoły statystyki matematycznej i biometrii, znanej w kraju i na świecie; wypromowanie 24 doktorów, z których 11 uzyskało tytuł profesora; opublikowanie ponad 150 prac w znaczących czasopismach zagranicznych i polskich, w tym 2 pozycji książkowych opublikowanych przez wiodące wydawnictwo naukowe Springer Verlag New York; popularyzowanie statystyki matematycznej i biometrii wśród matematyków poprzez zainicjowanie powstania, a następnie kierowanie Komisją Statystyki Matematycznej przy Komitecie Nauk Matematycznych Polskiej Akademii Nauk; popularyzowanie metod statystycznych i biometrycznych wśród studentów i pracowników naukowych z różnych dziedzin nauki poprzez wykłady i konsultacje; inicjacja i aktywne wspieranie międzynarodowej współpracy między polskimi statystykami i ich kolegami z różnych krajów świata, w szczególności z Holandii, Francji, Włoch, Wielkiej Brytanii, Niemiec, Japonii i Portugalii.

Można stwierdzić, że w swym dorobku naukowym w statystyce matematycznej, biometrii i doświadczalnictwie jest niestrudzonym kontynuatorem myśli naukowej zapoczątkowanej w Polsce przez Jerzego Spławę-Neymana. Obecnie Prof. Tadeusz Caliński jest emerytowanym profesor zwyczajnym Uniwersytetu Przyrodniczego w Poznaniu.

Profesor Malay Ghosh urodził się 15 kwietnia 1944 r. w Kalkucie w Indiach. Ukończył studia I i II stopnia w zakresie statystyki odpowiednio w 1962 i 1964 w Calcutta University w Indiach oraz w 1969 roku studia doktorskie w zakresie statystyki w University of North Carolina, USA. Aktualnie pracuje w Department of Statistics, University of Floryda, USA, gdzie aktualnie zajmuje stanowisko profesora wybitnego (distinguished professor).

Profesor Ghosh jest znanym w światowym środowisku statystyków specjalistą z zakresu statystyki, a w szczególności wnioskowania bayesowskiego, zwłaszcza subdziedziną zwaną empirical Bayes. Zajmuje się również tzw. statystyką małych obszarów. Ponadto, trzykrotnie brał udział w konferencji Survey Sampling in Economic and Social Research w Katowicach w latach 2011, 2010, 2008, jako gość z zaproszonym referatem (invited speaker). Jego osoba przyciągała innych uczestników konferencji z zagranicy. Wiele osób korzystało z jego cennych konsultacji i opinii o pracach. Dorobek naukowo-badawczy profesora Malaya Ghosha obejmuje kilka książek i około 300 artykułów. Jego prace charakteryzują się niezwykłą oryginalnością oraz wysokim poziomem naukowym. Profesor jest promotorem około 40 dysertacji. Oprócz pracy w University of Floryda Malay Ghosh był profesorem wizytującym wielu innych renomowanych uniwersytetów. Kierował i kieruje wieloma grantami naukowo-badawczymi dotyczących zarówno teorii statystyki, jak i jej zastosowaniem. Jest członkiem komite-

tów redakcyjnych wielu renomowanych czasopism oraz członkiem honorowym wielu znaczących organizacji i stowarzyszeń zawodowych.

Do najważniejszych publikacji Prof. Ghosh'a można zaliczyć: Small Area Estimation: An Appraisal (with J.N.K. Rao) (1994) Statistical Science, V9, No. 1; Second Order Probability Matching Priors (with R. Mukerjee) (1997) Biometrika, V84, No. 4; Generalized Linear Models for Small Area Estimation (with K. Natrarajan, T.W.F. Stroud and B. Carlin) (1998) Journal of the American Statistical Association, V93, No. 1.

Prof. Zdzisław Hellwig urodził się w 1925 r. w Dokszycach. Był żołnierzem IV brygady ułanów Armii Krajowej. W 1946 r. rozpoczął studia ekonomiczne w Wyższej Szkole Handlowej we Wrocławiu. Tytuł magistra uzyskał w 1952 r. w SGPiS (Szkoła Główna Planowania i Statystyki) w Warszawie. Tytuł profesora zwyczajnego otrzymał w 1972 r. Pełnił wiele funkcji; był m.in. kierownikiem Katedry Statystyki w WSE i AE w latach 1962-1995, dyrektorem Instytutu i kierownikiem Katedry Statystyki i Cybernetyki Ekonomicznej w latach 1969-1995, prorektorem ds. nauki AE w latach 1962-1965, 1968-1970 oraz 1979-1981, a także ekspertem UNESCO w Paryżu 1968-1974. Profesor wykładał w Nigerii na Uniwersytecie w Ibadanie w roku 1965/1966 i Tohoku University w Japonii w 1997 r. Był członkiem Centralnej Komisji Kwalifikacyjnej do Spraw Kadr Naukowych. Profesor Zdzisław Hellwig jest twórcą szkoły naukowej statystyki, ekonometrii i cybernetyki. Wypromował 34 doktorów, spośród których 15 osób uzyskało tytuł naukowy profesora. W 1996 r. otrzymał prestiżową nagrodę naukową Prezesa Rady Ministrów za wybitny dorobek naukowy. W uznaniu zasług naukowych został wyróżniony doktoratami honorowymi Akademii Ekonomicznej w Krakowie (1985 r.) oraz Uniwersytetu Karola w Pradze ze szczególnym podkreśleniem zasług dla rozwoju ekonomii i rozwijaniu kontaktów międzynarodowych w nauce (1994 r.). W statystyce opracował w 1969 roku m.in. Metodę Hellwiga, zwaną również metodą optymalnego wyboru predyktant, metodą wskaźników pojemności informacji – formalna metoda doboru zmiennych objaśniających do modelu statystycznego.

Prof. Zdzisław Hellwig jest autorem kilkuset publikacji i prac naukowych, do najważniejszych należą: Elementy rachunku prawdopodobieństwa i statystyki matematycznej (1959), Regresja liniowa i jej zastosowania w ekonomii (1960), Przyczynek do teorii regresji (Zeszyty Naukowe WSE, Wrocław 1961), Linear Regression and Its Application to Economics (Pergamon Press, Oxford, 1963), Aproksymacja stochastyczna (1965), Problem optymalnego wyboru predykant (Przegląd Statystyczny Nr 3-4, 1968), On the measurement of stochastical dependence (Zastosowanie Matematyki Vol. 10, 1969), Efekt katalizy jego wykrywanie i usuwanie (Przegląd Statystyczny Nr. 2, 1977), Elementy rachunku ekonomicznego (1976), Przechodność skorelowania między zmiennymi losowymi i płynące stąd wnioski ekonometryczne (Przegląd Statystyczny Nr 1, 1976).

Prof. Jana Jurečková urodziła się 20 września 1940 w Pradze. Tytuł magistra uzyskała w Uniwersytecie Karola w Pradze w 1962 roku. Następnie zdobyła tytuł doktora statystyki w 1967 roku w Czechosłowackiej Akademii Nauk. Profesor Jana

Jurečková pracę w Katedrze Prawdopodobieństwa i Statystyki Matematycznej na Wydziale Matematyki i Fizyki Uniwersytetu Karola w Pradze rozpoczęła w 1964 roku. W 1984 uzyskała stopień naukowy DrSc., a w 1992 roku została mianowana przez prezydenta Czechosłowacji do rangi profesorów tytularnych. Aktualnie pracuje w Jaroslav H'ajek Centrum Statystyki Teoretycznej i Stosowanej. Prof. Jana Jurečková jest autorką wielu publikacji (ponad 120 artykułów naukowych), w większości w czasopismach o statystyce i teorii prawdopodobieństwa i współautorem kilku monografii. Jej zainteresowania badawcze obejmują szeroki obszar wnioskowania statystycznego m.in.: procedury statystyczne oparte na szeregach, szczegółowe procedury statystyczne oparte na tak zwanych M-statystykach i L-statystykach, procedury statystyczne oparte na skrajnych obserwacjach próbek, asymptotyczne metody statystyki matematycznej, szacowanie parametrów dla próbki skończonej.

Osiągnięcia Profesor Jurečkovej przyczyniły się znacząco do rozwoju statystyki i teorii prawdopodobieństwa. Działalność zawodowa Prof. Jany Jurečkovej kształtuje się następująco: od 2003 r. jest członkiem Towarzystwa Naukowego w Republice Czeskiej, jest członkiem ISI (Międzynarodowego Instytutu Statystycznego), współpracuje z IMS (Instytut Statystyki Matematycznej). Była członkiem redakcji następujących czasopism: Annals of Statistics w latach 1979-1980 i 1992-1997, Journal of American Statistical Association (2005-2008), Sankhya (2006-2011).

Wybrane publikacje Prof. Jurečkovej: Robust Statistical Procedures: Asymptotics and Interrelations, J. Wiley, New York 1996; Adaptive Regression, Springer, New York 2000; Robust Statistical Methods with R, Chapman & Hall/CRC, Boca Raton 2006; Methodology in Robust and Nonparametric Statistics, Chapman & Hall/CRC, Boca Raton, London 2012.

Prof. Wojtek Janusz Krzanowski w 1967 roku w University of Leeds uzyskał tytuł B. Sc., a w 1968 Dip. Math. Stat. w University of Cambridge. Stopień Ph. D. otrzymał w 1974 roku w University of Reading. Jego praca naukowa realizowana była w następujących jednostkach naukowo-badawczych: Rothamsted Experimental Station, Harpenden, Wielka Brytania (Scientific Officer 1968-1971); RAF Institute of Aviation Medicine, Farnborough, Wielka Brytania (Senior Research Fellow 1971-1974); University of Reading, Wielka Brytania (Lecturer 1974-1980, Senior Lecturer 1980-1987, Reader 1987-1990); University of Exeter, Department of Mathematical Sciences, Exeter, Wielka Brytania Professor, Dziekan Wydziału Matematyki Uniwersytetu w Exeter 1998-2002; Redaktor Journal of the Royal Statistical Society, Series C (1991-1995); członek redakcji Journal of Classification (1984-); członek redakcji Journal of the Royal Statistical Society, Series B (1990-1992); członek następujących instytucji: Biometric Society, Classification Society, International Statistical Institute, Polskim Towarzystwie Biometrycznym, Royal Statistical Society.

Prof. Wojtek Krzanowski jest autorem 5 książek i ponad 100 artykułów w czasopismach naukowych oraz rozdziałów w książkach, m.in.: "Principles of Multivariate Analysis: a User's Perspective", Oxford University Press, 1988, 2000; "Multivariate Analysis Part 1: Distributions, Ordination and Inference"; "Multivariate Analysis

Part 2: Classification, Covariance Structures and Repeated Measurements" (współaut. Edward Arnold 1994/1995); "Recent Advances in Descriptive Multivariate Analysis", University Press, Oxford 1995; "An Introduction to Statistical Modeling, Edward Arnold, 1998.

Prof. Wojtek Krzanowski od 2005 roku jest emerytowanym profesorem statystyki w Exeter.

Prof. Kazimierz Zając urodził się 20 września 1916 roku w Krośnie. W czasie II wojny światowej prowadził tajne nauczanie i uczestniczył w pracach Polskiego Czerwonego Krzyża, organizując pomoc na rzecz oficerów polskich przebywających w obozach jenieckich. Został zaprzysiężony jako żołnierz Armii Krajowej – Obwód Krosno, pod pseudonimem „Konrad". Pełnił funkcję adiutanta dowódcy placówki Centrum Krosno ppor. Moskala (pseudonim „Zrąb"). W latach 1938/39 oraz 1945-1945/46 studiował w Akademii Handlowej w Krakowie na kierunku konsularnym. W latach 1947-1948 odbywał studia doktoranckie w Szkole Głównej Handlowej w Warszawie, uczestnicząc w seminarium doktorskim z ekonomii u prof. Edwarda Lipińskiego. Stopień kandydata nauk ekonomicznych (był to ówczesny odpowiednik doktora nauk ekonomicznych) otrzymał 26 czerwca 1958 r. W 1965 r. został kierownikiem Katedry Statystyki, którą kierował aż do przejścia na emeryturę w 1986 roku. Tytuł naukowy profesora nadzwyczajnego otrzymał w roku 1971, a tytuł naukowy profesora zwyczajnego w roku 1976.

Prof. Kazimierz Zając jest pionierem badań statystyczno-ekonometrycznych dochodów i wydatków ludności oraz płac w Polsce, a podstawowe wyniki prezentowane są m.in. w książkach: „Ekonometryczne metody ustalania rejonów konsumpcyjnych" (współautor: B. Podolec, 1978), „Metody badania usług rynkowych", (praca zbiorowa pod red. K. Zająca, 1982). Jego drugim nurtem badawczym były badania demograficzne. Z połączenia badań demograficznych z badaniami społeczno-ekonomicznymi wywodzi się monografia „Rozwój demograficzny a rozwój gospodarczy" (współautor: A. Sokołowski,1987), „Metody taksonomiczne w badaniach społeczno-ekonomicznych" (współautorzy: J. Pociecha, B. Podolec, A. Sokołowski; 1988). Z dziedziny statystycznych metod kontroli jakości opublikował szereg artykułów naukowych oraz książkę „Statystyczne metody kontroli jakości" (współautorzy: J. Cyran, J. Steczkowski; 1973). W latach 1965-1995 był członkiem Rady Naukowej GUS, przyczyniając się do wytyczania kierunków badań statystycznych oraz rozwoju metodologii i praktyki statystyki publicznej w Polsce oraz przez wiele lat zasiadał w Radzie Naukowej „Wiadomości Statystycznych". W latach 1980-1986 pełnił funkcję dziekana Wydziału Ekonomiki Obrotu Akademii Ekonomicznej w Krakowie. Przez kilka kadencji był członkiem Rady Głównej Polskiego Towarzystwa Statystycznego oraz przewodniczącym Rady Oddziału PTS w Krakowie. W roku 2000 wyróżniony został godnością Członka Honorowego PTS.

Prof. Kazimierz Zając wypromował 38 doktorów i sprawował opiekę merytoryczną nad 22 habilitantami. Jest autorem lub współautorem wielu skryptów i podręczników do statystyki. Wśród nich wyróżnia się podręcznik *Zarys metod statystycznych* (PWE,

Warszawa), mający 5 wydań, na którym wychowało się wiele roczników studentów krakowskich i innych uczelni polskich.

Dla polskich statystyków i ekonometryków Profesor Zając istnieje „od zawsze". W swoim długim życiu wnosi dla kolejnych pokoleń pracowników naukowo-dydaktycznych polskich uczelni ekonomicznych nie tylko swoją wiedzę i doświadczenie naukowe, inspirujące do podejmowania interesujących i ważnych tematów badań. Swoim osobistym przykładem życzliwości dla innych, otwartości i akceptacji odmiennych poglądów oraz podejść naukowych stwarzał standardy postępowania w życiu naukowym polskich uczelni.

Profesor Kazimierz Zając zmarł 6 maja 2012 r.

Profesor Ryszard Zieliński urodził się 1 lipca 1932 roku w Warszawie. W 1955 roku zdobył tytuł magistra w Szkole Głównej Planowania i Statystyki na Wydziale Finansów i Statystyki w Warszawie. Studia te ukończył w 1964 roku z dyplomem magistra matematyki w Uniwersytecie Warszawskim. Tytuł profesora nauk matematycznych uzyskał w 1988 roku.

W jego bogatym dorobku naukowym można znaleźć prace przede wszystkim z zakresu statystyki matematycznej i takich jej działów jak statystyczna kontrola jakości, statystyka asymptotyczna, statystyka odporna, statystyka nieparametryczna oraz generatory liczb losowych. Wśród wielu Jego osiągnięć należy zwrócić uwagę na zaproponowaną przez Niego definicję odporności statystyk przedstawioną w pracy z 1983 roku (Robust statistical procedures: a general approach, Lecture Notes in Mathematics 982, "Stability problems for stochastic models" Eds. V. V. Kalashnikov and V. M. Zolotarev, Springer-Verlag 1983, 283-295). Jego nowatorskie podejście dało impuls do rozwoju odpornych metod estymacji i weryfikacji hipotez statystycznych. Profesor R. Zieliński zajmował się również nieparametryczną estymacją kwantyli rozkładów prawdopodobieństwa oraz szeroko rozumianym prawdopodobieństwem sukcesu. Jego prace znajdują zastosowania m.in. w problematyce estymacji wartości narażonej na ryzyko (V@R).

Profesor Ryszard Zieliński opublikował łącznie ponad 100 oryginalnych prac naukowych (pierwszą w 1956 roku, ostatnia ukazała się w 2012 roku). W kilkudziesięciu pracach popularno-naukowych w bardzo przystępny sposób prezentował metody statystyczne oraz ich zastosowania. Jego podręcznik do statystyki dla szkół średnich (Rachunek prawdopodobieństwa z elementami statystyki matematycznej) był wielokrotnie wznawiany i do dzisiaj jest wykorzystywany w dydaktyce przedmiotu. Napisał także bardzo ceniony podręcznik z zaawansowanej statystyki matematycznej (Siedem wykładów wprowadzających do statystyki matematycznej. Biblioteka Matematyczna Tom 72. PWN, Warszawa 1990). Był też tłumaczem kilku książek z angielskiego, rosyjskiego i niemieckiego. Spośród nich należy wymienić książkę C. R. Rao, Modele liniowe statystyki matematycznej, PWN, Warszawa, 1982. Autor kilku monografii poświęconych metodom Monte Carlo oraz generatorom liczb losowych. Część z nich tłumaczona była na język niemiecki. Ponadto Autor bardzo szeroko wykorzystywanych Tablic Statystycznych (PWN 1972, 1990).

Profesor Ryszard Zieliński zmarł nieoczekiwanie 30 kwietnia 2012.

Prof. Witold Klonecki był profesorem w Instytucie Matematyki i Informatyki Politechniki Wrocławskiej, wybitnym specjalistą w zakresie statystyki matematycznej i zastosowań matematyki w genetyce, zasłużonym nauczycielem wielu pokoleń matematyków, autor poczytnego skryptu „Statystyka dla inżynierów" PWN 1999.

Profesor Witold Klonecki wspólnie z Profesorem Kazimierzem Urbanikiem oraz Profesorem Czesławem Ryll-Nardzewskim, był współzałożycielem czasopisma o randze międzynarodowej pt. „Probability and Mathematical Statistics".

Inicjator corocznych konferencji statystycznych w Polsce z udziałem gości zagranicznych. Utrzymywał ścisłe kontakty z Profesorem Jerzym Spławą-Neymanem oraz z Departamentem Statystyki w Berkeley (Kalifornia). Był współorganizatorem konferencji poświęconych pamięci Jerzego Spławy-Neymana z okazji 100-lecia urodzin Jerzego Spławy-Neymana zorganizowanej przez Polskie Towarzystwo Statystyczne w 1994 roku. Był wychowawcą wielu profesorów Statystyki, którzy prowadzili i prowadzą swe badania zarówno w kraju jak i za granicą.

Prof. Witold Klonecki zmarł 10 sierpnia 2012 r., pozostanie w naszej pamięci jako polski kontynuator Szkoły Statystycznej Jerzego Spławy-Neymana, legendarnego współtwórcy współczesnej Statystyki.

Przyznanie Medalu im. Jerzego Spławy-Neymana jest wyrazem czci i wdzięczności całego polskiego środowiska naukowego statystyków polskich za wybitne osiągnięcia naukowe, dydaktyczne oraz wysokie zaangażowanie w kształcenie młodych statystyków w Polsce i za granicą.

Kapituła medalu wyróżniła jeszcze dwie osoby, mianowicie: Prof. Grahama Kaltona oraz Prof. Calyampudi Radhakrishna Rao, którym nie przekazała odznaczenia ze względów organizacyjnych, stąd też, nie prezentujemy ich sylwetek, a dokonamy tego po wręczeniu medalu.

UCHWAŁA KONGRESU STATYSTYKI POLSKIEJ

I. Uczestnicy Kongresu Statystyki Polskiej zebrani w Poznaniu na Jubileuszu 100-lecia Polskiego Towarzystwa Statystycznego

- oddają cześć założycielom i członkom Towarzystwa, którego działalność w latach 1912-1939 walnie przyczyniła się do powstania i rozwoju II Rzeczypospolitej;
- składają hołd członkom PTS, ofiarom terroru w czasie II wojny światowej i represji po jej zakończeniu;
- wyrażają uznanie członkom i władzom Towarzystwa z lat 1947-1955, kontynuatorom jego misji i dorobku mimo trudności w następnych latach, inicjatorom reaktywacji PTS w roku 1981, a także działającym w późniejszym okresie członkom, Radom Oddziałów i Radzie Głównej PTS – za twórczy wkład w rozwój Towarzystwa i skarbnicy wiedzy statystycznej.

II. Zwracając się ku przyszłości uczestnicy Kongresu Statystyki Polskiej uznają, że Polska u progu XXI staje przed poważnymi problemami rozwojowymi związanymi m.in. ze starzeniem się społeczeństwa i nasilającymi się procesami migracyjnymi. Sprostanie wynikającym stąd wyzwaniom wymaga podejmowania przez organy państwa właściwych decyzji w wielu sferach życia społecznego i gospodarczego, do czego niezbędne są rzetelne dane statystyczne.

Za szczególnie istotne uczestnicy Kongresu uważają podjęcie i rozwiązanie następujących zagadnień:

- opracowanie analiz ostrzegawczych dotyczących niekorzystnych tendencji w przebiegu procesów demograficznych, gospodarczych i społecznych oraz wskazywanie możliwości ich przezwyciężenia;
- rozwijanie powszechnej edukacji statystycznej całego społeczeństwa;
- likwidowanie „białych plam" w systemie statystyki publicznej przez powierzenie Głównemu Urzędowi Statystycznemu funkcji współgestora wszystkich rejestrów administracyjnych, koordynującego zakres i sposób gromadzenia danych statystycznych przez poszczególne resorty i instytucje publiczne;

- umożliwienie w celach badawczych niekomercyjnego dostępu do nieidentyfikowalnych danych zawartych w rejestrach urzędowych oraz w bankach danych gromadzonych przez statystykę publiczną;
- stworzenie jak najlepszych warunków dla dalszego rozwoju badań statystycznych – zarówno praktycznych, jak i dotyczących ich podstaw metodologicznych.

III. Uczestnicy Kongresu uznają, że jego przebieg przyczynił się do:
- prezentacji dorobku w zakresie teorii i praktyki statystycznej oraz wskazania kierunków jego rozwoju;
- postawienia właściwych diagnoz wielu procesów społeczno-gospodarczych, co daje podstawę do budowy użytecznych prognoz tych procesów;
- integracji środowiska statystyków zajmujących się zagadnieniami teoretycznymi i praktycznymi;
- zwiększenia nacisku na konieczność intensywnego kształcenia statystyków i upowszechniania wiedzy statystycznej;
- zwiększenia zainteresowania władz państwowych wynikami prac całego środowiska statystyków polskich;
- lepszego pokazania dorobku statystyki polskiej, w tym Polskiego Towarzystwa Statystycznego, na forum międzynarodowym.

Jednocześnie uczestnicy Kongresu pragną podkreślić, że doniosłość dotychczasowych dokonań statystyków polskich jest wielkim wyzwaniem i zobowiązaniem do kontynuowania tego dzieła w XXI wieku.

Poznań, 20 kwietnia 2012 roku

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

WALENTY OSTASIEWICZ

ROZWÓJ MYŚLI STATYSTYCZNEJ W POLSCE W XIX WIEKU

1. RODOWÓD STATYSTYKI

To co dzisiaj nazywamy statystyką ma dość skomplikowaną historię. Mimo iż oficjalna, udokumentowana na piśmie, historia dyscypliny o tej nazwie jest niezbyt długa, to jednak jej korzenie sięgają antycznej Grecji. Statystyka nowożytna pojawia się wraz z rozwojem nauki nowożytnej. Data narodzin nauki nowożytnej jest dokładnie znana. Jest to wigilia św. Marcina 10 listopada 1619 roku. Wówczas to Kartezjusz w objawieniu poznał podstawy zdumiewającej nauki (*mirabilis scientiae fundamenta*). Rok później, czyli w 1620 F. Bacon publikuje *Novum Organum*, sam autor jak i jego dzieło miały duży wpływ na rozwój statystyki jako nauki.

Dnia 20 listopada 1660 roku Herman Conring wygłasza pierwszy wykład na uniwersytecie Helmstadt na temat statystyki. Był to wykład pod tytułem *Notitia rerum politicarum nostri aevi celeberrimavum.* Czyli wykład o przedstawieniu stanu państwa. Naukę o państwie w sensie platońsko-arystotelesowskim nazywano bowiem polityką.

Naukę Conringa kontynuował Gottfried Achenwall (1719-1772). Dnia 7 września 1748 roku obronił on pracę habilitacyjną „*Notitam rerum publicarum academiis vindicatum, consentiente ordine philosophorum ampulissimo, praeses Gottfried Achenwall, pro lao in facultate philosophica obtinendo, ad diem VII Septembris, MDCCXLVIII disputatione publica defendet respondente Joanne Justo Henne, Gottinga"* (7.IX.1748) i w tym samym roku rozpoczął wykłady (w języku niemieckim) na Uniwersytecie w Marburgu.

Na podstawie notatek do wykładu w 1749 roku wydał pracę *Abriss der neuesten Staatswisenschaft der heutigen vornehmsten europäischen Reiche und Republiken 1749*. We wstępie do tej pracy po raz pierwszy pojawia się słowo *Statistik, r*ozumiane jako wiedza o państwie. W języku niemieckim określana ona jest też za pomocą słów *Staatskunde* lub *Staatswissenschaft*. Etymologia słowa statystyka jest jednak pochodzenia włoskiego. Wywodzi się ono od słowa *stato* czyli państwo. W języku łacińskim na określenie państwa zamiennie używano takich słów jak *status*, *respublica*, *civitas* lub *imperium*. Wiedzę o państwie określano więc po łacinie jako *ratio status*, po włosku zaś jako *regione di stato*. Na język angielski tłumaczono dosłownie jako *knowledge of the state*. Samego słowa *statistica* po raz pierwszy użył Ghislini już w 1589 roku w określeniu *scienza statistica,* czyli w postaci przymiotnikowej jako nauka statystyczna. W takim samym sensie słowa tego używa H. Kołłątaj w swym dziele *Stan oświecenia w Polsce*.

Słowo statystyka przez niektórych było ostro krytykowane jako „słowo zepsute", barbarzyńskie, *vox hybrida.* Mimo krytyki i niechęci do niego, już w 1779 pojawia się w języku francuskim jako *statistique*, w 1791 w Szkocji jako *statistics*.

W Polsce słowo to w postaci rzeczownika pojawia się po raz pierwszy w 1809 roku. Rozumiane jest w duchu nauki *Conringa-Achenwalla* czyli w sensie arystotelesowskiej polityki. Zmiana następuje dopiero w drugiej połowie XIX wieku głównie za sprawą genialnego Belga A. Queteleta. W celu wyjaśnienia tej zmiany trzeba się cofnąć z powrotem do wieku XVII. A dokładniej do roku 1662. Otóż w tym roku skromny kupiec, samouk, John Graunt przedstawił znamienitemu Towarzystwu naukowemu *Royal Society* pracę pod tytułem „*Natura and political observations upon the bills of mortality, chiefly with reference to the government, religion, trade, growth, air, diseases, etc., of the city of London, by Captain J.G. presented in 1662 to Royal Society*. W pracy tej po raz pierwszy zastosował nieznaną dotychczas metodę myślenia o państwie, którą dzisiaj nazywa się wnioskowaniem statystycznym.

Pod wpływem tej pracy, w 1679 roku William Petty opublikował pracę pod tytułem *Political Arithmetick, or a discourse concerning the Entent and Value of Landes, People, Buildings; Husbandry, Manufacture, Commerce, Fishery, Artizans, Seamen, Soldiers; Public Revenues, Interests, Taxes, Superlucration, Registries, Banks; Valuation of Men, Increasing of seamen*, dając tym samym nazwę dziedzinie nauki o państwie określanej jednakże nie opisowo lecz za pomocą miar i wag czyli za pomocą liczb. W odróżnieniu od historycznej szkoły Conringa-Achenwalla, dziedzinę zapoczątkowaną przez Graunta i rozwiniętą przez Süssmilcha i Queteleta nazywano arytmetyką polityczną, która później rozwijana była jak statystyka matematyczna.

2. POCZĄTKI ROZWOJU STATYSTYKI W POLSCE

Analizę historyczną wygodnie przedstawia się w określonych cenzusach czasowych. W całym rozwoju cywilizacji polskiej Hugo Kołłątaj wydzielił cztery okresy (por. Kołłątaj, 2003):

1. Od wprowadzenia chrześcijaństwa do założenia Akademii Krakowskiej w 1364 r.
2. Od 1364 do roku 1564 (sprowadzenie Jezuitów do Polski).
3. Od 1564 do roku 1773 (skasowanie tego zakonu).
4. Od 1773 roku do 1795 (wymazanie Polski z mapy Europy).

Na użytek tego artykułu podzielmy historię statystyki w Polsce według kryterium historycznego wyznaczonego rozbiorami i wojnami, czyli na okres przedrozbiorowy i trzy okresy porozbiorowe:

1. Okres do I wojny światowej.
2. Okres międzywojenny.
3. Okres po II wojnie światowej.

Wykaz wybitnych postaci związanych z rozwojem myśli statystycznej w Polsce przedstawiony jest w tabeli 1.

Tabela 1.

Wybitne postacie związane z rozwojem myśli statystycznej w Polsce

Okres	Autorzy
Przed rozbiorami	Haur, Jan z Wschowy, Maciej z Miechowa
Do I wojny	S. Staszic, J. Śniadecki, W. Załęski, W. Surowiecki, Z. Korzybski, M. Marassé, Z. Rewkowski, A. Danielewicz, W. Gosiewski, S. Dickstein, J. Czekanowski
Międzywojenny	J. Buzek, Waściszewski, J. Neyman
Po II wojnie	S. Szulc, M. Fisz, Z. Sadowski, Z. Pawłowski, A. Zeliaś, R. Zieliński, R. Bartoszyński, S. Zubrzycki, W. Oktaba, W. Klonecki, K. Zając, T. Caliński

Źródło: opracowanie własne.

Myśl statystyczna w Polsce zaczyna się jednak rozwijać dopiero gdy Polska przestała istnieć na mapie Europy. W XVIII wieku w Europie istniały już cztery wyraźnie zarysowane szkoły statystyczne.

Pierwszą z nich, najbardziej znaną, była niemiecka szkoła państwoznawców. Drugą zaś, angielska szkoła arytmetyków politycznych stosujących metody ilościowe do analiz zjawisk gospodarczych. Trzecią szkołę stworzyli Duńczycy, ze względu na szerokie stosowanie metod graficznych i tabelarycznych, nazywano ich tabelerystami. Czwartą szkołę, zapoczątkowaną w Belgii, nazywaną statystyką moralną, rozwijano także we Francji.

W Polsce w tym czasie panowała dynastia Wettynów. W języku polskim nie powstawały prawie żadne prace. Królowie polscy nie znali języka polskiego. Całe szkolnictwo opanowane było przez zakon Jezuitów z łacińskim językiem nauczania. Analfabetyzm wśród drobnej szlachty sięgał 92%, natomiast wśród magnatów i szlachty bogatej analfabetów było nieco mniej, bo ok. 40%. Podczas gdy w Europie od XVII wieku na dobre utrwalił się już absolutystyczny system rządów, tymczasem w Polsce wciąż panowała tzw. „wypaczona anarchicznie demokracja". Statystyka, jak wiadomo, rodziła się głównie z potrzeb państwowości, a ponieważ w Polsce jej nie było, stąd też i zapotrzebowania na naukę potrzebną do mądrego gospodarowania także nie było.

Analizując stan gospodarki polskiej w końcu XVIII wieku oraz polskie piśmiennictwo ekonomiczno-statystyczne z tego okresu, M. Marassé stwierdził krótko: „nie można bowiem taić, że nie ma statystyki polskiej w naukowym słowa tego znaczeniu". Statystyka taka zaczyna się rozwijać dopiero w XIX wieku, a dokładniej, w tzw. okresie wiedeńskim czyli w okresie Królestwa Polskiego. W całym stuleciu od końca XVIII wieku prac statystycznych w języku polskim w zasadzie nie było. Nie było bowiem w owym czasie w Polsce statystyków w takim sensie w jakim oni występowali w Niemczech (Prusach), Anglii czy we Francji. W Polsce byli to głównie oświeceni historycy i ekonomiści. Pierwsi traktowali statystykę jako ważne źródło informacji do refleksji historycznych. Warto pamiętać, że w XVIII wieku popularne było traktowanie statystyki jako nieruchomej historii (*stillstehende Geschichte*), lub historii w spoczynku, zaś historię traktowano jako statystykę w ruchu (*fortlaufende Statistik*). Ekonomiści natomiast, wykorzystywali statystykę do analiz gospodarczych. To właśnie ekonomiści

polscy publikowali przy okazji prace „czysto" statystyczne, głównie o charakterze informacyjnym i przeglądowym. Dużo takich prac ukazała się w czasopiśmie *Ekonomista*, które powstało w 1865 roku.

W drugiej połowie XIX wieku świadomość statystyczna w krajach zachodniej Europy była bardzo wysoka. Odbywały się międzynarodowe zjazdy statystyczne, znane były prace matematyków francuskich z rachunku prawdopodobieństwa, matematyzacja wiedzy o społeczeństwie była na tak wysokim poziomie, że statystykę traktowano jako fizykę społeczną. Polscy ekonomiści i statystycy stan tej wiedzy przedstawiali dość rzetelnie. Publikacji jednak było niezbyt dużo. Prawie każda opublikowana praca była wydarzeniem jednostkowym, nie była ani kontynuowana ani rozwijana przez innych autorów.

Bardzo mało prac obcojęzycznych przełożono na język polski. Udało mi się ustalić tylko trzy przetłumaczone prace (por. Quetelet, 1874; Bleicher, 1919 i Yule,1921).

3. KRÓTKA CHARAKTERYSTYKA PIŚMIENNICTWA STATYSTYCZNEGO XIX WIEKU

Polskie piśmiennictwo statystyczne ma swój początek w XIX wieku. Rachunek prawdopodobieństwa, traktowany jako dział matematyki, wykładany był znacznie wcześniej, bo już na początku XVIII wieku. Pierwszym śladem zainteresowań rachunkiem prawdopodobieństwa w Polsce są *Dyskursy,* Jana Śniadeckiego, jako też i jego rękopis z 1790 roku p.t. *Rachunek Zdarzeń i Przypadków Losu*, niedawno odkryte przez W. Więsława (por. *Dzieje* , 2012). Pierwszą opublikowaną pracę z rachunku prawdopodobieństwa p.t. *O początkach i wzroście rachunków prawdopodobieństw*, opublikował Zygmunt Rewkowski w 1828 roku (por. Rewkowski, 2011).

Ograniczmy się jednak do charakterystyki piśmiennictwa o tym, co wówczas nazywano statystyką.

Niewątpliwie najpierwszą pracą są notatki wykładu W. Surowieckiego, które zachowały się do naszych czasów w postaci rękopisu i zostały opublikowane w (Surowiecki, 1957). Rękopis zaś, znajduje się w Bibliotece PAN w Krakowie.

Wawrzyniec Surowiecki (1769-1827) zapisał się w historii nauki polskiej jako wybitny publicysta i ekonomista. W historii rozwoju myśli statystycznej w Polsce przysługuje mu zaszczytne miano tego, kto po raz pierwszy wygłosił wykład ze statystyki. Wykład ten jako *Statystyka Księstwa Warszawskiego* był wygłoszony w Szkole Prawa i Administracji w Warszawie w roku akademickim 1812/13.

Plan tego wykładu podany był w następującej postaci:

1. Topografia z jeografią.
2. Ludność.
3. Przymioty i stan ziemi Księstwa.
4. Płody natury.
5. Płody przemysłu.
6. Handel.
7. Instrukcja.

8. Religia.
9. Rząd Krajowy.
10. Bogactwa, dochody i wydatki kraju.

W spisie literatury „potrzebnej do ułożenia statystyki Księstwa" nie ma jednak prac dobrze znanych statystyków takich jak Achenwall czy Quetelet.

Wykład miał prawdopodobnie inny cel niż wprowadzenie studentów w istotę problematyki nauki o państwie. Jako orędownik postępu społecznego i prawdziwy patriota, z zacięciem publicystycznym przedstawiał wiedzę o państwie takim jakim powinno ono być. Na początku swego wykładu zwracając się do studentów stwierdza:

„Wiecie, że kraj ten, niedawno pod obcym panowaniem i fałszywym imieniem, jest częścią wielkiego niegdyś państwa polskiego, które po różnych przygodach i burzach politycznych za zmową trzech ościennych sąsiadów haniebnie rozszarpane zostało. Z podań i dziejów znacie dawno jego znaczenie, zamożność, zamożność i potęgę; z własnej pamięci przypominacie sobie jego spodlenie, nikczemność i poniewierkę. Chcecie, żebym wam powiedział, kto był przyczyną tych nieszczęsnych igrzysk, na które był wystawiony? Niestety! My sami!".

Analizując tabele „rządu dzisiejszego" S. Wawrzyniec wnioskuje, że „w całym Księstwie znajduje się samych chrześcijan około 4 025 357" odtrąciwszy Niemców, Rusinów, Litwinów, szacuje, że „Polaków właściwych" w Księstwie było 3 426 123.

Doliczywszy tych żyjących pod obcym panowaniem, W. Surowiecki uzyskuje liczbę Polaków wynoszącą 4 486 123 osoby. W wykładzie Surowieckiego na uwagę zasługuje wnikliwa prezentacja stosunków etniczno-wyznaniowych. Z właściwą sobie pasją publicysty przedstawia sytuację Żydów zamieszkujących tereny państwa polskiego. Z wielką odwagą demaskuje księży, „którzy przez najokrutniejsze potwarze i czernidła podżegali lud do nienawiści przeciw Żydom". W. Surowiecki jest niewątpliwie osobą, o której powinniśmy pamiętać.

Niżej podana jest charakterystyka ważniejszych prac, które ukazały się drukiem w XIX wieku. Rozpocznijmy od chronologicznie najwcześniejszej pracy, której autorem jest Felicjan Kozłowski.

F. Kozłowski (1805-1870) jako nauczyciel w Gimnazjum Guberialnym w Warszawie wydał w 1838 roku wielkich rozmiarów pracę p.t. „*Rys statystyki ogólnej porównawczej pod względem darów przyrodzenia, przemysłu pierwotnego, rękodzielnego, fabrycznego, handlu i kultury państw Europy*". Jak sam zaznacza, że jest to pierwsza w Polsce praca statystyczna. „Gdy w literaturze zagranicznej osobliwie niemieckiej wyszło już wiele dzieł statystyki ogólnej państwa Europy, a w naszej żadnego jeszcze w tym rodzaju nikt nie wydał, przeto przenikniony ważnością przedmiotu, postanowiłem z najnowszych pisarzy niemieckich i francuskich zebrane główne data statystyczne, tyczące się darów przyrodzenia, ludności, przemysłu i kultury państwa Europy, i te w systematycznym porządku wystawić…". Całość opisu jest bardzo lakoniczna, bez podawania źródeł. Na każdy problem wymieniony w tytule pracy poświęca się parę zdań. Weźmy dla przykładu rozdział p.t. „Ludność w Europie podług wyznań". Treść jego jest następująca: Przyjmując ludność Europy na 225 000 000, wypada z niej Chrześcijan

219 milionów, Mahometanów 4 miliony, Żydów 1 $^{1}/_{2}$ miliona, Pogan na północy około 2 000. Chrześcijanie znowu dzielą się na Katolików w liczbie 132 milionów, Luteranów 25 milionów, Reformowanych 10 milionów, Episkopalnych czyli Anglikańskiego wyznania 14 milionów, Prezbiteryanów, Kongrecyonistów i z innych sekt mniejszych w zachodnim kościele 5 milionów. Chrześcijan Greckiego wyznania z Armeńczykami rachują 31 milionów, a mniejszych sekt we wschodnim kościele 2 miliony".

Autorem pierwszych prac w języku polskim odpowiadających nawet dzisiejszym rygorom prac naukowych jest Mieczysław Marassé (1840-188). Zmarł on wcześnie, bo w 40 roku życia, pozostawił jednak po sobie bardzo cenny dorobek. Dorobek ten został pośmiertnie wydany przez jego brata Adama w 1886 roku w postaci dwóch tomów p.t. „*Dzieła ekonomiczno-polityczne i statystyczne*" (por. Marassé, 1886). Wydanie to, poprzedzone jest obszernym wstępem p.t. „Pogląd na ekonomiczne i statystyczne prace ś.p. Mieczysława Marasse'go" napisanym w 1882 roku przez Tadeusza Pilata, w którym szczegółowo scharakteryzowane jest całe pisarstwo Marasségo.

Oprócz dwóch obszernych artykułów statystycznych o charakterze sprawozdawczym, w 1866 roku M. Marassé opublikował książkę p.t. „O pojęciu i zadaniu statystyki". Jest to pierwsza książka w języku polskim, w której przedstawiona została historia statystyki. Statystykę Marassé traktuje jako jedną całość złożoną z dwóch części. Jedną z nich jest umiejętność statystyczna drugą zaś, statystyka administracyjna. Statystyka tak rozumiana stanowi podstawę nauk społecznych. Marassé tak oto to wyraża: „Bez statystyki nie zrobimy ani kroku, ona wszystkie teorie gospodarstwa narodowego poprowadzić musi, bo te dopiero z niej wypłynąć mogą". M. Marassé nie jest li tylko autorem pierwszej polskiej pracy o historii statystyki, ale jest też pierwszym autorem, który nie tylko pisał *o statystyce* ale i stosował ją. Nadzorował on bowiem pierwsze w Galicji badanie statystyczne gospodarstw rolnych, które było przeprowadzone w 1847 roku. Szczegółowo przedstawił plan badania, krytycznie analizując jego słabe strony. W wyniku przeprowadzonych badań okazało się, że około 60% areału gruntów uprawnych należy do małych gospodarstw, nie przekraczających 50 morgów. Na ogólną liczbę 800 000 „własności ziemskich" aż 216 000 gospodarstw nie przekracza 2 morgów, ok. 440 000 posiada od 5 do 50 morgów. M. Marassé analizując sytuację społeczną kreśli słowa aktualne i w czasach obecnych. Oto one:

„Wiek wolnej konkurencji jest zarazem wiekiem ekonomicznej zawisłości dla tych, którzy z wolności korzystać nie umieją. A pewno nie ma gorszej niewoli, niż zawisłość materyalnie, gorszego i nielitowściowszego tyrana niż Kapitał". Na ogólną liczbę 5 368 702 mieszkańców Galicji w 1847 roku aż 1 750 000 żyło w niepewności „z dnia na dzień".

Kolejną postacią, której przysługuje poczesne miejsce na kartach historii rozwoju myśli statystycznej w Polsce jest Józef Deskur. Niestety, niewiele jest dostępnej o nim informacji, na próżno szukać tego nazwiska w Słowniku (1998). A przecież to jego, a nie kogo innego, można w pełni nazwać prekursorem stosowania metod Queteleta na gruncie polskim. W 1867 wydał on niewielkich rozmiarów książkę p.t. „*Zadania naukowe i życiowe statystyki, rzecz napisana z powodu Wykazów Statystycz-*

nych Sądowo-karnych, na rok 1865". Na początku tej rozprawy prezentuje pokrótce istotę statystyki, ale bardziej z punktu widzenia uczonego-filozofa aniżeli praktyka-ekonomisty. Następnie analizuje, na wzór analiz dokonywanych przez Queteleta, wyniki wykazu sądowo-karnego jaki Komisja Rządowa Sprawiedliwości wydała w 1866 roku. Rok tego wydania J. Deskur uważa za ważne wydarzenie w dziejach statystyki polskiej, a przez to i dla całej nauki. Rezygnując ze szczegółowej prezentacji wyników uzyskanych przez Deskura zwróćmy tylko uwagę na pewne ich osobliwości.

Wszystkich czynów karygodnych odnotowano 25 719, w tym 10 467 wykroczeń objętych ustawą gminną. Wyrokami sadowymi skazano 5 497 osoby, w tym 4 422 mężczyzn i 1 075 kobiet. Rozkład według wieku był następujący:

Do lat 14 skazano 52 osoby,
Od 14 do 21 skazano 469 osób,
Od 21 do 60 skazano 4 754 osoby,
Powyżej 60 lat skazano 222 osoby.

Jako kary wymierzano między innymi takie jak: mieszkanie na Syberii, poprawcze aresztanckie roty, dom poprawy, dom roboczy i inne.

Autorami pierwszych książek na temat statystyki byli Witold Załęski oraz Zdzisław Korzybski.

W. Załęski (1836-1908) wydał dwie książki, jedną pod tytułem „*Kilka słów o teorii statystyki*" o objętości 76 stron wydał w 1868 roku. Drugą, pod tytułem „*Teoria statystyki w zarysie, część I. Zasady ogólne i część historyczna*", o objętości 303 strony wydał w 1884 roku. Natomiast Z. Korzybski (1834-1896) książkę swoją pod tytułem „*Wstęp do teorii statystyki*" wydał w 1870 roku. Prace obu autorów powstały więc mniej więcej w tym samym czasie, są też do siebie podobne. Zawierają dużo ciekawego materiału historycznego, dość rzetelnie prezentowane są poglądy na statystykę wyrażone przez znanych zachodnich statystyków. Są to jednak prace o statystyce a nie ze statystyki.

W. Załęski traktuje metodę statystyczną jako jedną z metod poszukiwań naukowych. W sposób precyzyjny określa ją jako postrzeżenie systematyczne wieloliczne. W celu dokładniejszej prezentacji metody cały świat dzieli na „świat przyrodzenia" i „świat ludzki, czyli moralny i intelektualny obejmujący Boga i duszę ludzką".

Ponieważ w świecie ludzkim zdarzenia zależą od przyczyn stałych oraz przypadku, to stąd wynika główne zadanie statystyki, a mianowicie zadaniem tym jest usunięcie przypadku i uzyskanie przez to prawa statystycznego. W obu swych książkach W. Załęski drobiazgowo omawia metody wnioskowania, zwracając szczególną uwagę na metody J. S. Milla, dzieło którego rekomenduje wśród kilku innych jako te, które mogą posłużyć do bliższego poznania teorii statystyki. Podziela tez jego opinię o tym, że rachunku prawdopodobieństwa nie należy stosować do rozwiązywania problemów społecznych, w szczególności sądowniczych.

Na uwagę zasługuje to, że oprócz prac Quteleta omawia prace Guerry'ego. Obaj oni tworzyli mniej więcej w tym samym czasie obaj byli pionierami i obaj stworzyli

epokowe dzieła. Ale Quetelet, w dużej mierze dzięki sprawnej auto-promocji, był osobistością bardzo znaną. Andre-Michel Guerry prowadził życie o wiele bardziej ciche i skromne, o jego życiu najwięcej się dowiadujemy z nekrologu napisanego przez jednego z jego przyjaciół. W swym epokowym dziele Guerry analizuje dane 226000 przestępstw kryminalnych popełnionych w ciągu 25 lat we Francji i Anglii, oraz 85000 przypadków samobójstw. Jak sam Guerry szacuje, gdyby zebrane przez niego dane ułożyć jedna po drugiej , to zajęły by one 1170 metrów. W. Załęski przytacza tylko mały fragment tej pracy dotyczący samobójstw klasyfikowanych według ich liczby w poszczególnych miesiącach. Stwierdza między innymi, że „maksimum samobójstw nie przypada zatem na dnie mgliste jesieni w czasie żałoby natury, ale przeciwnie na przesilenie letnie koło św. Jana, kiedy dnie są najdłuższe i słońce jest w całym blasku"(por. Załęski, 1884).

Najważniejszą i najcenniejszą cechą pracy Załęskiego, na którą warto zwrócić uwagę jest prezentacja systemów statystyki administracyjnej w 24 krajach. Szczególnie cenne są także informacje o dziewięciu kongresach statystycznych. Po kolei omawianych, począwszy od kongresu w Brukseli, który się odbył w 1853 roku, a skończywszy na Kongresie w roku 1876, który się odbył w Budapeszcie. Prawdopodobnie jest to jedyne (do dni dzisiejszych) polskie źródło informacji o tych Kongresach.

Podobną do tych dwóch prac Załęskiego jest praca Z. Korzybskiego. Odznacza się ona jednak większą rzetelnością źródłową i większą klarownością języka. Historia statystyki przedstawiona przez Korzybskiego przewyższa znacznie uwagi na ten temat czynione przez Załęskiego. Analizują etymologię słowa „statystyka" wyraźnie oddziela od siebie dwa różne znaczenia łacińskiego słowa „*status*" oznaczającego albo *stan* albo *państwo*, a nigdy stan państwa.

W pierwszym zdaniu swego dzieła, Korzybski stwierdza: „Mało zaprawdę jest nauk, któryby jak Statystyka stanowiły tak wielki przedmiot niezgody między uczonymi, a to tak pod względem ściśle naukowego określenia samego przedmiotu, jako też zadania nauki i samej nawet metody przedstawiania". Szczegółowo przedstawia więc różne koncepcje rozumienia statystyki z perspektywy jej rozwoju historycznego. Dość dokładnie prezentowane są prace 16 najbardziej znanych statystyków XIX wieku. Wymieńmy niektórych z nich, podając hasłowo stanowisko każdego z autorów.

1. Moreau de Jones: statystyka jest nauką faktów przyrodzonych, społecznych i politycznych wyrażonych cyframi.
2. Ludwik Wołowski: statystyka daje nam poznać bieg i rozwój zjawisk społecznych za pomocą cyfr.
3. Dufau: statystyka jest nauką praw przewodniczących rozwojowi faktów społecznych (w oryginale: *théorie de l'étude des lois d'aprés lesquelles les dévelopment les faitas sociaux*).
4. Knies: zadaniem matematycznej szkoły statystycznej jest „wyprowadzanie praw wielkiej liczby", zaś zadaniem statystyki Conringo-Achenwallowskiej jest przedstawienie stanu państwa. Pogodzenie (i połączenie w jeden system) tych dwóch szkół Knies uważał za rzecz niemożliwą.

Dość szczegółowo omawia prace Queteleta, którego uważa, tak zresztą jak prawie wszyscy współcześni mu autorzy, „za gwiazdę przewodnią dzisiejszej matematycznej szkoły". W pracy Skrzywana (1954) Quetelet określony jest jako współczesny twórca przesądów statystycznych. „Jako twórca nie był on postępowy, , nawiązywał do tradycji i to tradycji niegodnej kontynuacji, oryginalne myśli jego z małymi wyjątkami były nieudane i szkodliwe dla rozwoju statystyki"(por. Skrzywan, 1954). Opinia taka mogła być konsekwencją opinii K. Marksa, który już w 1869 roku napisał, że Quetelet nie potrafił wyjaśnić zaobserwowanych regularności zjawisk społecznych, w badaniach swych nie czynił postępu, rozszerzał li tylko materiał swoich obserwacji.

Statystyce polskiej Korzybski poświęca mało miejsca, niecałe dwie strony. Zaznacza, że kierunek nadany statystyce przez Achenwalla znalazł w Polsce pewien odgłos w postaci „maleńkiej broszury *O statystyce Polski*", której autorem był Stanisław Staszic. Z. Kobrzycki daje następującą ocenę tej pracy: „W pracy tej zresztą cyfra występuje tylko gdzieniegdzie, a owa broszurka, gdyby nie tytuł jaki nosi, mogłaby być wzięta, szczególnie dzisiaj, prędzej za jakiś polityczny memoriał, jak za systematyczną pracę". Z. Korzybski deklaruje swoje własne stanowisko. Statystykę uważa za bardzo ważną i potrzebną naukę przedmiotem, której ma być państwo i społeczeństwo. Przy czym najważniejszym celem państwa ma być szczęście obywateli. Siebie określa jako „przyjaciela postępu" i uważa, że „bez przymusowego nauczania, zupełnej wolności prasy i samorządu gminy, prawdziwy, to jest trwały postęp jest niemożliwy". Prawie identycznie brzmią sztandarowe słowa obecnego ruchu na rzecz trwałego (zrównoważonego) postępu społecznego.

4. PIŚMIENNICTWO NA TEMAT HISTORII STATYSTYKI W POLSCE

Naśladując Marasségo można powiedzieć, że nie można taić, iż zwartych opracowań na temat historii statystyki w Polsce nie ma. Po raz pierwszy najwięcej na temat historii statystyki w ogóle napisał właśnie Marassé w 1866 roku. Dwa lata później Załęski tak oto pisze (por. Załęski, 1868):

„Polaków niewielu pisało o statystyce: Fryderyk hr. Skarbek mówi o niej w swojem gospodarstwie narodowem stosowanem, Warszawa 1860 r." pojmuje ją jednak li tylko jako naukę znajomości kraju. Z nowszem pojęciem spotykamy się w pracach: Marassé „O pojęciu i zadaniu statystyki. Kraków 1866". Descur, „Zadanie życiowe statystyki. Warszawa 1866" i J. Obiezierski, „Statystyka wobec zagadnień wyższego porządku", w Ekonomiście, 1866, II, str. 209 i n."

Po prawie 150 latach możemy ze smutkiem powiedzieć prawie to samo: Polaków niewielu pisało o historii rozwoju statystyki w Polsce.

Jeśli chodzi o prace polskich autorów, wydanych współcześnie (w XX wieku), to na uwagę zasługuje opracowanie Skrzywana (1954), wydane przez PWN jako „materiały do wykładów" w Szkole Głównej Planowania i Statystyki. Zawiera ono dużo materiału o historii statystyki i nauki w ogóle. O historii statystyki w Polsce materiału jest tam niewiele. Wymieniane są w zasadzie tylko nazwiska i opublikowane przez nich prace.

Pierwszą wydaną w Polsce i najbardziej kompletną „ewidencję" polskich statystyków wraz z ich biogramami stanowi opracowanie p.t. *Sylwetki statystyków polskich* (por. Sylwetki, 1984), przygotowane przez Stanisława Kwiatkowskiego i wydane w Łodzi w 1984 roku. Wartościową i ważną cechą tego wydawnictwa jest podana na końcu opracowania lista 84 pozycji źródeł bibliograficznych. Rozszerzoną wersję tego opracowania wydano w Warszawie w 1998 roku w postaci Słownika biograficznego (por. Słownik, 1998). W wielu przypadkach podawane jednak tam informacje nie są należycie dokumentowane, często z pominięciem źródeł i odnośników do prac oryginalnych. Z lektury tego Słownika dowiadujemy się na przykład, że mieliśmy ponad 35 prekursorów statystyki, wielu pionierów i założycieli szkół statystycznych. Praca Załęskiego z 1884 roku została tam nazwana pierwszym polskim podręcznikiem statystyki matematycznej, i informację tę powielano w innych opracowaniach. Praca ta nie jest jednak w żadnym sensie podręcznikiem statystyki, a tym bardziej statystyki matematycznej. Poza elementarnymi rozważaniami o losowaniu kul z urny, nie ma w niej bowiem żadnych informacji o prawdopodobieństwie i jego zastosowaniu do wnioskowania. Statystyka matematyczna była w owych czasach chyba dobrze znana. Świadczyć o tym może wydana w tym samym roku, czyli w roku 1884, praca p.t. *Z dziedziny statystyki matematycznej*, której autorem jest Bolesław Danielewicz. Dotyczy ona co prawda tylko problematyki konstrukcji tablic trwania życia, ale stosowany w niej aparat probabilistyczny poziomem swym dorównuje, a nawet przewyższa, poziom wymagany w dzisiejszych programach nauczania statystyki na kierunkach ekonomicznych.

Najlepiej udokumentowane, pisane profesjonalnie i dobrą polszczyzną są pojedyncze artykuły autorstwa Juliusza Łukasiewicza. Wiele z nich opublikowanych zostało w serii o nazwie *Biblioteka Wiadomości Statystycznych*. Na pewno było by wielce wskazanym wydać je w zwartej i usystematyzowanej postaci, na przykład jako odrębny tom tej *Biblioteki*.

Niewątpliwie, najcenniejszym źródłem wiedzy o polskich pracach statystycznych jest *Wybór pism statystyków polskich,* którego dokonał E. Rosset, zaś przedmową opatrzył S. Konferowicz (por. Rozwój, 1968).

Teksty profesjonalnych historyków nauk matematycznych ,dotyczące w szczególności teorii prawdopodobieństwa, publikowane były w czasopiśmie *Antiquitates mathematicae,* wydawanie którego zostało niestety po 2009 z niewiadomych przyczyn zaniechane. Inne źródło informacji o historii statystyki stanowi seria wydawnicza *Studia i materiały z dziejów nauki polskiej* wydawanej pod egidą PAN od 1957 roku. Od 1992 seria ta wydawana jest w zmienionej wersji.

5. NIECO KRYTYKI O PIŚMIENNICTWIE HISTORYCZNYM

Ponieważ w początkowym okresie rozwoju o statystyce pisali ekonomiści, to informację o pracach statystycznych można znaleźć w opracowaniach dotyczących historii myśli ekonomicznej w Polsce. Niestety, dotychczasowe opracowania na temat myśli ekonomicznej przesiąknięte są ideologią. Prace statystyczne były oceniane z punktu

widzenia obowiązującej doktryny. Weźmy dla przykładu pracę wybitnego statystyka, i chyba ekonomisty także, M. Marasségo. Z jego prac jeszcze dzisiaj można dowiedzieć się bardzo wiele, a czyta się z prawdziwą przyjemnością, gdyż jak to już dawno temu zauważył Pilat, że prace te cechuje „jasność argumentacji i styl przyjemny" (por. Marassé, 1886). Niestety, prace Marasségo nie zasługiwały na upowszechnienie gdyż mimo iż on zapoczątkował w Polsce rozwój kierunku historycznego w ekonomii, to jednak „pozostając pod wyraźnym wpływem teoretyka niemieckiej szkoły historycznej W. Roschera, stara się w licznych, zresztą bardzo powierzchownych pracach, uzasadnić tezę o zależności myśli ekonomicznej od rozwoju historycznego" (por. Guzicki, 1969). Jeszcze ostrzej oceniony został Korzybski. „Prymitywizm wywodów Korzybskiego i upraszczanie myśli nawet takich wulgaryzatorów jak Say i Bastiat, świadczą o głębokim upadku polskiej myśli ekonomicznej tak chlubnie rozwijanej półwieku wcześniej" (por. Guzicki, 1969).

Nie tylko w powojennej ocenie spotkała go niesprawiedliwość. Nie są znane (przynajmniej autorowi tego tekstu) przyczyny trudności z jakimi się spotkał Korzybski ze strony Rady Wydziału Prawa i Administracji w Szkole Głównej w Warszawie, w uzyskaniu stanowiska profesora ekonomii. Stanowisko takie uzyskał, i jak sam zaznacza, „dla przyczyn odemnie niezależnych tylko przez jeden semestr wykładałem, a właściwie zacząłem wykładać Statystykę... opuszczając Katedrę wydaję ów wstęp do teorii, który może będzie mógł uczącej się młodzieży dalsze studia ułatwić"(por. Korzybski, 1870). ów wstęp, to świetna praca p.t. *Wstęp do teorii statystyki. Część pierwsza, Rys historyczny i ogólne zasady.* Praca Z. Korzybskiego gdyby była dzisiaj ponownie wydana, to mogła by i dzisiaj „uczącej się młodzieży dalsze studia ułatwić", zaś wielu współcześnie piszącym o statystyce, stanowiłaby wzorzec pisarstwa.

Zupełnie inaczej oceniany jest D. Krysiński. Przez zwolenników ideologii marksistowskiej, przyznano mu jedno z czołowych miejsc w „kształtowaniu polskiej myśli ekonomicznej" (por. Guzicki, 1969). Z kolei statystycy uważają go za prekursora badań budżetów rodzinnych w Polsce. O badaniach takich chyba nic nie wiadomo, wiadomo zaś, że w pracy p.t. *O arytmetyce politycznej ze względu na te nauki, które jej szczególniejszą są podstawą,* wydanej w 1814 roku D. Krysiński zdecydowanie występuje przeciwko matematyzacji nauk społecznych. O samej arytmetyce politycznej pisze, że „skromne jest jej przeznaczenie. Jest to tylko zbiór sposobów obrachowywania przez przybliżenie tych politycznych przedmiotów, które za pomocą tylko rachunku mogą nam dać naprzód jakoweś do postępowania wskazówki i chronić tym samym od błędów, które zawsze drogo przypłacają narody".

Oprócz Krysińskiego, szczególnie wysoko ceniony przez ekonomistów jest F. Skarbek (1792-1866). Jego główne dzieło studiował bowiem K. Marks i powoływał się na nie w swym *Kapitale* (por. Skarbek, 1957). W słownikach statystycznych Skarbek określany jest pionierem polskiej myśli statystycznej, który wprowadził nowe elementy naukowe do rozwoju myśli statystycznej. Statystykę rzeczywiście docenia i pisze o niej we *Wstępie* (i tylko tam) do swego dwutomowego dzieła *Ogólne zasady gospodarstwa narodowego*. Oto całość tekstu poświęconego statystyce (por. Skarbek, 1957):

„Do tych wszystkich oddziałów nauki gospodarstwa narodowego dodać należy jeszcze jedną nader ważną, która się wspiera na nauce gospodarstwa narodowego i jest zarazem dla tejże punktem wsparcia do dalszego posuwania się w zawodzie udoskonalenia własnego; a tą jest *statystyka*. Przekonaliśmy się wyżej, iż nauka gospodarstwa narodowego jest wypływem obserwacji zdarzeń w towarzyskim życiu narodów przytrafiających się; statystyka zaś jest, że tak rzekę, ciągłą tychże zdarzeń kontrolą, która dostarcza materiałów do sprawdzenia dawniej czynionych spostrzeżeń i do czynienia nowych. Jej związek z całą nauką gospodarstwa narodowego jest ścisły i wzajemny; bo jeżeli zebranie i zachowanie faktów najważniejszą innym częściom tejże nauki świadczy przysługę, sama nie mogłaby istnieć albo przynajmniej być tyle pożyteczną, ile być może, gdyby się na zasadach nauki gospodarstwa narodowego nie opierała."

F. Skarbek dwa lata studiował w Paryżu. Do Polski wrócił w 1811 roku. We Francji w tym czasie statystyka była bardzo dobrze znana, zarówno administracyjna, jak i matematyczna. W wyniku rewolucji w 1789 roku we Francji dokonywano ogromnej reformy administracyjnej, a to wymagało wielu danych. Angażowano do tego wszystkich uczonych. Od 1803 roku w Paryżu istniało aktywnie działające towarzystwo statystyczne (*Société de Statistique*).

6. ZAKOŃCZENIE

Niżej podane jest krótkie schematyczne zestawienie w porządku chronologicznym ważniejszych polskich dokonań w zakresie statystyki na tle osiągnięć statystyki światowej.

Akadademia Krakowska	1364	
	1660	Conring, Graunt
	1693	Halley
	1712	Arbuthnot, deMoivre, Bernoulli
	1740	Anchersen, Süssmilch
	1749	Achenwall: wykład po niemiecku (Statistik)
Śniadecki	1790	Achenwall (siódme wydanie jego pracy)
	1791	Sinclair: statistics (po angielsku)
Staszic	810	Malthus, Fourier
Surowiecki:wykład po polsku	1812	
Rewkowski	1828	
	1834	Królewskie Towarzystwo Statystyczne
Kozłowski	1835	Quetelet,
	1853	Kongres statystyczny
Surowiecki, Marassé, Załęski	1860	

Korzybski	1870	
Kleczyński, Danielewicz	1880	
Gargas, Gosiewski	1900	
Danielewicz & Dickstein	1910	Yule
PTS	1912	

W powyższym zestawieniu daty zostały podane w pewnym zaokrągleniu. Zestawienie to należy odczytywać w sposób następujący: w 1660 roku pojawiły się prace Conringa i Graunta, w 1810 roku ukazują się prace Malthusa i Fouriera, w tym samym czasie ukazują się publikacje Staszica i Śniadeckiego.

Celem tego zestawienia jest zademonstrowanie w prosty i poglądowy sposób polskiego zapóźnienia, o którym pisali: Marassé, Załęski, Korzybski, zaś Obiezierski w 1866 roku ujął to następująco: „kraj nasz niezawodnie należy do najbardziej zacofanych na drodze racjonalnych badań społecznych", do których zaliczał między innymi statystykę. Być może jedną z przyczyn nielicznych polskich publikacji było to, „że ściśle naukowe studia znajdują u nas bardzo nieliczne grono pracowników, a książki pisane nie dla samej tylko rozrywki, lecz wymagające od czytelnika dokładnego obeznania z elementarnemi pewnikami nauk przyrodzonych i ścisłych – nie popłacają w księgarniach" (por. Obiezierski, 1866).

PODZIĘKOWANIA

Dziękuję dwóm anonimowym Recenzentom za wytknięcie mi mankamentów w pierwotnej wersji artykułu i sugestie jego poprawy, dziękuję także Pani Elżbiecie Rogalskiej za okazaną mi pomoc w dostępie do publikacji źródłowych.

Uniwersytet Ekonomiczny we Wrocławiu

LITERATURA

[1] Bleicher H., (1919), *Statystyka*, Warszawa.
[2] Danielewicz D., (1884), *Z dziedziny statystyki matematycznej*, Warszawa.
[3] Deskur J., (1867), *Zadania naukowe i życiowe statystyki, rzecz napisana z powodu Wykazów Statystycznych Sądowo-karnych, na rok 1865*, Drukarnia Gazety Warszawskiej, Warszawa.
[4] *Dzieje matematyki polskiej*, (2012), praca zbiorowa pod red. W. Więsława, Instytut Matematyczny Uniwersytetu Wrocławskiego, Wrocław, Guzicki L., S. Żurawicki, (1969), *Historia polskiej myśli ekonomicznej do roku 1914*, PWE, Warszawa.
[5] Kołłątaj H., (2003), *Stan oświecenia w Polsce w ostatnich latach panowania Augusta III (1750-1764)*, Ossolineum, Wrocław.
[6] Korzybski Z., (1870), *Wstęp do teorji statystyki,* Nakładem autora, Drukarnia K. Kowalewskiego, Warszawa.
[7] Krysiński D., (1814), O arytmetyce politycznej ze względu na te nauki, które jej szczególniejszą są podstawą.
[8] Marassé M, (1886), *Dzieła ekonomiczno-polityczne i statystyczne*, Nakładem A. Marasse'go, Drukarnia Uniwersytetu Jagiellońskiego, Kraków.

[9] Obiezierski J., (1866), *Statystyka wobec zagadnień wyższego porządku*, Ekonomista, II, 1866, 209-236.
[10] Quetelet A., (1874), *Układ społeczny i jego prawa*, Nakładem Księgarni Celsa Lewickiego i S-ki, Warszawa.
[11] Rewkowski Z. (2011), *Pamiętniki,* Tom I, II, Na prawach rękopisu, opracował Więsław W., Instytut Matematyczny Uniwersytetu Wrocławskiego, Wrocław.
[12] *Rozwój polskiej myśli statystycznej*, (1968), Wybór pism statystyków polskich, PWE, Warszawa.
[13] Skarbek F., (1957), *Ogólne zasady nauki gospodarstwa narodowego*, PWN Warszawa, (wydana praca z 1820 roku).
[14] Skrzywan W., (1954), *Historia statystyki*, Warszawa, *Słownik biograficzny statystyków polskich*, (1998), GUS i PTS, Warszawa.
[15] Surowiecki W.,(1957), *Wybór pism*, PWN, Warszawa.
[16] *Sylwetki statystyków polskich*, (1984), Wojewódzki Urząd Statystyczny w Łodzi i PTS Oddział w Łodzi, Łódź.
[17] *Wokół Bernoullich*, (2006), Praca zbiorowa pod redakcją W. Więsława, Lublin.
[18] Yule G., U., (1921), *Wstęp do teorii statystyki*, Warszawa.
[19] Załęski W., (1868), *Kilka słów o teorii statystyki,* Drukarnia Gazety Polskiej, Warszawa.
[20] Załęski W., (1884), *Teoria statystyki w zarysie, Część I. Zasady ogólne i część historyczna*, Biblioteka Umiejętności Prawnych, Warszawa.

ROZWÓJ MYŚLI STATYSTYCZNEJ W POLSCE W XIX WIEKU

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

W artykule tym przedstawiony jest krótki rys historyczny rozwoju statystyki. W szczególności wyjaśniony jest rodowód słowa statystyka. Na tle ogólnego rozwoju przedstawione zostały warunki społeczno-polityczne w jakich powstawała statystyka polska. Pobieżnie omówiono prace dotyczące statystyki, jakie ukazały się w zwartej postaci w XIX wieku. Organizacja statystyki publicznej i literatura na ten temat, jako osobna problematyka, w artykule tym nie jest omawiana. Dokonano zaś krótkiej analizy krytycznej istniejącego polskiego piśmiennictwa na temat rozwoju myśli statystycznej w Polsce.

Słowa kluczowe: statystka, historia

THE DEVELOPMENT OF STATISTICAL THOUGHT IN POLAND IN 19^{th} c

A b s t r a c t

This paper presents a short overview of the history of statistics in Poland. The most significant books published during XIX century are briefly described. Organization and publications concerning the official statistics are beyond the scope of this paper. A critical description of the existing polish literature concerning the history of statistics is also included into this paper.

Key words: statistics, history

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

TADEUSZ CALIŃSKI

ROZWÓJ I OSIĄGNIĘCIA W BIOMETRII POLSKIEJ

Mówiąc biometria, mamy na myśli dyscyplinę naukową zajmującą się zastosowaniami metod matematycznych i statystycznych w rozwiązywaniu problemów biologicznych, w szerokim tego słowa znaczeniu, a zwłaszcza w planowaniu i analizie eksperymentów. W Polsce pionierami biometrii było dwóch wybitnych uczonych.

Jednym z nich był antropolog, Jan Czekanowski (1882-1965), autor pierwszego podręcznika biometrii w języku polskim (Czekanowski, 1913). Więcej o nim w pracy Krzyński (2009).

Drugim był chemik, agrotechnik i hodowca roślin, Edmund Załęski (1863-1932), autor pierwszego podręcznika polskiego z metodyki doświadczalnictwa rolniczego (Załęski, 1927). Więcej o nim i jego uczniach w pracy Schmidta (1971).

Idee Edmunda Załęskiego zostały podjęte i rozwinięte przez kilku sławnych uczonych polskich. O ile teoretyczny rozwój tych idei zawdzięczamy głównie jednemu z twórców nowoczesnej statystyki matematycznej, Jerzemu Spławie-Neymanowi (1894-1981), o tyle praktyczne wcielenie tych myśli do metodyki doświadczalnictwa rolniczego i biometrii jest zasługą przede wszystkim Stefana Barbackiego (1903-1979), bezpośredniego ucznia Załęskiego.

Stefan Barbacki od 1945 roku związany był z Poznaniem, gdzie w maju owego roku habilitował się, a następnie zorganizował na Wydziale Rolniczo-Leśnym Uniwersytetu Poznańskiego Katedrę Doświadczalnictwa Rolniczego i Biometrii. Tu rozwinął swoje przedwojenne znaczne osiągnięcia i przemyślenia dotyczące planowania eksperymentów i analizy statystycznej ich wyników. Jego działalność naukowa i pedagogiczna przyczyniła się do stworzenia poznańskiej szkoły statystyki matematycznej i biometrii (patrz Caliński, 1995; Caliński i inni, 2004).

Podobnie rodziły się zainteresowania biometrią w kilku innych polskich ośrodkach naukowych. Należy tu wymienić zwłaszcza następujące trzy osoby i ich szkoły.

Mikołaj Olekiewicz (1896-1971), absolwent i doktorant Uniwersytetu Columbia w Nowym Jorku, był inicjatorem lubelskiej szkoły biometrycznej, rozwiniętej na Akademii Rolniczej (obecnie Uniwersytecie Przyrodniczym) w Lublinie przez jego przyjaciela i ucznia Wiktora Oktabę (1920-2009) i nadal rozwijanej przez uczniów Wiktora Oktaby (patrz Oktaba, 1973, 1998; Wesołowska-Janczarek, 2002). Zorganizowana przez niego katedra będzie miała w tym roku 60-lecie.

Zygmunt Nawrocki (1910-1978), kontynuator myśli Jerzego Spławy-Neymana, był współtwórcą warszawskiej szkoły statystyki matematycznej i biometrii, rozwijanej nadal przez jego następców i uczniów na kilku wydziałach Szkoły Głównej Gospodarstwa

Wiejskiego (SGGW) w Warszawie (patrz Laudański, 2006). W roku 2008 ufundowano w tej uczelni tablice pamiątkowe poświęcone Jerzemu Spławie-Neymanowi i Zygmuntowi Nawrockiemu, jako prekursorom statystyki i biometrii w SGGW (patrz Laudański, 2008).

Julian Perkal (1913-1965), z wrocławskiej szkoły zastosowań matematyki Hugona Steinhausa (1887-1972), skupił wokół siebie grono wybitnych matematyków i przyrodników różnych specjalności, inspirując ich do stosowania metod statystycznych w badaniach eksperymentalnych. Hugo Steinhaus w swoich wspomnieniach (Steinhaus, 1992) pisze o nim: „znałem go bardzo dobrze, bo razem uprawialiśmy zastosowanie matematyki do medycyny, antropologii, antropometrii, pediatrii, stomatologii, transportu, komunikacji etc.".

Nic dziwnego, że właśnie Julianowi Perkalowi powierzono zorganizowanie towarzystwa biometrycznego w Polsce. Najpierw, w 1958 roku, jako Sekcji Biometrycznej Polskiego Towarzystwa Przyrodników im. Kopernika. Następnie, w 1961 roku, Sekcja ta przekształciła się w Polskie Towarzystwo Biometryczne (PTB). Towarzystwo to swoją rozwijającą się szybko aktywnością organizacyjną skupiło wielu naukowców i praktyków z rozmaitych dziedzin nauk przyrodniczych (w szerokim tego słowa znaczeniu, z naukami medycznymi włącznie) zainteresowanych zastosowaniami metod matematycznych i statystycznych w pracach badawczych.

W roku 1964 pod redakcją Juliana Perkala zaczęło wychodzić czasopismo *Listy Biometryczne*, organ PTB. Ukazywały się w nim artykuły przedstawiające matematyczne podstawy biometrii oraz omawiające metody przydatne w projektowaniu badań i analizowaniu ich wyników. Publikowano również sprawozdania z walnych zgromadzeń i sesji naukowych PTB, także materiały z rozmaitych konferencji tematycznych i kursów szkoleniowych organizowanych przez PTB. Niebawem zaczęły pojawiać się prace omawiające konkretne zastosowania metod biometrycznych w różnych dziedzinach badań. Publikowano również interesujące przeglądy dorobku biometrii w rozmaitych naukach (patrz np. Bogdanik, 1976; Oktaba, 1976; Welon, 1976).

Z czasem zaczęły tworzyć się pewnego rodzaju specjalizacje zastosowań biometrii. Wyraźnym przykładem tego było uruchomienie w 1970 r. (dzięki staraniom szkoły lubelskiej) serii corocznych konferencji pod nazwą *Colloquium Metodologiczne z Agrobiometrii*. Materiały z tych kolokwiów były publikowane w kolejnych tomach. W roku 2011 ukazał się tom 41, pod zmienioną nazwą *Colloquium Biometricum*. Konferencje te po pewnym czasie zaczęły bowiem przyciągać naukowców z poza nauk rolniczych, na przykład z nauk medycznych, co rozszerzyło zakres zainteresowań. Co więcej, kolokwia te nabrały charakteru międzynarodowego. Także Listy Biometryczne zmieniły nazwę na *Biometrical Letters* i coraz częściej zawierają artykuły autorów lub współautorów zagranicznych.

W istocie, biometria polska stosunkowo szybko wyszła ze swym dorobkiem naukowym poza granice Polski. Już w latach 50-tych i 60-tych pojawiają się pierwsze prace biometryków polskich publikowane w czasopismach zagranicznych. Prace te były

charakterystyczne dla podejmowanych wówczas badań, zwłaszcza w takich kierunkach jak:

– zastosowanie procesów stochastycznych w biologii i medycynie (patrz Urbanik, 1956),

– opracowanie wskaźników przyrodniczych (tzw. „wskaźników Perkala") i ich zastosowanie w analizie danych wielocechowych (patrz np. Perkal, Szczotka, 1960),

– zastosowanie metod biometrycznych w hodowli roślin (patrz np. Elandt, 1960),

– rozwój metod statystycznych, ze szczególnym uwzględnieniem planowania i analizy doświadczeń rolniczych (patrz np. Elandt, 1961, 1962, 1964; Elandt, Andrew, 1964; Oktaba, 1968, 1969), także serii doświadczeń (patrz np. Elandt, 1963; Caliński, 1966),

– zastosowanie analizy skupień w doświadczalnictwie (patrz np. Caliński, 1969).

Najwięcej z tych prac wyszło spod pióra Reginy Elandt (1918-2011), uczennicy Stefana Barbackiego.

Poczynając od lat 70-tych prace autorów polskich pojawiają się co raz częściej w zagranicznych czasopismach statystycznych i biometrycznych o zasięgu międzynarodowym, takich jak *Annals of Statistics*, *Applied Statistics*, *Australian & New Zealand Journal of Statistics*, *Biometrical Journal*, *Biometrics*, *Biometrika*, *Canadian Journal of Statistics*, *Communications in Statistics*, *Environmetrics*, *Computational Statistics and Data Analysis*, *Journal of Agricultural, Biological and Environmental Statistics*, *Journal of Multivariate Analysis*, *Journal of the Royal Statistical Society*, *Journal of Statistical Planning and Inference*, *Linear Algebra and Its Applications*, *Metrika*, *Sankhyā: The Indian Journal of Statistics*, *Scandinavian Journal of Statistics*, *Statistical Papers*, *Statistics and Probability Letters*, *Utilitas Mathematica* i w wielu innych. Dzisiaj już niemal w każdym czasopiśmie publikującym prace biometryczne można znaleźć artykuły autorów lub współautorów polskich.

Jednym z najczęstszych autorów polskich w tych czasopismach był Jerzy Baksalary (1944-2005), związany z poznańską szkołą statystyki matematycznej i biometrii, a od 1988 roku z Zieloną Górą. Jest autorem lub współautorem ponad 170 publikacji naukowych. Większość z nich ukazała się w wiodących czasopismach z algebry liniowej, statystyki matematycznej i biometrii. Wśród jego licznych (43) współautorów występują doktorzy honoris causa uczelni poznańskich, Leo C. A. Corsten z Holandii i Calyampudi Radhakrishna Rao z Indii i USA. W 2005 r. ukazał się specjalny tom (vol. 410) czasopisma *Linear Algebra and Its Applications* poświęcony Jerzemu Baksalaremu.

Warto także wspomnieć, że w wielu prestiżowych czasopismach specjalistycznych z różnych nauk szczegółowych znajdujemy prace, których współautorami są biometrycy polscy. Na przykład w czasopismach publikujących prace z genetyki, takich jak *Acta Biochimica Polonica*, *Archiv für Tierzucht*, *Genetics*, *Heredity*, *Journal of Animal and Feed Sciences*, *Journal of Animal Breeding and Genetics*, *Journal of Applied Genetics*, *Nucleic Acids Research*, *PLoS Biology*, *Theoretical and Applied Genetics*, znajdują się prace biometryczne autorów lub współautorów polskich, często cytowane w literaturze.

Wiele prac opublikowanych w wymienionych wyżej czasopismach to rezultat szeroko zakrojonej współpracy międzynarodowej, często w ramach międzynarodowych projektów badawczych. Jako przykład można tu wymienić prace, współautorów polskich, dotyczące metod biometrycznych stosowanych w badaniach ekspresji genów, realizowanych wspólnie z ośrodkami naukowymi we Włoszech (patrz np. Sari-Gorla i inni, 1997, 1999; Frova i inni, 1999), w Holandii, Niemczech i Wielkiej Brytanii (np. Kaufmann i inni, 2009) oraz w Szwajcarii (np. Okoniewski i inni, 2012).

O międzynarodowym znaczeniu biometrii polskiej świadczy również to, że od roku 1992 istnieje Polska Grupa Narodowa w ramach The International Biometric Society.

Podsumowując można stwierdzić, że polscy matematycy i przyrodnicy z różnych dziedzin wiedzy i praktyki badawczej nie zawiedli ojców biometrii w Polsce, Jana Czekanowskiego i Edmunda Załęskiego.

Uniwersytet Przyrodniczy w Poznaniu

LITERATURA

[1] Bogdanik T., (1976), Przegląd dorobku biometrii w naukach medycznych w Polsce w ostatnim 30-leciu, *Listy Biometryczne*, Nr 51-54, 29-47.

[2] Caliński T., (1966), On the distribution of the F-type statistics in the analysis of a group of experiments, *Journal of the Royal Statistical Society, Series B*, 28, 526-542.

[3] Caliński T., (1969), On the application of cluster analysis to experimental results, *Bulletin of the International Statistical Institute*, 42, 101-103.

[4] Caliński T., (1995), Statystyka matematyczna i biometria w ośrodku poznańskim, [w:] Palka Z. (red.), *Poznańska szkoła matematyczna* (s. 31-40), Wydawnictwo Naukowe UAM, Poznań.

[5] Caliński T., Dobek A., Kaczmarek Z., (2004), Stefan Barbacki i jego szkoła biometryczna w Poznaniu, [w:] Krajewski P., Zwierzykowski Z., Kachlicki P. (red.), *Genetyka w ulepszaniu roślin użytkowych* (s. 197-205), Instytut Genetyki Roślin PAN, Poznań.

[6] Czekanowski J., (1913), *Zarys metod statystycznych w zastosowaniach do antropologii*, Nakładem Towarzystwa Naukowego Warszawskiego, Warszawa.

[7] Elandt R.C., (1960), Biometric methods in plant breeding, *Acta Agronomica Academiae Scientiarum Hungaricae*, 10, 69-74.

[8] Elandt R.C., (1961), The folded normal distribution: Two methods of estimating parameters from moments, *Technometrics*, 3, 551-562.

[9] Elandt R.C., (1962), Exact and approximate power function of the non-parametric test of tendency, *Annals of Mathematical Statistics*, 33, 471-481.

[10] Elandt R.C., (1963), Optimal and sufficient allocation of multiple varietal experiments, *Biometrics*, 19, 615-628.

[11] Elandt R.C., (1964), Applicability of the extended method of parabolic curves in the analysis of agricultural data, *Sankhyā: The Indian Journal of Statistics, Series B*, 26, 201-216.

[12] Elandt R.C., Andrew G. M., (1964), Tables for application of the method of parabolic curves to a certain balanced systematic arrangement, *Sankhyā: The Indian Journal of Statistics, Series B*, 26, 17-28.

[13] Frova C., Krajewski P., di Fonzo N., Villa M., Sari-Gorla M., (1999), Genetic analysis of drought tolerance in maize by molecular markers. I. Yield components, *Theoretical and Applied Genetics*, 99, 280-288.

[14] Kaufmann K., Muiño J.M., Jauregui R., Airoldi C.A., Smaczniak C., Krajewski P., Angenent G.C., (2009), Target genes of the MADS transcription factor SEPALLATA3: Integration of development and hormonal pathways in the *Arabidopsis* flower, *PLoS Biology*, 7, 854-875.
[15] Krzyśko M., (2009), Jan Czekanowski anthropologist and statistician, *Acta Universitatis Lodziensis, Folia Oeconomica*, 228, 21-32.
[16] Laudański Z., (2006), Katedra Biometrii, [w:] Łabętowicz J. (red.), *100-lecie Wydziału Rolnictwa i Biologii 1906-2006*, *Tom 1* (s. 195-222), Wydawnictwo SGGW, Warszawa.
[17] Laudański Z. (2008). Prekursorzy statystyki i biometrii w SGGW, *Agricola*, *Pismo SGGW w Warszawie*, Nr 70, 9-10.
[18] Okoniewski M.J., Leśniewska A., Szabelska A., Zyprych-Walczak J., Ryan M., Wachtel M., Morzy T., Schäfer B., Schlapbach R., (2012), Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage. *Nucleic Acids Research*, 40 (9), e63, doi: 10.1093/nar/gkr1249.
[19] Oktaba W., (1968), A note on estimating the variance components by the method two of Henderson, *Biometrische Zeitschrift*, 10, 97-108.
[20] Oktaba W., (1969), Generalized inverses of matrices in a fixed model, *Biometrische Zeitschrift*, 11, 228-251.
[21] Oktaba W., (1973), Mikołaj Olekiewicz (1896-1971), *Wiadomości Matematyczne*, 16, 79-85.
[22] Oktaba W., (1976), Agro-biometria w Polsce, *Listy Biometryczne*, Nr 51-54, 10-28.
[23] Oktaba W., (1998), *Dziennik i wspomnienia*, Wydawnictwo Akademii Rolniczej w Lublinie, Lublin.
[24] Perkal J., Szczotka F., (1960), Eine neue Methode der Analyse eines Kollektivs von Merkmalen, *Biometrische Zeitschrift*, 2, 108-116.
[25] Sari-Gorla M., Caliński T., Kaczmarek Z., Krajewski P., (1997), Detection of QTL × environment interaction in maize by a least squares interval mapping method, *Heredity*, 78, 146-157.
[26] Sari-Gorla M., Krajewski P., di Fonzo N., Villa M., Frova C., (1999), Genetic analysis of drought tolerance in maize by molecular markers. II. Plant height and flowering, *Theoretical and Applied Genetics*, 99, 289-295.
[27] Schmidt S., (1971), Istotne aspekty w rozwoju zastosowań statystyki matematycznej w doświadczalnictwie rolniczym w Polsce (I), *Roczniki Nauk Rolniczych*, 79-G, 7-26.
[28] Steinhaus H., (1992), *Wspomnienia i zapiski*, Aneks Publisher, London.
[29] Urbanik K., (1956), On a problem concerning birth and death processes, *Acta Mathematica Academiae Scientiarum Hungaricae*, 7, 99-106.
[30] Welon Z., (1976), Biometria a antropologia w Polsce, *Listy Biometryczne*, Nr 51-54, 1-9.
[31] Wesołowska-Janczarek M., (2002), *50 lat Katedry Zastosowań Matematyki Akademii Rolniczej w Lublinie (1952-2002)*, Wydawnictwo Akademii Rolniczej w Lublinie, Lublin.
[32] Załęski E., (1927), *Metodyka doświadczeń rolniczych*, Wydawnictwo Rozpraw Biologicznych Nr 1, Lwów.

ROZWÓJ I OSIĄGNIĘCIA W BIOMETRII POLSKIEJ

Streszczenie

Biometria jest dyscypliną naukową zajmującą się zastosowaniami metod matematycznych i statystycznych w rozwiązywaniu problemów biologicznych, zwłaszcza w planowaniu i analizie eksperymentów. W Polsce pionierami biometrii było dwóch wybitnych uczonych. Antropolog, Jan Czekanowski (1882-1965), oraz chemik, agrotechnik i hodowca roślin, Edmund Załęski (1863-1932). Jednym z uczniów Załęskiego był Stefan Barbacki, uczony o ogromnych osiągnięciach w zakresie rozwoju metodyki doświadczalnictwa rolniczego i biometrii. Jego działalność przyczyniła się do stworzenia poznańskiej szkoły

statystyki matematycznej i biometrii. Podobnie rodziły się zainteresowania biometrią w innych polskich ośrodkach naukowych, zwłaszcza w Lublinie, dzięki Mikołajowi Olekiewiczowi i Wiktorowi Oktabie, w Warszawie, dzięki Jerzemu Spławie-Neymanowi i Zygmuntowi Nawrockiemu, oraz we Wrocławiu, dzięki Hugonowi Steinhausowi i Julianowi Perkalowi.

Słowa kluczowe: biometria, planowanie i analiza doświadczeń

THE DEVELOPMENT AND ACHIEVEMENTS IN POLISH BIOMETRY

Abstract

Biometry is a branch of science which deals with applications of mathematical and statistical methods to biological problems, particularly to the design and analysis of experiments. Two prominent scientists are considered as pioneers of biometry in Poland. An anthropologist, Jan Czekanowski (1882-1965), and a chemist, agricultural researcher and plant breeder, Edmund Załęski (1863-1932). One of Załęski's followers was Stefan Barbacki, a scientist of great achievements in the development of agricultural research methodology and biometry. His activity contributed essentially to the formation of the Poznań school of mathematical statistics and biometry. The interest in biometry in other Polish scientific centers was initiated in a similar way. Particularly in Lublin, due to Mikołaj Olekiewicz and Wiktor Oktaba, in Warsaw, due to Jerzy Spława-Neyman and Zygmunt Nawrocki, and in Wrocław, due to Hugo Steinhaus and Julian Perkal (1913-1965).

Key words: biometry, design and analysis of experiments

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

KRZYSZTOF JAJUGA

ROZWÓJ POLSKIEJ MYŚLI STATYSTYCZNEJ W NAUKACH EKONOMICZNYCH

Kongres Statystyki Polskiej jest okazją do przedstawienia historii osiągnięć polskich statystyków. W tym nurcie mieści się też ten artykuł. Jest w nim zawarta syntetyczna prezentacja rozwoju polskiej myśli statystycznej w naukach ekonomicznych. Główna uwaga skoncentrowana jest na postaciach tych polskich statystyków, którzy związani byli, i są, z naukami ekonomicznymi. Rozważania tego artykułu w pierwszej części korzystają z pracy Domańskiego (Domański, 2008).

1. POLSCY STATYSTYCY W NAUKACH EKONOMICZNYCH PRZED 1945

Na wstępie należy zaznaczyć, iż rozwój statystyki w dawnych latach był w dużym stopniu determinowany potrzebami gospodarki. W działalności makroekonomicznej istotna bowiem jest ewidencja ludności oraz aktywów gospodarczych. Czołowi polscy statystycy, związani z naukami ekonomicznymi bądź aktywnie uczestniczący w życiu gospodarczym, wnieśli wkład właśnie w obszarze szeroko rozumianej ewidencji. Warto jednak zaznaczyć, że przy okazji tejże ewidencji powstawały osiągnięcia o charakterze teoretycznym.

Fryderyk Józef Moszyński (ur. 14.03.1738 w Dreźnie, zm. 21.01.1817 w Kijowie) jest znany jako autor konstytucji sejmowej „Lustracja dymów i podanie ludności", która została uchwalona 22 czerwca 1789 roku. Konstytucja ta ustanowiła pierwszy spis powszechny, który został przeprowadzony w tym samym roku. Spis nie obejmował szlachty i duchowieństwa, dotyczył liczby ludności oraz struktury społeczno-zawodowej. Uważa się zatem Fryderyka Moszyńskiego za ojca statystyki państwowej. W trakcie Sejmu Czteroletniego Moszyński przedstawiał też projekty reform gospodarczych. Oprócz tego ten wybitny statystyk jest autorem metody ustalania wymiaru podatku na cele wojskowe. W metodzie tej wartość dóbr jest wyliczana na podstawie aktów sprzedaży oraz liczby dymów (dziś byłaby to pewnie liczba domów). Moszyński w 1788 roku ustalił liczbę ludności Polski na 7 354 620 osób, przyjmując założenie, iż jednemu dymowi odpowiada 6 osób.

Stanisław Staszic (ur. 6.11.1755 w Pile, zm. 20.01.1826 w Warszawie) w statystyce jest znany jako autor rozprawy wydanej w 1807 pt. „O statystyce Polski, krótki rzut wiadomości potrzebnych tym, którzy ten kray chcą oswobodzić i tym, którzy w nim chcą rządzić". W rozprawie tej po raz pierwszy pojawia się słowo „statystyka" w języku

polskim. Oprócz tego praca ta zawiera propozycje reform gospodarczych, analizę sytuacji wewnętrznej i zewnętrznej Polski, wszystko to wspierane danymi statystycznymi. Warto podać tytuły poszczególnych części tej pracy (oryginalne nazewnictwo):

„Rozległość polskiej ziemi topograficznie uważana"; „Departamenta w równinach i krajach zapadłych"; „Departamenta w krajach zapadłych"; „Rzeki spławne w Polsce"; „Jaką ilość ziemi zajmują miasta, drogi, błota, lasy, role i łąki"; „Wiele zboża wychodzi z Polski"; „W jakim stanie miasta"; „Jaka ludność całej Polski"; „Jak wiele Polska wysławić i utrzymać może teraz wojska"; „Jakie podatki Polska teraz składać może"; „Jakie są stosunki polityczne Polski z Francją"; „Stosunki handlowe Polski z Francją".

W pracy tej Staszic podaje, że w 1776 roku ludność Polski liczyła około 14 milionów osób (liczba ta wyraźnie różni się od podanej przez Moszyńskiego) i szacuje jej powiększenie o kolejny milion w ciągu półwieku.

Wawrzyniec Surowiecki (ur. 4.08.1769 w Imielenku, zm. 10.06.1827 w Warszawie) był pierwszym polskim wykładowcą statystyki i kierownikiem pierwszej katedry statystyki, utworzonej w 1811 roku w Szkole Głównej Prawa i Administracji w Warszawie. Warto wymienić tu dwie prace napisane przez Wawrzyńca Surowieckiego. Pierwsza praca to „O upadku przemysłu i miast w Polsce" z 1810 roku, w której rozpowszechniał zasady ekonomii wprowadzone przez Adama Smitha oraz proponował reformy rynkowe w polskim przemyśle. Druga praca to „Statystyka Księstwa Warszawskiego" z 1812 roku, będąca skryptem napisanym na podstawie własnych wykładów. W tej drugiej pracy przedstawił przedmiot i zadania statystyki.

Dominik Krysiński (ur. 1785, zm. 17.04.1983 w Warszawie) jest autorem dwóch prac: „O ekonomii politycznej" z 1812 oraz „O arytmetyce politycznej" z 1814. Krysiński traktował statystykę jako metodę tzw. arytmetyki politycznej, która to metoda służyła do opisu zbiorowości. Uważał, że w analizie zjawisk gospodarczych należy uwzględniać charakterystyki jakościowe. Przez praktyków uważany jest za prekursora badań budżetów gospodarstw domowych.

Józef Kleczyński (ur. 27.10.1841 w Ihnatowie, zm. 21.09.1900 w Zakopanem) napisał pracę „Miejskie biura statystyczne", w której proponował tworzenie lokalnych urzędów statystycznych. Jest też autorem opublikowanej w 1879 po niemiecku pracy „O obliczaniu liczby ludności w latach między spisami". Był redaktorem czasopisma „Statystyka miasta Krakowa". Ponadto napisał inne prace z zakresu organizacji statystyki, np. „Statystyka gospodarstwa gminnego", „Spisy ludności w Rzeczypospolitej Polskiej", „Spis ludności diecezji krakowskiej z roku 1787".

Adam Bolesław Danielewicz (ur. 23.12.1846 w Koźminku, zm. 25.06.1935 w Warszawie) zajmował się przede wszystkim zagadnieniami statystyki ubezpieczeniowej oraz demografii. Jest autorem pracy z 1896 pt. „Podstawy matematyczne ubezpieczeń życiowych", oraz innych prac z zakresu ubezpieczeń życiowych, jak: „Zasady taryf ubezpieczeń życiowych" z 1878. Ponadto w 1878 napisał artykuł „W przedmiocie badań

tablic śmiertelności", w którym przedstawił tablice umieralności dla Warszawy stosując metodę Halleya.

Jan Wiśniewski (ur. 16.10.1904 w Jadowie, zamordowany w Charkowie wiosną 1940) zajmował się teoretycznymi i praktycznymi zagadnieniami statystyki. Jeśli chodzi o badania teoretyczne, to najbardziej jest znany z artykułów opublikowanych w słynnym „Journal of American Statistical Association".

W artykule z 1930 roku pt. „A note on interpolation", Jan Wiśniewski podaje wzór interpolacyjny dla rozwiązania następującego zagadnienia: znając średnie arytmetyczne funkcji w podprzedziałach argumentu, znaleźć przybliżoną wartość funkcji dla dowolnego argumentu należącego do całego przedziału będącego sumą tych podprzedziałów. W artykule z 1931 roku pt. „Extension of Fisher's formula number 353 to three or more variables" (jako ciekawostkę warto wspomnieć, że w tym samym numerze tego renomowanego czasopisma ukazał się artykuł Ludwika Landau o organizacji polskiej statystyki) Jan Wiśniewski uogólnił na przypadek więcej niż dwóch zmiennych dwa wyniki dotyczące indeksów statystycznych podane przez Irvinga Fishera.

W artykule z 1934 pt. „Pitfalls in the computation of the correlation ratio" Jan Wiśniewski zajmuje się problemem obciążeń w obliczeniach współczynnika korelacji, gdy przedział zmienności jednej ze zmiennych staje się wąski. Prowadził też badania nad rozkładem dochodów w Polsce za pomocą rozkładu logarytmiczno-normalnego i rozkładu Pareto oraz badania dynamiki cen oraz cykli koniunkturalnych.

Kolejne dwie osoby mają zasługi przede wszystkim w obszarze organizacji statystyki.

Ludwik Krzywicki (ur. 21.08.1859 w Płocku, zm. 10.06.1941 w Warszawie) jest współautorem aktu Rady Regencyjnej z 13 lipca 1918 roku o utworzeniu i organizacji Głównego Urzędu Statystycznego. Akt ten umożliwił powstanie pierwszego w niepodległej Polsce urzędu statystycznego. Ludwik Krzywicki potem współorganizował powstanie Głównego Urzędu Statystycznego i był jego wicedyrektorem w latach 1918-1925.

Edward Szturm de Sztrem (ur. 18.07.1885 w Sankt Petersburgu, zm. 9.09.1962 w Warszawie) był dyrektorem Głównego Urzędu Statystycznego w latach 1929-1939 oraz Prezesem Polskiego Towarzystwa Statystycznego po jego reaktywacji w 1937. Prowadził badania w zakresie statystyki rolnej, warto tu wymienić pracę z 1927 pt. „Kształtowanie się cen na ważniejsze artykuły rolne w Polsce". Zainicjował też badania ankietowe dotyczące płac.

2. BADANIA STATYSTYCZNE W NAUKACH EKONOMICZNYCH PO 1945

Przedstawienie najważniejszych osiągnięć polskich statystyków po drugiej wojnie światowej rozpoczniemy od krótkiej prezentacji uczonych, którzy zajmowali się statystyką jeszcze przed wojną, zaś kontynuowali je po odzyskaniu niepodległości.

Stefan Szulc (ur. 19.12.1881 w Prażuchach, zm. 12.10.1956 w Warszawie) jest uważany za współtwórcę powojennego systemu statystyki w Polsce. Pełnił funkcję Prezesa Głównego Urzędu Statystycznego w latach 1945-1949. W zakresie działalności naukowej i dydaktycznej Stefan Szulc znany jest jako autor dwutomowej pracy „Metody statystyczne", wydanej w 1952 roku.

Wincenty Styś (ur. 30.07.1903 w Husowie, zm. 21.04.1960 we Wrocławiu) jest autorem wielu badań w zakresie statystyki rolnej, z których część powstała jeszcze przed wojną, np. „Rozdrabnianie gruntów chłopskich w byłym zaborze austriackim od roku 1787 do 1931" (Towarzystwo Naukowe we Lwowie, 1934), „Wpływ uprzemysłowienia na ustrój rolny" (Towarzystwo Naukowe we Lwowie, 1936), „Drogi postępu gospodarczego wsi. Studium szczegółowe na przykładzie zbiorowości próbnej wsi Husowa" (Towarzystwo Naukowe we Lwowie, 1938). Po wojnie kontynuował badania w obszarze statystyki wsi, publikując pracę „Współzależność rozwoju rodziny chłopskiej i jej gospodarstwa" (Wrocławskie Towarzystwo Naukowe, Wrocław, 1959). Wincenty Styś był współorganizatorem w 1947 i jednym z rektorów wrocławskiej uczelni ekonomicznej (najpierw Wyższa Szkoła Handlowa, obecnie Uniwersytet Ekonomiczny).

Oskar Lange (ur. 27.07.1904 w Tomaszowie Mazowieckim, zm. 2.10.1965 w Londynie) zasłynął najbardziej w teorii ekonomii, jednak znane są również jego związki ze statystyką. Jest współautorem wydanego w 1952 roku podręcznika „Teoria statystyki". W 1955 opublikował, w czasopiśmie „Colloquium Mathematicum" wydawanym przez Instytut Matematyczny PAN, artykuł pt. „Statistical estimation of parameters in Markov process". Jednak największe związki Oskara Langego ze statystyką wiążą się z popularyzacją narzędzi statystycznych oraz ekonometrycznych na polskich uczelniach w latach pięćdziesiątych oraz w pierwszej połowie lat sześćdziesiątych.

Ustrój polityczny Polski, który dominował do końca lat osiemdziesiątych ubiegłego stulecia, nie sprzyjał rozwojowi nauk ekonomicznych, a w związku z tym również zastosowaniom metod statystycznych w badaniach ekonomicznych. Niemniej jednak obserwowany był rozwój w zakresie samych metod i narzędzi statystycznych. Dopiero po roku 1990 nastąpił bardziej dynamiczny rozwój w zakresie zastosowań metod statystycznych w badaniach zjawisk ekonomicznych, a także we wspomaganiu decyzji.

Analiza dorobku polskich statystyków działających w obszarze nauk ekonomicznych pozwala na wyodrębnienie kilku głównych obszarów badawczych. Są nimi:

– Klasyfikacja i analiza danych wielowymiarowych;
– Wnioskowanie statystyczne;
– Statystyka małych obszarów;
– Badania empiryczne.

Z uwagi na charakter artykułu, który dotyczy polskich statystyków, poniższy opis nie obejmuje bogatego dorobku polskiej ekonometrii, która dynamicznie rozwija się co najmniej od połowy lat sześćdziesiątych ubiegłego stulecia i zaowocowała wieloma osiągnięciami zarówno w zakresie metod, jak i zastosowań.

3. KLASYFIKACJA I ANALIZA DANYCH WIELOWYMIAROWYCH

Pionierem badań w tym zakresie w polskim środowisku ekonomistów jest Zdzisław Hellwig, który w 1968 roku opublikował w „Przeglądzie Statystycznym" artykuł pt. „Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr". W tym artykule zaproponował metodę wzorca rozwoju, służącą do porządkowania liniowego obiektów, charakteryzowanych wieloma zmiennymi. Ponadto w tym artykule rozpowszechnił w środowisku ekonomistów metodę taksonomii wrocławskiej, która jak wiadomo pozwala na utworzenie klas obiektów charakteryzujących się zbliżonymi wartościami wielu zmiennych.

Idea klasyfikacji (*cluster analysis*) polega najczęściej na utworzeniu klas obiektów na podstawie odległości między obiektami w przestrzeni wielowymiarowej, tak aby odległości między obiektami należącymi do tej samej klasy były jak najmniejsze, a odległości między obiektami należącymi do różnych klas były jak największe.

Na początku obszarem zainteresowań polskich ekonomistów były jedynie metody klasyfikacji i porządkowania liniowego, z czasem jednak badania objęły inne metody analizy danych, takie jak: analiza głównych składowych, analiza korelacji kanonicznej, analiza danych jakościowych itp. Główną cechą charakterystyczną badań było zastosowanie podejścia opisowego (deterministycznego), czyli niezakładającego losowości wektora zmiennych (a w związku z tym konkretnej postaci rozkładu wielowymiarowego). Metody klasyfikacji i analizy danych na początku były rozwijane głównie w uczelniach ekonomicznych w ośrodku wrocławskim (także jeleniogórskim), krakowskim i katowickim, później również w innych ośrodkach (np. łódzki, gdański).

Warto wymienić przynajmniej część osób, które wniosły wkład w rozwój tych metod:

Ośrodek wrocławski: Zdzisław Hellwig (metoda pojemności informacji wyboru zmiennych regresyjnych, zagadnienie przechodniości korelacji), Stanisława Bartosiewicz (metoda wyboru zmiennych regresyjnych), Maria Cieślak, Danuta Strahl, Marek Walesiak, Tadeusz Borys, Krzysztof Jajuga;

Ośrodek krakowski: Józef Pociecha, Aleksander Zeliaś, Kazimierz Zając, Jan Steczkowski, Andrzej Sokołowski, Tadeusz Grabiński;

Ośrodek katowicki: Elżbieta Stolarska, Józef Kolonko, Krzysztof Zadora, Andrzej Barczak, Eugeniusz Gatnar.

Rozwój badań w zakresie klasyfikacji i analizy danych wiąże się niewątpliwie z Sekcją Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego (SKAD). W latach 1979-1987 Akademia Ekonomiczna (obecnie Uniwersytet Ekonomiczny) we Wrocławiu, pod patronatem Komitetu Statystyki i Ekonometrii PAN, zorganizowała pięć konferencji naukowych poświęconych zagadnieniom taksonomicznym.

Z inspiracji profesorów Zdzisława Hellwiga, Kazimierza Zająca i Józefa Pociechy powstał w 1988 zamiar utworzenia ogólnopolskiej Sekcji Taksonomicznej przy Polskim Towarzystwie Statystycznym. W wyniku podjętych w tym względzie działań

Rada Główna Polskiego Towarzystwa Statystycznego w dniu 9 listopada 1988 r. podjęła decyzję o powołaniu Sekcji Taksonomicznej PTS. Sekcja Taksonomiczna zorganizowała w 1989 pierwszą konferencję naukową nt. „Taksonomia: teoria i jej zastosowania". W 1992 roku zmieniono nazwę Sekcji Taksonomicznej PTS na Sekcję Klasyfikacji i Analizy Danych PTS. W 1994 Sekcja Klasyfikacji i Analizy Danych PTS została przyjęta do International Federation of Classification Societies (IFCS). Obecnie jest to czwarte co do liczebności towarzystwo naukowe działające w ramach tej federacji.

Ważnym wydarzeniem dla Sekcji SKAD PTS było powierzenie przez Zarząd IFCS organizacji światowej konferencji IFCS. Organizacji konferencji, która miała miejsce w Krakowie w dniach 16-19 lipca 2002, podjął się Andrzej Sokołowski. Za program naukowy tej konferencji odpowiadał Krzysztof Jajuga. Część referatów przedstawionych w trakcie konferencji została opublikowana w książce wydanej przez Springer-Verlag. Warto też wspomnieć, że w latach 1979-2012 Sekcja zorganizowała w kraju 26 konferencji poświęconych klasyfikacji i analizie danych.

4. WNIOSKOWANIE STATYSTYCZNE

Badania naukowe w zakresie wnioskowania statystycznego prowadzone są od wielu lat. Wiele metod powstało jeszcze przed 1990. Można powiedzieć, że badania te były rozwijane w każdym większym ośrodku związanym z naukami ekonomicznymi.

Lista osób, które prowadziły i prowadzą badania w tym obszarze jest duża, warto wskazać na czołowe postacie:

Zdzisław Hellwig – test zgodności dla małej próby; Czesław Domański – badania nad testami opartymi na statystykach rangowych, analiza mocy testów, wartości krytyczne testów normalności; Andrzej Tomaszewicz – wnioskowanie statystyczne przy nieklasycznych założeniach modeli liniowych; Jerzy Greń – metody estymacji bayesowskiej; Ryszard Zasępa, Jan Kordos, Czesław Bracha, Janusz Wywiał – metoda reprezentacyjna; Wiesław Sadowski, Jerzy Greń – statystyczna teoria gier i decyzji; Zbigniew Czerwiński – zależność stochastyczna; Zbigniew Maria Pawłowski – prognozy statystyczne.

5. STATYSTYKA MAŁYCH OBSZARÓW

Ten obszar badań rozwinął się głównie po roku 1990, co wiąże się po pierwsze z rozwojem statystyki regionalnej, po drugie, z zwiększoną współpracą międzynarodową w tym obszarze. Niewątpliwie pionierem badań w zakresie statystyki małych obszarów w Polsce jest Jan Kordos. Obecnie badania są intensywnie rozwijane w ośrodku poznańskim (Jan Paradysz, Grażyna Dehnel, Elżbieta Gołata), w szczególności są to prace dotyczące dobroci estymacji.

6. BADANIA EMPIRYCZNE

Badania empiryczne były prowadzone również przed rokiem 1990. Tutaj należy wymienić prace nad indeksami statystycznymi (Kazimierz Romaniuk, Władysław Welfe) oraz badania płac i budżetów gospodarstw domowych (Jan Kordos, Kazimierz Zając, Andrzej Luszniewicz, Mirosław Krzysztofiak).

Po roku 1990 znacznie wzrosła dostępność danych statystycznych, z których może korzystać środowisko ekonomistów. Mowa tu przede wszystkim o danych finansowych, marketingowych, regionalnych. Wszystko to spowodowało, że znacznie wzrosła waga badań empirycznych.

Lista osób oraz ośrodków, które intensywnie rozwijają badania w poszczególnych obszarach empirycznych, jest następująca:

– Badania jakości życia: Tomasz Panek, zespół Katedry Statystyki Uniwersytetu Ekonomicznego we Wrocławiu pod kierunkiem Walentego Ostasiewicza;

– Badania regionalne: Danuta Strahl, ośrodek poznański (Jan Paradysz z zespołem), ośrodek krakowski (Kazimierz Zając, Barbara Podolec, Aleksander Zeliaś z zespołem);

– Badania sondażowe: Mirosław Szreder;

– Badania marketingowe: ośrodek krakowski, ośrodek jeleniogórski.

Najważniejsze tendencje, jakie mogą być zaobserwowane w badaniach statystycznych prowadzonych przez badaczy związanych z naukami ekonomicznymi w ostatnich kilkudziesięciu latach, są następujące:

– równomierny rozwój nurtu stochastycznego (rozkłady zmiennych losowych) oraz opisowego w badaniach metodycznych;

– mocny nurt aplikacyjny, wynikający ze zwiększonej dostępności oraz wzrostu znaczenia danych regionalnych, finansowych i marketingowych;

– rozwój badań ekonometrycznych, które w sposób naturalny wyłoniły się pół wieku temu z badań statystycznych;

– bardzo bogata działalność dydaktyczna, wynikająca przede wszystkim ze wzrostu liczby studentów w latach dziewięćdziesiątych ubiegłego stulecia.

Uniwersytet Ekonomiczny we Wrocławiu

LITERATURA

[1] Domański C., (2008), Zasłużeni statystycy dla nauki, referat wygłoszony z okazji 90-lecia GUS, http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Zasluzeni_statystycy_dla_nauki.pdf.

ROZWÓJ POLSKIEJ MYŚLI STATYSTYCZNEJ W NAUKACH EKONOMICZNYCH

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

W artykule przedstawione są postacie i osiągnięcia polskich statystyków prowadzących badania w naukach ekonomicznych. Odrębnie analizowano okres przed 1945 oraz po 1945. Analiza tego drugiego okresu prowadzona jest na podstawie klasyfikacji obejmującej kilka obszarów badawczych.

Słowa kluczowe: statystyka, klasyfikacja, analiza danych wielowymiarowych, wnioskowanie statystyczne, statystyka małych obszarów

THE DEVELOPMENT OF POLISH STATISTICS IN ECONOMIC SCIENCES

A b s t r a c t

The paper presents main scientific achievements of Polish statisticians conducting research in economic sciences. Two periods are distinguished: before and after 1945. Analysis for the second period is conducted in the framework of classification into several research areas.

Key words: statistics, classification, multivariate data analysis, statistical inference, small area statistics

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

JAN KORDOS

WSPÓŁZALEŻNOŚĆ POMIĘDZY ROZWOJEM TEORII I PRAKTYKI BADAŃ REPREZENTACYJNYCH W POLSCE

1. WSTĘP

Praktyczne problemy napotkane w czasie projektowania, realizacji i analizy badań reprezentacyjnych wpłynęły w znacznej mierze na rozwój teorii tych badań. Z drugiej strony, teoria badań reprezentacyjnych wpływała na praktykę tych badań, co prowadziło często do ich udoskonalenia (O'Muircheartaigh, Wong, 1981; Rao, 2005). Wydaje się więc celowe, aby prześledzić w jakim stopniu występowała współzależność pomiędzy rozwojem teorii i praktyki badań reprezentacyjnych w Polsce, uwzględniając postępy w tej dziedzinie badań na świecie.

Szeroki rozwój badań reprezentacyjnych w ostatnich latach związany jest w znacznym stopniu z rozwojem teorii, na której opierają się badania reprezentacyjne jak również z rozwojem swoistej sztuki związanej z przygotowaniem i prowadzeniem tych badań (Deming, 1950; Hansen et al., 1953; Kruskal i Mostller, 1979-1980; Mahalanobis, 1944; O'Muircheartaigh et al.,1981; Rao, 2005; Smith, 1976, 1994). Chodzi tu nie tylko o rozwój teorii, na której opiera się metoda reprezentacyjna, ale także o inne istotne komponenty badań reprezentacyjnych, w których znaczną rolę odgrywa nie tylko rachunek prawdopodobieństwa i statystyka matematyczna, ale także prakseologia, socjologia i psychologia (Cochran, 1977; Hansen et al.,1953; Kish, 1965; Kordos, 2000a; O'Muircheartaigh et al., 1981; Särndal et al., 1992; Yates, 1980; Zasępa, 1972). Dotyczy to takich komponentów badań jak: a) wybór metody zbierania danych, a w szczególności projektowanie kwestionariuszy, b) szkolenie personelu przeprowadzającego badanie, c) kontrola badania na różnych etapach jego przygotowania, realizacji i opracowania, d) badania na poszczególnych etapach źródeł błędów, ich rodzajów i rozmiarów, a także możliwości ich minimalizacji, e) metod przetwarzania danych, f) metod analizy uzyskanych wyników oraz ich prezentacji i rozpowszechniania.

Dotychczas nie ma jednolitej teorii badań statystycznych, a poszczególne komponenty projektowania badań opierają się na procedurach praktycznych uznawanych za odpowiednie w danych warunkach lub na fragmentarycznych rozwiązaniach teoretycznych, które przystosowuje się do potrzeb praktycznych (Hansen et al.,1953; Kish, 1965, 1987, 1996). Na mocnych podstawach teoretycznych opierają się plany losowania próby, ustalania potrzebnej jej liczebności w konkretnym badaniu, choć zwykle dla wybranych parametrów, wybór metod estymacji i oceny rozmiarów błędów losowych

(Bracha, 1996; Hansen et al., 1953; Smith, 1976, 1994; Zasępa, 1972). Tym zagadnieniom statystycy teoretycy i metodolodzy poświęcali najwięcej uwagi. Inne komponenty projektowania badań reprezentacyjnych opierają się w większym stopniu na sztuce niż nauce.

Z wielu zagadnień metodologicznych i tematów badań reprezentacyjnych wymienimy tu najważniejsze, które dotyczą (Smith, 1976, 1994; Kordos, 2000a): a) wyboru próby, b) metod estymacji, c) oceny błędów losowych i sposobów ich obliczenia, d) metod zbierania danych oraz e) oceny wszelkiego rodzaju błędów nielosowych.

W okresie późniejszym nabrały dużego znaczenia, takie tematy jak: (i) korekta danych i ich imputacja oraz kalibracja, (ii) koordynacja badań, (iii) estymacja dla małych obszarów, (iv) badania powtarzalne, (v) wykorzystanie komputerów. Naszą uwagę skupimy głównie na wykazaniu współzależności pomiędzy rozwojem teorii i praktyki badań reprezentacyjnych w GUS.

Za początek badań reprezentacyjnych w statystyce publicznej w Polsce uważa się rok 1933, w którym J. Neyman opublikował monografię poświęconą zastosowaniu *metody reprezentacyjnej* do przyśpieszenia wyników spisu ludności z 1931 r. (Neyman, 1933). Tak więc, badania reprezentacyjne w sensie losowego wyboru próbki zostały zapoczątkowane w Polsce przez J. Neymana (1933) oraz J. Piekałwiewicza (1934). Współzależność tutaj jest bardzo scisła, gdyż Neyman opracował zarys teorii i praktyki, a Piekałkiewicz zaprezentował i zaanalizował uzyskane wyniki z badania.

W połowie lat dwudziestych ubiegłego wieku J. Neyman zajmował się już teorią metody reprezentacyjnej, a szczególnie wyborem małych próbek ze skończonych populacji (Neyman, 1925). Sławę światową przyniósł mu dopiero słynny artykuł opublikowany w języku angielskim w 1934 r. (Neyman, 1934), ale główne myśli tego artykułu zostały opublikowane najpierw w języku polskim w 1933 r. (Neyman, 1933). Rozważymy tu więc badania częściowe oparte na próbkach, których jednostki do badania były wybrierane według *procedury probabilistycznej.* Zostaną więc pominięte badania monograficzne i inne badania częściowe, do których jednostki badania zostały wybierane w sposób niezgodny z tą procedurą badawczą.

Na początku XX wieku pod pojęciem metody reprezentacyjnej kryła się procedura wyboru próby z całej populacji, która umożliwiała uzyskanie pewnego rodzaju miniatury populacji. Próba mogła być wybrana zarówno w sposób celowy jak i losowy. Dopiero fundamentalna praca J. Neymana z 1934 r. (Neyman, 1934) położyła podstawy pod losowy wybór próby. Przedstawimy więc badania reprezentacyjne, w których jednostki badania zostały wybrane według określonej procedury probabilistycznej, umożliwiającej uogólnienie uzyskanych wyników z próby na całą populację. Metoda reprezentacyjna zdobywała sobie stopniowo uznanie w statystyce publicznej, chociaż ze znacznymi oporami (Kish, 1996). Rozwijała się więc nie tylko teoria badań reprezentacyjnych, ale także, jak już wspomniano, sztuka ich projektowania i realizacji w różnych dziedzinach badań statystycznych (Kordos, 2000a). Szersze zastosowania metody reprezentacyjnej w statystyce oficjalnej w Polsce rozpoczęły się po II wojnie światowej. Znaczną rolę odegrał tu także J. Neyman, który w czasie swych wizyt w Polsce w latach 1950

i 1958 udzielał konsultacji dla GUS w zakresie zastosowania metody reprezentacyjnej w badaniach statystycznych (Fisz, 1950a; Zasępa, 1958). Szczególną role odegrały tu przede wszystkim:

a) Komisja Matematyczna GUS działająca w latach 1950-1993, w skład której wchodzili specjaliści z ośrodków akademickich i przedstawiciele GUS (Kordos, 1975, 2012a);
b) Pracownia Metod Matematycznych, a następnie Pracownia ds. Statystyki Matematycznej Zakładu Badań Statystyczno-Ekonomicznych GUS w latach 1966-2004;
c) Zespół ds. badań reprezentacyjnych w Departamencie Programowania i Organizacji Badań GUS w latach 1995-2004, a od 2005 r. Wydział Operatów Statystycznych w Departamencie Metodologii, Standardów i Rejestrów GUS.

Metodologią badań reprezentacyjnych zajmowały się także w ciągu ostatnich 50-ciu lat pracownicy naukowi wyższych uczelni oraz jednostek naukowo-badawczych, których prace dotyczyły głównie planów prób, schematów losowania, liczebności prób i ich lokalizacji, metod estymacji oraz oceny błędów losowych i nielosowych. Łącznie literatura przedmiotu jest bardzo bogata i liczy kilkaset pozycji. Z konieczności w tym artykule okolicznościowym ograniczyłem się tylko do wybranych pozycji charakteryzujących prezentowane kierunki badań i praktycznych zastosowań. Z tych względów wiele pozycji teoretycznych z metody reprezentacyjnej nie mogły być tu uwzględnione.

Szczególną uwagę poświęcę:

1. pracom teoretycznym i metodologicznym rozwijanym przez polską statystykę akademicką i w ośrodkach naukowo-badawczych, mających istotny wpływ na praktykę statystyczną,
2. badaniom powtarzalnym w czasie, a głównie *metodzie rotacyjnej,*
3. badaniom jakości danych, a szczególnie metodom prodzadzącym do zwiększenia ich dokładności,
4. metodom estymacji dla małych obszarów.

Przed tym jednak przedstawię zarys rozwoju teorii i praktyki badań reprezentacyjnych na świecie, aby na tym tle przedstawić rozwój badań reprezentacyjnych prowadzonych Polsce.

2. ZARYS ROZWÓJU TEORII I PRAKTYKI BADAŃ REPREZENTACYJNYCH NA ŚWIECIE

Należy podkreślić, że *Międzynarodowy Instytut Statystyczny (MIS)* uznał metodę reprezentacyjną w statystyce oficjalnej dopiero w 1925 r., a przecież już w 1891 r. w spisie ludności w Norwegii wykorzystano metodę badania częściowego, która właśnie została nazwaną *metodą reprezentacyjną.* Metoda ta była zaprezentowana na sesji Międzynarodowego Instytutu Statystycznego w Bernie w 1895 r. przez A. Kiaera (1897). Dlatego rok 1895 uważany jest jako początek badań reprezentacyjnych w statystyce oficjalnej. A. L. Bowley – pierwszy profesor statystyki w London School of Economics and Political Science – przez swoje prace o doborze próby (Bowley, 1926) oraz o pomiarze (Bowley, 1915) był pierwszą z głównych postaci rozwoju naukowe-

go doboru próby, tj. opartego na zasadach rachunku prawdopodobieństwa. Gdy dziś powszechnie uznaje się zasady losowego pobierania próby w badaniach społecznych i gospodarczych, warto zauważyć, że propozycja Kiaera na sesji MIS w 1895 r. nie była przychylnie przyjęta (Kish, 1996). Kiaer dowodził, że badanie częściowe może dać rzetelne wyniki, pod warunkiem, że obserwacje tworzą reprezentatywny obraz całej badanej zbiorowości. Próba powinna, jego zdaniem, odzwierciedlać populację ze względu na ważne cechy. Uważano wtedy, że reprezentatywność próby może być osiągnięta na dwa sposoby:

a) przez wybór zrandomizowany (losowy) z całej populacji oraz

b) przez wybór celowy, który odzwierciedla populację ze względu na wybrane zmienne.

W pierwszych dekadach XX wieku badania reprezentacyjne były przeprowadzane przy wykorzystaniu jednej z wyżej podanych metod wyboru, a do połowy lat trzydziestych ubiegłego wieku, na pytanie która z nich jest „lepsza", nie znano odpowiedzi. Dopiero fundamentalny wkład w rozwój teorii metody reprezentacyjnej wniósł, wspomniany już, *polski statystyk Jerzy Neyman* w swoim słynnym artykule z 1934 r. (Neyman, 1934), który zmienił teoretyczne podstawy wnioskowania na podstawie badań częściowych, wprowadzając błędy losowe oparte na randomizacji rozkładu. Wprowadził *optymalną lokalizację próby* w losowaniu warstwowym, a także *przedziały ufności* (Neyman, 1935, 1938). Użył on terminu „*reprezentatywny"* w nowym znaczeniu. Warto dodać, że w literaturze anglojęzycznej termin „*metoda reprezentacyjna*" (representative method) obecnie jest rzadko używany, ale w języku polskim jest dalej dominującym. Wybór losowy w warstwach jest, według Neymana, reprezentatywny, nawet gdy frakcje losowania w poszczególnych warstwach znacznie się różnią. Uświadomienie sobie tego, że równe prawdopodobieństwo wyboru nie jest konieczne dla ważności wnioskowania, było wielkim krokiem naprzód. Neyman zalecał podejście probabilistyczne (tj. losowe) przy wyborze próby, a krytycznie odnosił się do celowego wyboru próby.

Od tego czasu teoria i praktyka badań reprezentacyjnych rozwijały się intensywnie (Kish, 1996; O'Muircheartaigh, et al., 1981;. Smith, 1994; Rao, 2005). W latach 1950-tych prawie zakończono rozwój teorii randomizacji, a zainteresowanie statystyków przesunęło się stopniowo z *błędów losowych* w kierunku *błędów nielosowych*, do czego w szczególnym stopniu przyczynił się M.H. Hansen i jego współpracownicy (Hansen i Hurwitz, 1943, 1946; Hansen, Hurwitz i Madow, 1953; Hansen, Hurwitz i Bershad, 1961). W tym czasie obserwuje się znaczny rozwój badań reprezentacyjnych, chociaż nie zawsze zgodnie z aktualnym rozwojem teorii. Niejednokrotnie praktyka wyprzedzała teorie (Kruskal i Mosteller, 1979-1980; O'Muircheartaigh, et al., 1981; Smith, 1994). Dotyczy to szczególnie początków lat siedemdziesiątych. Trudno tu ocenić wkład w tym okresie poszczególnych autorów, dlatego ograniczę się do omównienia ważniejszych kierunków prac badawczych.

Wprawdzie badania reprezentacyjne na szerszą skalę rozpoczęły się po 1945 r., to jednak w pierwszej połowie XX stulecia warto także odnotować kilka faktów. Należy wspomnieć o badaniach prowadzonych i opisanych przez A.L. Bowleya'a (1913, 1926)

w Anglii, a także badania bezrobocia prowadzone w latach 1934-1942 w USA (Frankel i Stock, 1942), które w 1943 r. przekształciły się w sławne badania siły roboczej prowadzone przez Biuro Spisów USA.

Po 1945 r. w wielu krajach nastąpił intensywny rozwój badań reprezentacyjnych. W tym czasie ukazało się 5 podręczników z metody reprezentacyjnej w języku angielskim: Yates (1949), Deming (1950), Cochran (1953), Hansen et al. (1953) oraz Sukhatme i Sukhatme (1970). Po ukazaniu się wymienionych podręczników zorganizowano wiele kursów z metody reprezentacyjnej w USA, Wielkiej Brytanii, w Indiach i wielu innych krajach.

W USA badania prowadzone przez Biuro Spisów przekształciły się w Bieżące Badania Ludności (*Current Population Surveys*), które pod kierunkiem M.H. Hansena wykształciło wielu specjalistów w dziedzinie badań reprezentacyjnych na poziomie krajowym i międzynarodowym.

Instytut Statystyczny w Indiach, kierowany przez Mahalanobisa (1944, 1952), stał się drugim centrum badań reprezentacyjnych o znaczeniu międzynarodowym. Indyjskie Narodowe Badania Reprezentacyjne (Indian National Sample Surveys) zdobyły sławę międzynarodową (Suthatme i Sukhatme, 1970).

Do rozpowszechnienia się badań reprezentacyjnych, szczególnie w krajach rozwijających, przyczyniło się FAO (Rzym) i ONZ (Nowy Jork), które podjęły prace już w 1950 r. (O'Muircheartaigh et al., 1981, Zarkovich, 1966). Z ramienia FAO pracowałem jako ekspert i konsultant z badań reprezentacyjnych w Etiopii, Chińskiej Republice Ludowej oraz Nepalu, a w latach 200-2005 byłem konsultantem UN Statistics Division z zakresu badań reprezentacyjnych.

Obecnie większość urzędów statystycznych prowadzi badania reprezentacyjne, chociaż ich rozwój jest dość zróżnicowany z punktu widzenia tempa i jakości.

Dość zróżnicowany rozwój badań reprezentacyjnych był w ośrodkach uniwersyteckich. Warto tu wspomnieć o dwóch ośrodkach: Centrum Badań Reprezentacyjnych (Sample Surveys Center) na uniwersytecie Iowa (Iowa State University), zorganizowane przez Snedecora i Cochrana (1977), oraz Centrum Badawcze na Uniwersytecie Michigan (Survey Research Center) kierowane przez L. Kisha (Kish, 1965, 1996, 1987).

W aspekcie międzynarodowym znaczną rolę odgrywają także sesje naukowe *Międzynarodowego Instytutu Statystycznego (MIS)*, które odbywają się co dwa lata. Na sesjach tych prezentowane są zarówno osiągnięcia teoretyczne jak i praktyczne uzyskane w tej dziedzinie przez ośrodki badawcze różnych krajów, a także przez urzędy statystyczne. Istotną rolę w rozwoju metodologii badań reprezentacyjnych odgrywa także jedna z sekcji MIS: *International Association of Survey Statisticians (IASS),* która jest organizatorem lub współorganizatorem wielu konferencji międzynarodowych poświęconych różnym aspektom badań reprezentacyjnych. W Polsce organizacja ta była współorganizatorem międzynarodowej konferencji poświęconej metodom estymacji dla małych obszarów w 1992 r. (Kalton et al. 1993).

W ostatnim okresie Eurostat podjął intensywne prace nad integracją i harmonizacją badań reprezentacyjnych prowadzonych w statystyce społecznej i ekonomicznej

(Eurostat, 1996, 1997). Poświęca się wiele uwagi ocenie jakości uzyskanych wyników oraz sposobom ich prezentacji (Eurostat, 2007, 2009).

Zostanie dalej przedstawiony, w ogólnym zarysie, rozwój badań reprezentacyjnych w Polsce, rozpoczynając od prac metodologicznych prowadzonych w ośrodkach akademickich i naukowo-badawczych oraz ich wpływ na praktykę badań reprezentacyjnych w GUS. Wspomnę o ważniejszych badaniach reprezentacyjnych prowadzonych przez GUS, rozważę także niektore aspekty badań reprezentacyjnych, takie jak *badania powtarzalne*, *jakość uzyskanych wyników* oraz *metody estymacji dla małych obszarów*.

3. WPŁYW NIEKTÓRYCH STATYSTYKÓW NA ROZWÓJ TEORII I PRAKTYKI BADAŃ REPREZENTACYJNYCH W POLSCE W POCZĄTKOWYM OKRESIE

W Polsce rozwijano zarówno teorię jak i praktykę badań reprezentacyjnych. W początkowym okresie szczególny wpływ na rozwój tych badań miały prace badawcze kilku naukowców z kraju i z zagranicy. Znany jest wpływ Jerzego Neymana, który był kilkakrotnie omawiany przy różnych okazjach (Fisz, 1950a; Kordos, 2009; Zasępa 1958). Znacznie mniej znany jest wpływ profesorów: Marka Fisza i Hugona Steinhausa, którzy oddziaływali w pierwszym okresie na rozwój badań reprezentacyjnych. Istotny wpływ mieli także niektórzy statystycy z innych krajów, z których należy wymienić, moim zdaniem, przede wszystkim Deminga (1950), Daleniusa (1957), Yatesa (1949), Hansena et al. (1953), Cochrana (1977) Kisha (1965) i Zarkowicha (1966).

3.1. WPŁYW PROFESORA MARKA FISZA

Oprócz Jerzego Neymana, znaczny wpływ na rozwój badań reprezentacyjnych w Polsce w początkowym okresie miał także Marek Fisz, który współpracował z GUS oraz brał udział w konsultacjach z J. Neymanem i opisał ich wyniki (Fisz, 1950a). W Komisji do Spraw Statystyki Matematycznej GUS powołanej zarządzeniem wewnętrznym Prezesa GUS, której przewodniczył prof. S. Szulc, Marek Fisz pełnił funkcję sekretarza Komisji[1]. Następnie brał intensywny udział w szkoleniach pracowników GUS z metody reprezentacyjnej, w których były wykorzystane jego programy szkoleniowe, skrypty, a nastepnie jego podręcznik pt. „*Rachunek prawdopodobieństwa i statystyka matematyczna*". Profesor Marek Fisz miał także pewien wpływ na podjęcie w Głównym Urzedzie Statystycznym badań w zakresie jakości danych statystycznych (Fisz, 1950b).

3.2. WPŁYW PROFESORA HUGONA STEINHAUSA

Szczególne znaczenie i wpływ na podjęcie przez GUS badań w zakresie jakości danych oraz badań reprezentacyjnych miał profesor Hugo Steinhaus. Muszę tu odwo-

[1] Zarządzenie wewnętrzne Prezesa GUS Nr 87 z dnia 31 grudnia 1949 r. w sprawie powołania Komisji do Spraw Statystyki Matematycznej.

łać się do własnych doświadczeń w tym zakresie. W latach 1953-1955 studiowałem na Uniwersytecie Wrocławskim, a profesor Steinhaus prowadził dla nas wykłady ze *statystycznej kontroli jakości (SKJ)* oraz seminarium z rachunku prawdopodobieństwa. Uczestniczyłem w tym czasie również w seminariach prowadzonych przez prof. H. Steinhausa i prof. J. Perkala w Instytucie Matematycznym PAN we Wrocławiu nt. *zastosowań przyrodniczych i gospodarczych matematyki*. Pod kierunkiem prof. H. Steinhausa napisałem pracę magisterską pt. *„Przedziały ufności i tolerancji dla rozkładu normalnego",* którą obroniłem w czerwcu 1955 r. Prof. H. Steinhaus miał wpływ na trzy zagadnienia związane z badaniami GUS: a) badanie jakości danych, b) losowanie próbek do badań, oraz c) wycenę statystyczną, które przedstawię w skrócie.

3.2.1. BADANIE JAKOŚCI DANYCH STATYSTYCZNYCH

Jak już wspomniałem, na wykładach ze *statystycznej kontroli jakości,* prof. H. Steinhaus niejednokrotnie podkreślał, że metody te mogą być stosowane także w badaniach statystycznych w celu zbadania ich dokładności, gdyż wyniki tych badań mogą być obciążone różnego rodzaju błędami. Starałem się przenieść pewną wiedzę zdobytą na studiach w mojej pracy w GUS, którą rozpocząłem we wrześniu 1955 r., a współpracowałem dodatkowo również z Komisją Matematyczną GUS, na której przedstawiłem swoje uwagi nt. badania jakości w statystyce w oparciu o wykłady prof. H. Steinhausa ze statystycznej kontroli jakości i mojej pracy z tego zakresu. Dlatego Komisja Matematyczna GUS skierowała mnie dodatkowo do współpracy z wydziałem kontroli badań statystycznych GUS, a dotyczyło to głównie kontroli jakości sprawozdawczości statystycznej. Współpraca ta trwała kilka lat, a następnie zająłem się także jakością innych badań statystycznych prowadzonych przez GUS, które przedstawię w syntetycznym ujęciu w dalszej części artykułu.

3.2.2. LOSOWANIE PRÓBEK – WYKORZYSTANIE TABLIC LICZB LOSOWYCH

W badaniach reprezentacyjnych wystepuje problem uzyskania próbki możliwie najbardziej losowej. Wykorzystuje się tu różne mechanizmy losowania. Pomocnym narzędziem w tym względzie są *tablice liczb losowych,* które prof. Steinhaus także konstruował. Dyskutowaliśmy niejednokrotnie ten problem na seminarium prowadzonym przez Profesora. Starając się zaradzić rozmaitym mankamentom związanym z wyborem próbki, Profesor projektował zbudowanie tablicy zawierającej wszystkie liczby naturalne od 0000 do 9999, każdą dokładnie jeden raz, a powstałej przez wymieszanie tych liczb za pomocą odpowiedniego i wyraźnie określonego algorytmu. Jeden taki dość zawiły algorytm został użyty do zbudowania tablicy liczb „*przetasowanych*". Inny algorytm, korzystający ze złotego podziału odcinka, a dokładniej z reszty modulo 1 wielokrotności liczby złotej $a = (\sqrt{5} - 1)/2 = 0,618....$, został wykorzystany do zbudowania tablic liczb „*złotych"* i „*żelaznych"* (zob. Steinhaus, 1956). Właśnie „*tablice liczb żelaznych*" (TLŻ) zbudowane przez H. Steinhausa zostały wykorzysty-

wane w badaniach reprezentacyjnych prowadzonych przez GUS. Zastosowanie TLŻ w badaniach GUS zaproponował także prof. J. Neyman w czasie swojej konsultacji dla GUS w 1958 r. zamiast losowania systematycznego (zob. Zasępa, 1958). Uważaliśmy, że TLŻ są szczególnie przydatne w losowaniu sekwencyjnym, tj. gdy, z różnych przyczyn, należy dalej kontynuować losowanie próbki. Losowanie sekwencyjne stosowaliśmy, gdy założyliśmy określoną liczebność próby, a na skutek różnych przyczyn nie uzyskania odpowiedzi, trzeba było dalej kontynuować losowanie. Dotyczyło to badań budżetów gospodarstw domowych i innych badań społecznych. Stosowałem także TLŻ w badaniach rolniczych, gdy w latach 1974-1980 pracowałem w Etiopii jako ekspert FAO z zakresu badań reprezentacyjnych.

3.2.3. WYCENA STATYSTYCZNA – DOKŁADNOŚĆ METOD ZBIERANIA DANYCH STATYSTYCZNYCH

Prof. H. Steinhaus w czasie wykładów na temat statystycznej kontroli jakości podkreślał, że metody SKJ mogą być stosowane także do określenia jakości badań masowych, takich jak spisy ludności, sprawozdawczość statystyczna i badania statystyczne, w których metody zbierania danych statystycznych, z różnych powodów, nie zapewniają uzyskanie dokładnych danych. W takim przypadku proponował zastosowanie tzw. *wyceny statystycznej*, która, przeciętnie rzecz biorąc, da dokładniejsze wyniki niż metoda wywiadu indywidualnego. Jako przykład wymieniał uzyskanie informacji o dochodach gospodarstw rolnych lub pracujących na rachunek własny. Respondenci nie są w stanie podać dokładnej informacji, gdyż faktycznie jej nie znają lub z innych przyczyn podają informacje nieprawdziwe. Podają więc informacje obciążone, znacznie odbiegające od stanu faktycznego. Proponował zbudowanie pewnego rodzaju modelu uwzględniającego poszczególne komponenty związane z oszacowaniem dochodów danej jednostki lub zbiorowości. Następnie komponenty te byłyby szacowane na podstawie dostępnych dokumentów lub nawet przez rzeczoznawców. Wiedzieliśmy, że dochody podawane przez respondentów są znacznie zaniżane. Podejście Profesora próbowałem zastosować przy badaniu rozkładów dochodów ludności według wysokości zatrudnionej w gospodarce pozarolniczej (Kordos, 1959, 1963). Budowałem modele wykorzystujące dane z badań budżetów gospodarstw domowych, z badania rozkładów płac według wysokości, które w tym okresie były badaniami pełnymi, z danych z ZUS oraz ze spisów ludności. Podejmowaliśmy w Pracowni Regionalnej Zakładu Badań Statystyczno-Ekonomicznych GUS we Wrocławiu próby zastosowania metody wyceny statystycznej do oceny dochodów w gospodarstwach rolnych, ale prace w tym zakresie nie zostały zakończone. Dopiero po 2000 r. dowiedziałem się, że prof. H. Steinhaus zastosował wycenę statystyczną jako metodę odbioru towarów produkcji masowej, stosując zasadę „minimax" (zob. Steinhaus, 2000). Wydaje mi się, że podejście to można by zastosować w rozwijanym ostatnio globalnym zarządzaniu jakością (TQM), które wkracza również stopniowo do statystyki (Kordos, 2001b).

3.3. WPŁYW INNYCH STATYSTYKÓW

Oprócz wymienionych wyżej polskich statystyków na rozwój badań reprezentacyjnych w Polsce w pierwszym okresie poważny wpływ wywarło kilku statystyków, z których przede wszystkim należy wymienić Deminga (1950), Yatesa (1949), Hansena et al., (1953), Daleniusa (1957), Kisha (1965), Cochrana (1977) i Zarkovicha (1966). Osiągnięcia tych statystyków były intensywnie studiowane i dyskutowane na zebraniach i seminariach organizowanych przez Komisję Matematyczną GUS oraz Pracownię Metod Matematycznych ZBS-E GUS. Szczególną uwagę poświęcaliśmy pracy Deminga (1950, 1987), a głównie jego podejściu do projektowania badań reprezentacyjnych i odpowiedzialności statystyka za projektowanie próbki w całokształcie badania reprezentacyjnego (Kordos, 2009). Zauważyliśmy, że podejście to było zbliżone do podejścia prof. H. Steinhausa, który również kładł nacisk na jakość danych, sposoby wyboru próbki oraz metody zbierania informacji. Intensywnie studiowaliśmy także książkę Daleniusa (1957) dotyczącą teorii i praktyki badań reprezentacyjnych w Szwecji. Chodziło tu głównie o optymalną lokalizację próbki ze względu na kilka cech. Interesujące wyniki w tej dziedzinie uzyskał J. Greń (1964, 1966, 1969, 1970). Należy tu podkreślić, że w sprawie zastosowania optymalnej lokalizacji próbki w praktyce napotkaliśmy w tym okresie na pewną barierę ze strony opracowujących wyniki badań. Twierdzili, że występują trudności techniczne zastosowania proponowanych metod, tłumacząc to „*zmiennym przecinkiem*". Nie mogliśmy więc naszych osiągnięć teoretycznych w tym zakresie zastosować w praktyce. Problem ten został rozwiązany dopiero po wprowadzeniu w szerokim zakresie elektronicznego przetwarzania danych. Ze względów dydaktycznych cieszyły się popularnością podręczniki Cochrana (1977) i Kisha (1965), które były wykorzystywane zarówno w naszych pracach badawczych jak i szkoleniowych. Prof. S. Zarkovich (1966) przyczynił się w znacznym stopniu do pogłębiena badań w zakresie jakości danych (Kordos, 1973, 1987, 1988). W latach 1974-1980 pracowałem w Etiopii jako ekspert FAO, a prof. Zarkovich był moim przełożonym. Kładł szczególny nacisk na metody badania jakości danych statystycznych, a następnie uzyskane doświadczenie w tym zakresie przeniosłem w mojej pracy w GUS.

4. DZIAŁALNOŚĆ KOMISJI MATEMATYCZNEJ GUS

Jak już podkreśliłem, szczególną rolę w rozwoju badań reprezentacyjnych w statystyce publicznej miała Komisja Matematyczna GUS[2], która działała w latach 1950-1993, jako organ doradczy i opiniodawczy Prezesa GUS w zakresie wdrażania i stosowania do praktyki statystycznej metod matematyczno-statystycznych, a w szczególności

[2] Pierwsza Komisja została powołana zarządzeniem wewnętrznym Prezesa GUS Nr 87 z dnia 31 grudnia 1949 r. pod nazwą „*Komisja do Spraw Statystyki Matematycznej"* w składzie: prof. Stefan Szulc – przewodniczący, mgr Marek Fisz – sekretarz, dr Egon Vielrose, mgr Ryszard Zasępa – członkowie. Faktycznie działalność Komisji rozpoczęła się w 1950 r. i stopniowo rozszerzała zakres działania i skład osobowy.

metody reprezentacyjnej. Komisji Matematycznej GUS, w różnych okresach, przewodniczyli znani w Polsce statystycy: prof. Stefan Szulc, prof. Zbigniew Pawłowski, prof. Władysław Welfe i prof. Ryszard Zasępa. Stopniowo Komisja rozszerzała zakres swej działalności i od 1969 r. była Komisją Matematyczną międzyresortową, służącą nie tylko badaniom GUS, ale także innym resortom[3].

Wyniki działalności Komisji Matematycznej GUS były często publikowane w wydawnictwach GUS, a głównie w serii „*Biblioteka Wiadomości Statystycznych*". Publikacje te dotyczyły: a) zastosowania metod matematycznych w statystyce (GUS, 1969); b) metodologii badań reprezentacyjnych (GUS, 1970b, 1971a, 1971b, 1972, 1978, 1987b, 1989); c) wybranych problemów prognoz statystycznych (GUS, 1970a, 1979) d) zastosowania metod ekonometrycznych (GUS, 1976, 1979); e) perspektyw rozwoju statystyki GUS (1973, 1987a). Artykuły indywidualne były zwykle publikowane w „*Wiadomościach Statystycznych*" lub „*Przeglądzie Statystycznym*". W latach 1966-1969, jako sekretarz naukowy Komisji, publikowałem na łamach „*Wiadomości Statystycznych*" informacje pt. *„Z prac Komisji Matematycznej GUS*"[4]. W „*Przeglądzie Statystycznym"* omówiłem działalność Komisji za okres 25-lecia (Kordos, 1975). Znaczny udział w pracach Komisji Matematycznej GUS w latach 1972-1993 oraz w rozwoju badań reprezentacyjnych w Polsce, miał mgr Bronisław Lednicki, który w latach 1990-1993 pełnił funkcje sekretarza naukowego Komisji. Był autorem wielu ważnych badań reprezentacyjnych prowadzonych przez GUS (Lednicki, 1973, 1974, 1979, 1982, 1987, 1989). Najdłużej funkcję przewodniczącego Komisji pełnił prof. Ryszard Zasępa, który wniósł największy wkład w rozwój badań reprezentacyjnych w Polsce, a także organizował intensywne szkolenia pracowników GUS z zakresu badań reprezentacyjnych.

W Komisji Matematycznej GUS pełniłem różne funkcje od członka Komisji, sekretarza naukowego, zastępcy przewodniczącego do przewodniczącego Komisji w latach 1990-1993[5]. Szerzej działalność Komisji Matematycznej GUS w latach 1950-1993 przedstawawiłem w opublikowanym ostatnio artykule na łamach „*Wiadomości Statystycznych*" (Kordos, 2012 a).

5. ZARYS ROZWÓJU BADAŃ REPREZENTACYJNYCH W POLSCE

Pomimo podjęcia dość wcześnie prac w zakresie zastosowania metody reprezentacyjnej w Polsce, bo już w pierwszej połowie lat trzydziestych ubiegłego wieku, to jednak początkowo wykorzystanie tej metody badań w naszej praktyce statystycznej

[3] Zarządzenie, nr 38 Prezesa Głównego Urzędu Statystycznego z dnia 10 czerwca 1969 r. (znak: I-2-0200-38) w sprawie powołania Komisji Matematycznej przy Głównym Urzędzie Statystycznym.

[4] Patrz: J. Kordos, Z prac Komisji Matematycznej GUS, *Wiadomości Statystyczne*, z. 9, 1966, s. 20-22; z. 10, s. 22-24.; z. 12, s. 14-16; 1967, z. 3, s. 25-27; z. 6, s. 18-19; z. 9, s. 12-13; 1969, z. 5, s. 21-22.

[5] Zarządzenie Nr 31 Prezesa Głównego Urzędu Statystycznego a dnia 4 lipca 1990 r. w sprawie powołania Komisji Matematycznej *„Dziennik Urzędowy Głownego Urzędu Statystycznego",* Nr 12 (172), s. 333-334.

było dość ograniczone. Przed II wojną światową zastosowanie metody reprezentacyjnej ograniczyło się do przyśpieszenia opracowania wyników spisu ludności z 1931 r. (Neyman, 1933; Piekałkiewicz,1934). Zastosowania metody reprezentacyjnej w badaniach GUS po ostatniej wojnie były konsultowane, jak już wspomniałem, z prof. J. Neymanem w czasie jego wizyt w Polsce w latach 1950 i 1958. Brałem udział w konsultacjach w 1958 r. w ramach Komisji Matematycznej GUS, które trwały przez 6 tygodni. Było to dla mnie wielkim przeżyciem naukowym, a także jako najmłodszy z zespołu Komisji prowadziłem protokóły i dokładne notatki oraz miałem bezpośrednie kontakty z Profesorem w przygotowaniu i ocenie spotkań. Z prof. R. Zasępą podziwialiśmy nie tylko wielką wiedzę prof. Neymana i jego oryginalne podejście do rozwiązywania problemów statystycznych, ale także ogromną pracowitość. Niejednokrotnie kończyliśmy pracę w późnych godzinnych wieczornych, gdy była taka potrzeba. Swoje refleksje z tych spotkań przedstawiłem ostatnio w artykule w języku angielskim (Kordos, 2011).

5.1. BADANIA REPREZENTACYJNE DO 1989 ROKU [6]

W poprzednim systemie społeczno-gospodarczym, badania statystyczne były głównie oparte o pełną sprawozdawczość statystyczną, a w ograniczonym stopniu wykorzystywano badania częściowe. Metodę reprezentacyjną stosowano głównie w badaniach społecznych (badania budżetów gospodarstw domowych i badania warunków życia), a także w badaniach rolniczych. Badania te zostały opisane w kilku publikacjach Głównego Urzędu Statystycznego (GUS, 1978, 1987a, 1987b, 1989). Polscy statystycy rozwijali metodologię badań reprezentacyjnych w ośrodkach akademickich oraz jednostkach naukowo-badawczych GUS. Powstało wiele intersujących opracowań naukowych, które w zmacznym stopniu zostały wykorzystane w praktyce. Prace te wymagają specjalnego opracowania monograficznego, a tu wymienię tylko niektóre z nich.

Metoda reprezentacyjna była w Polsce twórczo rozwijana, wykładana na uczelniach o profilu ekonomicznym, a w 1962 r. został opublikowany w Polsce pierwszy podręcznik metody reprezentacyjnej (Zasępa, 1962), a później dalsze książki poświęcone tej metodzie badawczej (Pawłowski, 1972; Zasępa, 1972; Steczkowski, 1988; Bracha, 1996, Szreder, 2004; Wywiał, 2010). W 1987 r. została opublikowana moja monografia nt. *dokładności danych w badaniach społecznych* (Kordos, 1987), a w 1988 r. nt. *jakości danych statystycznych* (Kordos, 1988). Z inicjatywy Komisji Matematycznej GUS, pod kierunkiem jej przewodniczącego, prof. R. Zasępy, prowadzono dla pracownikow GUS intensywne szkolenia z metody reprezentacyjnej, co przyczyniło się do lepszego zrozumienia tej metody badania statystycznego w praktyce. Powstanie w 1966 r. Zakładu Badań Statystyczno-Ekonomicznych GUS (ZBS-E GUS) umożliwiło podjęcie prac naukowo-badawczych, wdrożenie a także doskonalenie tej metody badawczej w praktyce. Próbowaliśmy także przekazywać nasze doświadczenia w tym zakresie innym krajom, organizując w 1970 r. konferencję międzynarodową (GUS, 1971a i 1971b).

[6] Podano także publikacje wydane po 1989 roku.

Z publikacji powstałych w ośrodkach akademickich oraz naukowo-badawczych, które miały istotny wpływ na praktykę statystyczną, należałoby tu wymienić prace R. Zasępy (1958, 1962, 1968, 1972, 1993ab), Cz. Brachy (1972, 1994, 1996, 1998, 2003), Cz. Brachy et al. (2004a, 2004b), J. Grenia (1964, 1966, 1969, 1970), B. Lednickiego (1973, 1974, 1979, 1982, 1987, 1989), B. Lednickiego et al., (1994, 2003,), J. Kordosa (1967, 1968, 1971, 1974, 1982, 1985, 1991), L. Kursy, B. Lednickiego (2006), J. Pawłowskiej (1969), A. Szarkowskiego, J. Witkowskiego (1994) oraz J. Wywiała (1988, 1995, 1996, 1998).

Polscy specjaliści z badań reprezentacyjnych byli współorganizatorami i brali aktywny udział w trzech międzynarodowych konferencjach, które odbyły się w Polsce na początku okresu przejściowego do gospodarki rynkowej:

1) w 1991 r. nt. *pomiaru ubóstwa w okresie przemian gospodarczych w krajach Europy Wschodniej* (*poverty measurement for economies in transition in Eastern European Countrie*s), która odbyła się w Jachrance k. Warszawy w dn. 7-9 października 1991 r. (GUS, 1992);

2) w 1992 r. nt. *problemów statystyki małych obszarów i planowania badań* (*small area statistics and survey design),* która odbyła się w Warszawie w dn. 30 września – 3 października 1992 r. (Kalton et al., 1993);

3) w 1994 r. *międzynarodowa konferencja statystyczna dla upamiętnienia setnej rocznicy urodzin Jerzego Neymana* (*International Conference in Memory of the Hundredth Anniversary of the Birth of Jerzy Neyman)*, która odbyła się w dn. 25-26 listopada 1994 r. w Jachrance k. Warszawy.

W czasie tych międzynarodowych konferencji oraz spotkań z ekspertami Eurostatu i innych organizacji miedzynarodowcy okazało się, że polscy statystycy specjalizujący się w metodyce badań reprezentacyjnych byli dość dobrze przygotowani do przekształcenia polskiej statystyki do wymagań gospodarki rynkowej.

5.2. BADANIA REPREZENTACYJNE PO 1989 ROKU

Od 1989 r. rozpoczęto w GUS dostosowywanie statystyki publicznej do międzynarodowych standardów, a także podjęto starania zharmonizowania jej ze statystyką Unii Europejskiej (Kordos, 1998). Zostały zaadaptowane podstawowe klasyfikacje i nomenklatury, program badań został w znacznym stopniu zmieniony, zastosowano nowe metody i metodologie badań dostosowane do wymagań gospodarki rynkowej i standardów europejskich GUS (1995).

Przerwano większość sprawozdawczości statystycznej, a rozpoczęto nowe badania reprezentacyjne. Dostosowano *badania budżetów gospodarstwach domowych* do wymagań Eurostatu (Eurostat, 1997; GUS, 2011a) oraz podjęto nowe badania, takie jak: *badania aktywności ekonomicznej ludności* (Popiński, 2006; Szarkowski,Witkowski, 1994)*, badania przedsiębiorstw, koniunktury w przemyśle, budownictwie i handlu* (Barczyk et al., 1994)*, badania* w *rolnictwie* (Kordos, Kursa, 1997; Kursa, Lednicki, 2006), *badania zdrowia* (Bracha, 1998; GUS, 2006, 2011b), *badania warunków życia* (GUS,

2002, 2005a) i *badania wykorzystania czasu* (GUS, 2005b). Od 2005 r. wprowadzono Europejskie Badanie Dochodów i Warunków Życia (EU-SILC) przygotowane według koncepcji Eurostatu (GUS, 2008; Verma & Betti, 2006). Studiowano i wdrożono do praktyki metody estymacji precyzji dla złożonych schematów losowania (Kordos, Szarkowski, 1973; Popiński, 2006; Szarkowski, Witkowski, 1994; Kordos, Zięba-Pietrzak, 2010). Należy tu także wymienić prace z zakresu lokalizacji próby i statystyki małych obszarów, które wskazały kierunki doskonalenia praktyki statystycznej (np. Kozak, 2004, 2006; Kubacki, 2004, 2006; Niemiro et al., 2001, 2002, 2012).

5.3. REPREZENTACYJNE BADANIA POWTARZALNE W CZASIE – METODA ROTACYJNA

Większość badań statystyki publiczne są badaniami powtarzalnym w różnych okresach. Próbuje się tu stosować badania panelowe, w których wybrana jednostka badania bierze udział w badaniu w kilku kolejnych rundach, gdy głównym celem badania jest ocena zmian w kolejnych okresach. Jednakże w większości badań chodzi zarówno o badania zmian w czasie jak również o oceny w poszczególnych okresach. Kompromisem jest tu *metoda rotacyjna,* której na świecie, a także w Polsce, poświęcono wiele uwagi. Prace teoretyczne rozpoczęto w tej dziedzinie już na początku lat pięćdziesiątych ubiegłego wieku (Patterson, 1950). Z późniejszych prac warto tu wymienić prace Ecklera (1955), Rao i Grahama (1964), Kaltona i Kasprzaka (1987), Kaltona i Citro (1993).

W Polsce problematyce badań powtarzalnych poświęcono wiele uwagi, a szczególnie metodzie rotacyjnej, którą próbowano zastosować w kilku badaniach, a głównie w badaniach budżetów gospodarstw domowych (GUS, 1972, 1986, 1989; Kordos, 1967, 1971, 1982; Lednicki, 1982), w badaniu aktywności ekonomicznej ludności (Szarkowski i Witkowski, 1994; Popiński, 2006) oraz warunków życia GUS (2008). Należy także wspomnieć o pracach teoretycznych w zakresie metody rotacyjnej podjętych w ostatnich latach przez kilku statystyków (np. Ciepiela et al., 2012; Kowalczyk, 2003; Kowalski, 2006, 2009; Kowalski, Wesołowski, 2011; Wesołowski, 2010). Szerszego przeglądu prac teoretycznych w zakresie badań reprezentacyjnych w czasie, a szczgólnie metody rotacyjnej, prowadzonych w Polsce w ostatnich latach, a także ich praktycznych zastosowań, przedstawiłem w opracowaniu monograficznym (Kordos, 2012b).

5.4. PROBLEMY JAKOŚCI DANYCH W BADANIACH REPREZENTACYJNYCH

Prace nad jakością danych statystycznych podjęto w GUS dość wcześnie, bo już w końcu lat 1950-tych. Znaczny wpływ, jak już wspomniałem, miał tu prof. Hugo Steinhaus, którego propozycje w tym zakresie próbowano zastosować w statystyce publicznej. Pierwsze wyniki prac z zakresu analizy błędów losowych i nielosowych w Polsce przedstawiłem na sesji naukowej Międzynarodowego Instytutu Statystycznego w Wiedniu w 1973 r. (Kordos, 1973). Problematykę jakości danych statystycznych

przedstawiłem poźniej w dwóch monografiach (Kordos, 1987, 1988) i w kilku artykułach (Kordos, 1995, 2001a, 2007).

W ostatnich latach na świecie wiele uwagi poświęcano jakości danych statystycznych, a w ostatnim okresie – *jakości statystyki.* Wiąże się to ściśle ze stale rosnącym zapotrzebowaniem na rzetelne informacje statystyczne, które wykorzystywane są w szerokim zakresie w różnych dziedzinach życia gospodarczego i społecznego, a także do porównań międzynarodowych.

Szersze zainteresowanie jakością statystyki, a nie tylko danymi statystycznymi wynika także z rewolucji jakościowej w społeczeństwie. W przedsiębiorstwach przemysłowych, a także w innych jednostkach organizacyjnych wprowadzane jest stopniowo *globalne zarządzanie jakością*[7] *(Total Quality Management – TQM)*, które obejmuje wszystkie procesy i jednostki organizacyjne oddziaływujące na jakość produkcji lub usług. Organizacje statystyczne nie stanowią tu wyjątku. Przecież produkty statystyczne – dane statystyczne, ich interpretacje i analizy, wykorzystywane wykresy i diagramy, publikacje statystyczne oraz udzielane konsultacje w zakresie statystyki, podlegają podobnym zasadom i regułom procesów masowych jak produkty przemysłowe czy usługi.

Problematyką jakości danych zajmowało się kilku naszych statystyków (Kordos, 1973, 1987, 1988; 2001a, 2004, 2007; Szutkowska, 2012; Wierzchosławski, 1995; Zasępa, 1962, 1968, 1972, 1993ab).

W ostatnich latach wiele uwagi poświęca się problematyce jakości w statystyce. Ograniczę się tu tylko do wspomnienia ostatnich prac prowadzonych pod patronatem Eurostatu. Prace te dotyczą udoskonalenia jakości Europejskiego Systemu Statystycznego. Ważne jest, że pojęcie jakości w statystycznych organizacjach zmieniło się w czasie ostatniej dekady. Tak, więc dokładność nie jest już jedyną miarą jakości. W takim podejściu jakość może być zdefiniowana w wielu wymiarach, gdzie dokładność jest jednym z nich Eurostat (2007; 2009). GUS próbuje stopniowo uwzględnić te zalecenia (Szutkowska, 2012). W publikacjach GUS, w których prezentowane są wyniki z badań reprezentacyjnych, podawane są zwykl błędy standardowe lub względne błędy standardowe dla wybranych parametrów. Z błędów nielosowych podawane są niekiedy frakcje realizacji badania, tj. iloraz liczby zbadanych jednostek do liczby jednostek wylosowanych, lub frakcje braku odpowiedzi. Błędy odpowiedzi są rzadko badane, a próbę oszacowania błędów tego rodzaju podejmuje się tylko w kontrolnych badaniach spisów ludności. Jednakże oceny jakości ostatnich spisów ludności, w których zastosowano pospisowe badania kontrolne, tj. w latach 1978, 1988 i 2002, nie zostały opublikowane. Należy mieć nadzieję, że analiza jakości spisów z tych badań zostanie wkrótce opublikowana, łącznie z oceną jakości Narodowego Spisu Ludności 2011. W sprawie oceny jakości spisów ludności została ostatnio opublikowana specjalna monografia przez UN Statistics Division (2010).

[7] Warto zauważyć, że prof. W.E. Deming uważany jest za ojca TQM, który miał znaczny wpływ na rozwój badań reprezentacyjnych w Polsce w początkowym okresie. Omówiłem ten fakt w moim artykule Kordos (2009).

5.5. ESTYMACJA DLA MAŁYCH OBSZARÓW A BADANIA REPREZENTACYJNE

Problem ocen wartości parametrów dla małych obszarów występował w statystyce już od początku prowadzenia badań częściowych. Dla badań częściowych, prowadzonych metodą reprezentacyjną, w których jednostki próby losowano według określonego schematu, powstaje problem rzetelności ocen, szczególnie, gdy są one oparte na niewielkiej liczebności badanych jednostek. Niekiedy precyzja takich ocen jest bardzo niska i ich wartość informacyjna znikoma. Dotyczy to zwykle małych obszarów, takich jak podregiony, powiaty, miasta, a nawet niektórych województw, gdy próbka, na podstawie, której szacowane są wyniki, jest zbyt mała.

Wprawdzie metodami ocen dla małych obszarów zajmowano się już od dość dawna, to jednak szersze zainteresowania tymi problemami wśród statystyków obserwuje się w ciągu ostatnich 40 lat. Zaczęły ukazywać się prace metodyczne i naukowe w tym zakresie, a także rozpoczęto organizowanie międzynarodowych konferencji naukowych i seminariów, na których rozważano różne problemy związane z estymacją parametrów dla małych obszarów. Warto tu wymienić międzynarodowe sympozjum zorganizowane w 1985 r. w Ottawie (Platek, et. al., 1987), międzynarodowe konferencje naukowe, które odbyły się w 1992 r. w Warszawie (Kalton et al., 1993) i w Rydze (1999). Od 2005 roku organizowane są co dwa lata międzynarodowe konferencje poświęcone metodom estymacji dla małych obszarów, a dotychczas odbyły się konferencje: w Filandii (2005), we Włoszech (Piza, 2007), w Hiszpanii (Elche, 2009) oraz w Niemczech (Trier, 2011). Wiele artykułów prezentowanych na tych konferencjach opublikowano w naszym międzynarodowym czasopiśmie w języku angielskim *Statistics in Transition*. W latach 2002-2005 polscy statystycy, pod kierunkiem prof. J Paradysza i prof. E. Gołaty, uczestniczyli w międzynarodowym konsorcjum Eurarea (2005), sponsorowanym przez Eurostat. Wymienię tu tylko niektóre publikacje z tego zakresu wydanych w Polsce: Dehnel, (2010); Domański, Pruska (2001); Gołata (2004a, 2004b, 2012); Kalton, et al. (1993); Kordos (1991, 1999, 2004); Kordos, Paradysz (2000); Paradysz (1998); Pekasiewicz, Pruska (2002); Wesołowski (2004). Jednakże, do wdrożenia tej metody badawczej do praktyki statystycznej na szerszą skalę niezbędne są dalsze prace badawcze (Gołata, 2012).

5.6. MIĘDZYNARODOWE KONFERENCJE W KATOWICACH POŚWIĘCONE METODZIE REPREZENTACYJNEJ

Wiele interesujących opracowań teoretycznych i metodologicznych z metody reprezentacyjnej prezentowali polscy statystycy na różnych konferencjach międzynarodowych. Szczególne znaczenie dla rozwoju badań reprezentacyjnych w Polsce miały międzynarodowe konferencje organizowane przez Katedrę Statystyki Uniwersytetu Ekonomicznego w Katowicach, pod kierunkiem prof. Janusza Wywiała i jego współpracowników pt: *Metoda reprezentacyjna w badaniach ekonomiczno-społecznych* (*Survey Sampling in Economic and Social Research*).

Wykładowcami na tych konferencjach byli często światowej sławy eksperci, jak np. prof. Malay Ghosh z USA, prof. Jean-Claude Deville z Francji i prof. Yves Tillé ze Szwajcarii. Dotychczas zorganizowano 7 konferencji międzynarodowych, a wyniki konferencji z lat 2008 i 2009 zostały wydane w publikacjach zwartych (patrz Wywiał, Żądło, 2009; Wywiał, Gamrot, 2010). Ostatnia konferencja odbyła się w dn. 18-20 września 2011 r. Należy mieć nadzieję, że konferencje te będą w przyszłości kontynuowane.

5.7. REPREZENTACYJNY SPIS LUDNOŚCI 2011

Największym badaniem reprezentacyjnym w dziejach GUS był przeprowadzony w ramach Narodowego Spisu Powszechnego 2011 *reprezentacyjny spis ludności* (GUS, 2012). Spis reprezentacyjny został przeprowadzony na próbie losowej w ok. 20% mieszkań w skali kraju. Jednostką losowania było mieszkanie, a dokładniej jego adres. Operat losowania, który został utworzony w oparciu o rejestry i systemy informacyjne, został odpowiednio uporządkowany i podzielony na poszczególne warstwy.

Podstawowym celem spisu reprezentacyjnego realizowanego w ramach NSP 2011 było uzyskanie informacji o sytuacji społeczno-demograficznej na poziomie powiatu. Konieczne było więc dokonanie podziału założonej 20%-owej próby mieszkań dla Polski pomiędzy powiaty. Dokonano tego przy wykorzystaniu metody alokacji pierwiastkowej. Metoda ta stanowi kompromis pomiędzy alokacją proporcjonalną, a alokacją zapewniającą jednakową precyzję dla subpopulacji. Przy założeniu, że zastosowano by proporcjonalne losowanie próby, w każdym powiecie próba stanowiłaby 20% populacji. Ponieważ precyzja wyników tj. wielkość błędu losowego, zależy od liczebności próby, uzyskano by w efekcie niedostateczną precyzję dla wielu małych powiatów. Z kolei, w metodzie alternatywnej uzyskujemy, w przybliżeniu, jednakową precyzję wyników dla wszystkich powiatów, ale za cenę istotnego „spłaszczenia" liczebności próby. W efekcie liczebność próby, a tym samym pracochłonność przy realizacji spisu byłaby mało zróżnicowana pomiędzy dużymi i małymi powiatami. Z tych powodów jako metodę rozdziału próby przyjęto alokację pierwiastkową, w której np. liczba mieszkań losowanych w p-tym powiecie jest proporcjonalna do pierwiastka kwadratowego z populacyjnej liczby mieszkań. W celu wylosowania, w każdym z powiatów, próby o ustalonej wcześniej liczebności zastosowany został schemat losowania jednostopniowego warstwowego. Jednostki losowania – mieszkania, zostały przed losowaniem pogrupowane w warstwy w celu zwiększenia efektywności losowania. Zastosowano zróżnicowane podejście do warstwowania w zależności od typu powiatu i gminy. Schemat losowaniu próby oraz ogólny zarys metody estymacji wyników zostały przedstawione w publikacji GUS (2012).

Spis kontrolny do NSP 2011 przeprowadzono od 1 do 11 lipca 2011 r. Jego celem było sprawdzenie kompletności przeprowadzonego spisu, poprawności danych uzyskanych w spisie oraz ich zgodności ze stanem faktycznym. Spis kontrolny przeprowadzono w formie spisu reprezentacyjnego. Spośród 2744 tys. mieszkań, które wcześniej zostały wylosowane do spisu reprezentacyjnego, wylosowano 80 tys. mieszkań. Spisem

kontrolnym objęto mieszkania, w których respondenci dokonali samospisu przez Internet, mieszkania spisane bezpośrednio przez rachmistrzów spisowych lub telefonicznie przez ankieterów jak i takie, w których spis nie został przeprowadzony (z różnych powodów). Ze względu, jak twierdzą autorzy, na zapisy ustawowe spis kontrolny został przeprowadzony przez ankieterów poprzez telefon (metodą CATI). Spowodowało to objęcie spisem kontrolnym wyłącznie tych mieszkań, w których przynajmniej dla jednej osoby udało się ustalić numer telefonu, przy czym nie miało znaczenia, czy był to telefon stacjonarny, czy komórkowy. Formularz do spisu kontrolnego zawierał 14 pytań. Zwykle w badaniu kontrolnym stosuje się dokładniejsza metodę zbierania danych niż w spisie podstawowym. Wydaje się jednak wątpliwe, aby metoda wywiadu telefonicznego umożliwiła sprawdzenie dokładności udzielanych odpowiedzi w badaniu podstawowym (UN Statistics Division, 2010). Jednakże analizę krytyczną uzyskanych wyników należy przeprowadzić i wyciągnąć stąd odpowiednie wnioski na przyszłość.

6. UWAGI KOŃCOWE

Przedstawiłem tu w bardzo ogólnym zarysie, współzależność pomiędzy rozwojem teorii i praktyki badań reprezentacyjnych w Polsce. Zdaje sobie sprawę z subiektywizmu mojej oceny, gdyż w badaniach tych uczestniczyłem jako praktyk, zarówno w kraju jak i za granicą, ponad 50 lat. Uważam, że w okresie tym polscy statystycy, zarówno akademiccy jak i oficjalni, wnieśli istotny wkład w światowy rozwój badań reprezentacyjnych, pomimo znacznych trudności występujących w poszczególnych okresach. Brali aktywny udział w międzynarodowym życiu statystycznym, prezentując swój dorobek na międzynarodowych konferencjach, seminariach lub biorąc udział w konsorcjach, a także w różnych projektach międzynarodowych.

Wydaje się, że największym wyzwaniem jest stale rosnące zapotrzebowanie na różnego rodzaju rzetelne dane statystyczne w różnorodnych układach, porównywalne w czasie i przestrzeni. Powstaje poważny problem w jaki sposób można zaspokoić te potrzeby biorąc pod uwagę rozwój teorii badań statystycznych, dotychczasową praktykę statystyczną i istniejące różnego rodzaju rejestry administracyjne i inną dostępną dokumentację. Chodzi tu nie tylko o porównywalność między poszczególnymi krajami i regionami w układach europejskich zalecanych przez Eurostat, OECD i ONZ, ale także w układach dla poszczególnych krajów. Danych tych przecież nie można uzyskać z badań pełnych oraz istniejących rejestrów administracyjnych, więc niezbędne staje się szersze wykorzystanie badań reprezentacyjnych.

Badania te, jak wykazano powyżej, odegrały dotychczas ogromną rolę, jednakże wystąpiło wiele zaniedbań i uchybień, które powinny być stopniowo usuwane. Chodzi tu przede wszystkim o jakość tych badań, przystosowanie ich do analiz przekrojowych i longitudinalnych, a także o zwiększenie ich precyzji dla małych obszarów. Te zagadnienia nie zostały dotychczas dostatecznie rozwiązane. We wszystkich krajach obserwuje się ograniczenia środków finansowych na badania statystyczne i wydaje się, że tendencja ta utrzyma się dalej.

Poważnym dylematem jest w jaki sposób zaspokoić wymagania użytkownika. Przecież perspektywa *użytkownika* jest zupełnie inna niż statystyka. Użytkownik interesuje się udostępnionymi wynikami z badania, jeśli one spełniają jego osobiste pojęcia o jakości i użyteczności. Będzie on stawiał następujące pytania: (a) czy wyniki są porównywalne lub zgodne z informacjami z innych podobnych źródeł danych? (b) czy informacja jest odpowiednia i aktualna? (c) czy jest porównywalna z poprzednimi obserwacjami? (d) czy mogę zaufać procedurze pomiarowej, korektom ze względu na brak odpowiedzi i procedurze estymacji wyników? (e) czy dostępne dane są przygotowane w sposób łatwo zrozumiały, solidny i użyteczny dla moich celów?

Użytkownicy chcą *zaufania* do jakości danych statystycznych im dostarczanych. Zaufanie może być rozszerzone przez dostarczenie użytkownikom obiektywnych mierników dotyczących ważnych aspektów planu i procesu badania. Wykorzystuje się wiele mierników. Niektóre urzędy statystyczne podają w pewnych badaniach określone mierniki jakości danych. Praktyka w tym zakresie, tj. informowania użytkownika o jakości danych, jest różnorodna. Powstaje więc pytanie, czy ujednolicić informacje o badaniu dla potrzeb użytkownika? Wiadomo, że za wiele numerycznych szczegółów może być dla użytkownika raczej kłopotliwe niż objaśniające. Jaką praktykę przyjąć w tym zakresie? Na te pytania trzeba odpowiedzieć, ale nie ma gotowych recept. Należy podkreślić, że metodologia badań statystycznych, w tym badań reprezentacyjnych, jest podstawową częścią ogólnej efektywności urzędów statystycznych (Eurostat, 2007, 2009; Fellegi, 1999). Reputacja urzędu statystycznego zależy w poważnym stopniu od solidności metodologii statystycznej. Słyszy się niekiedy opinie, że w czasach ograniczeń budżetowych nie powinno się rozwijać metodologii badań, gdyż wiele z nowych badań nie będzie prowadzonych. Niejednokrotnie można się było przekonać, że ulepszenie metody badania może być ważnym czynnikiem ogólnej efektywności urzędu. Może to dotyczyć lepszego projektowania badania oraz udoskonalenia narzędzi pomiaru i przetwarzania. Podczas gdy metodologię można zniszczyć w przeciągu kilku miesięcy, to odbudowanie właściwej zdolności metodologicznej urzędu zabierze lata a nawet dekady (Fellegi, 1999). Główny Urząd Statystyczny powinien odpowiedzieć na te pytania. Podjęte prace w zakresie globalnego zarządzania jakością statystyki, zgodnie z koncepcją Eurostatu, powinny być transparentne dla wszystkich statystyków i użytkowników danych. Wiadomo, że rok 2013 będzie obchodzony jako *Międzynarodowy Rok Statystyki*, okazja ta powinna być wykorzystana do wzmocnienia naszej statystyki publicznej, zawodu statystyka, a szczególnie zwiększenia publicznej świadomości znaczenia i wpływu statystyki na wszystkie aspekty życia społecznego. Jakość badań reprezentacyjnych powinna być udoskonalana w całokształcie systemu statystyki publicznej.

Szkoła Główna Handlowa w Warszawie
Wyższa Szkoła Menedżerska w Warszawie

LITERATURA

[1] Barczyk A., Gaca I., Zagożdzińska I., (1994), Badanie koniunktury, *Zeszyty metodyczne i klasyfikacje GUS*, Warszawa.

[2] Bowley A.L., (1913), Working Class Households in Reading, *Journal of the Royal Statistical Society*, 672-691.
[3] Bowley A.L., (1915), *The Nature and Purpose of the Measurement of Social Phenomena,* P.S. King and Son, Ltd, London.
[4] Bowley A.L., (1926), Measurement of the Precision Attained in Sampling, *Bulletin of the International Statistical Institute*, 22 (1), 1-62.
[5] Bracha Cz., (1972), Reprezentacyjne badanie struktury zapasów w uspołecznionym handlu detalicznym, *Wiadomości Statystyczne*, 1, 21-25.
[6] Bracha Cz., (1994), Metodologiczne aspekty badania małych obszarów, *Z prac Zakładu Badań Statystyczno-Ekonomicznych GUS,* 43.
[7] Bracha Cz., (1996), *Teoretyczne podstawy metody reprezentacyjnej,* PWN, Warszawa.
[8] Bracha Cz., (1998), Schemat losowania próby i metoda estymacji w badaniu ankietowym „stan zdrowia ludności", *Wiadomości Statystyczne*, 3, 9-15.
[9] Bracha Cz., (2003), *Estymacja danych z Badania Aktywności Ekonomicznej Ludności na poziomie powiatów dla lat 1995-2002*, GUS, Warszawa.
[10] Bracha Cz., Jakubowski J., Szarkowski A., (2004a), Analiza porównawcza estymatorów regresyjnych w reprezentacyjnych badaniach statystycznych, *Seria Studia i Prace – Z prac Zakładu Badań Statystyczno-Ekonomicznych GUS i PAN*, 292, Warszawa.
[11] Bracha Cz., Lednicki B., Wieczorkowski R., (2004b), Wykorzystanie złożonych metod estymacji do dezagregacji danych z Badania Aktywności Ekonomicznej Ludności w roku 2003, *Studia i Prace – Z prac Zakładu Badań Statystyczno-Ekonomicznych GUS i PAN*, 299, Warszawa.
[12] Ciepiela P., Gniado K., Wesołowski J., Wojty M., (2012), Dynamic K-Composite estimator for an arbitrary rotation scheme, *Statistics in Transition-new series*, 13 (1), 7-20.
[13] Cochran W.G., (1977), *Sampling Techniques*, 3rd ed., Wiley, New York.
[14] Dalenius T., (1957), *Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice.* Uppsala.
[15] Dehnel G., (2010), *Rozwój mikroprzedsiębiorczości w Polsce w świetle estymacji dla małych domen*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.
[16] Deming W.E., (1950), *Some Theory of Sampling*, Wiley, New York.
[17] Deming W.E., (1987), On the Statistician's Contribution to Quality, *Bulletin of the International Statistical Institute, Proceedings of the 46th Session,* 2, 355-369.
[18] Domański Cz., Pruska, K., (2001), *Metody statystyki małych obszarów*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
[19] Eckler A.R., (1955), Rotation Sampling, *Annals of Mathemathical Statistics*, 26, 664-685.
[20] Eurostat (1996), *The Future of European Social Statistics: Use of Administrative Registers and Dissemination Strategies*, Luxembourg.
[21] Eurostat (1997), *Family Budget Surveys in the EC: Methodology and Recommendations for Harmonisation*, Population and Social Conditions 3, Methods E, Luxembourg.
[22] Eurostat (2007), *Handbook on Data Quality Assessment: Methods and Tools*, Luxembourg.
[23] Eurostat (2009*), Handbook on Quality Report,* Luxembourg.
[24] Fellegi I., (1999), Statistical Services – Preparing for the Future, *Survey Methodology,* 25 (2), 113-128.
[25] Fisz M., (1950a), Konsultacje prof. Neymana i wnioski z nich wypływające, *Studia i Prace Statystyczne*, 3-4, 14-27.
[26] Fisz M., (1950b), Kontrola jakości produkcji masowej na cechę ciągłą, *Studia i Prace Statystyczne*, 2, 123-160.
[27] Frankel L.R., Stock J.S., (1942), On the sample surveys of employment, *Journal of the Statistical American Association,* 37, 77-80.
[28] Gołata E., (2004a), *Pośrednia estymacja bezrobocia na lokalnym rynku pracy,* Wydawnictwo Poznańskiej Akademii Ekonomicznej, Prace habilitacyjne, 11, Poznań.

[29] Gołata E., (2004b), Problems of Estimate Unemployment for Small Domains in Poland, *Statistics in Transition*, 6 (5), 755-776.
[30] Gołata E., (2012), Data integration and small domain estimation in Poland – experiemnces and problems, *Statistics in Transition-new series*, 13 (1), 107-142.
[31] Greń J., (1964), O pewnych metodach wyznaczania lokalizacji próby w losowaniu warstwowym wieloparametrycznym, *Przegląd Statystyczny*, 3, 361-369.
[32] Greń J., (1966), O pewnym zastosowaniu programowania nieliniowego do metody reprezentacyjnej, *Przegląd Statystyczny*, 3, 361-369.
[33] Greń J., (1969), Efektywność powtarzalnych badań metodą reprezentacyjną, *Biblioteka Wiadomości Statystycznych,* 7, 97-104.
[34] Greń J., (1970), Wielowymiarowy estymator regresyjny średniej dla skończonej populacji, *Przegląd Statystyczny*, 1, 73-78.
[35] GUS (1969), Zastosowanie metod matematycznych w statystyce, *Biblioteka Wiadomości Statystycznych,* 7, Warszawa.
[36] GUS (1970a), Wybrane problemy prognoz statystycznych, *Biblioteka Wiadomości Statystycznych,* 11, Warszawa.
[37] GUS (1970b), Statystyczna ocena wyników badań budżetów rodzinnych, *Studia i Prace Statystyczne*, Warszawa.
[38] GUS (1971a), Badania statystyczne metodą reprezentacyjną w krajach socjalistycznych, *Biblioteka Wiadomości Statystycznych*, 14, Warszawa.
[39] GUS (1971b), Wybrane problemy metodologiczne badań reprezentacyjnych, *Biblioteka Wiadomości Statystycznych*, 15, Warszawa.
[40] GUS (1972), Eksperytentalne badania budżetów rodzinnych metodą rotacyjną, *Biblioteka Wiadomości Statystycznych,* 18, Warszawa.
[41] GUS (1973), Stan i perspektywy rozwoju statystyki w Polsce, *Biblioteka Wiadomości Statystycznych*, 17, Warszawa.
[42] GUS (1976), Statystyka i ekonometria w Polsce Ludowej, *Biblioteka Wiadomości Statystycznych*, 25, Warszawa.
[43] GUS (1978), Metodologia badan reprezentacyjnych w GUS – Prace Komisji Matematycznej, *Biblioteka Wiadomości Statystycznych*, 29, Warszawa.
[44] GUS (1979), Statystyka i ekonometria w Polsce Ludowej (wydanie drugie – zmienione), *Biblioteka Wiadomości Statystycznych*, 31, Warszawa.
[45] GUS (1986), Metodyka i organizacja badań budżetów gospodarstw domowych, *Zeszyty Metodologiczne,* 62, Warszawa.
[46] GUS (1987a), Problemy integracji statystycznych badań gospodarstw domowych, *Biblioteka Wiadomości Statystycznych*, 34, Warszawa.
[47] GUS (1987b), Zastosowanie metody reprezentacyjnej w badaniach statystycznych GUS (1981-1986), *Z prac Zakładu Badań Statystyczno-Ekonomicznych* GUS, 166, Warszawa.
[48] GUS (1989), Problemy badań statystycznych metodą reprezentacyjną, *Biblioteka Wiadomości Statystycznych,* 36, Warszawa.
[49] GUS (1992), *Poverty Measurement for Economies in Transition in Eastern European Countries*, Polish Statistical Association, Warsaw.
[50] GUS (1995), Rozwój metodologii badań statystycznych w Polsce, *Biblioteka Wiadomości Statystycznych,* 44, Warszawa.
[51] GUS (1998), Metodologia i organizacja mikrospisów, *Statystyka w Praktyce*, Warszawa.
[52] GUS (2002), Warunki życia ludności w 2001 r., *Studia i Analizy Statystyczne*, Warszawa.
[53] GUS (2005a), Warunki życia ludności w 2004 r., *Studia i Analizy Statystyczne,* Warszawa.
[54] GUS (2005b), Budżet czasu ludności, *Studia i Analizy Statystyczne*, Warszawa.
[55] GUS (2006), Stan zdrowia ludności Polski w 2004 r., *Informacje i Opracowania Statystyczne,* Warszawa.

[56] GUS (2008), *Incomes and Living Conditions of the Population in Poland (report from the EU-SILC survey of 2006),* Warsaw.
[57] GUS (2011a), Metodologia badania budżetów gospodarstw domowych, *Zeszyty Metodologiczne,* Warszawa.
[58] GUS (2011b), Stan zdrowia ludności Polski w 2009 r., *Informacje i Opracowania Statystyczne*, Warszawa.
[59] GUS (2012), *Raport wyników, Narodowy Spis Powszechny Ludności i Mieszkań 2011*, Warszawa.
[60] Hansen M.H., Hurwitz W.N., (1943), On the theory of sampling from a finate population, *Annals of Mathematical Statistics,* 14, 333-362.
[61] Hansen M.H., Hurwitz W.N., (1946), The problem of non-response in sample surveys, *Journal of the American Statistical Association,* 41, 517-529.
[62] Hansen M.H., Hurwitz W.N., Bershad M.A., (1961), Measurement errors in censuses and surveys, *Bulletin of the International Statistical Institute,* 38, 359-374.
[63] Hansen M.H., Hurwitz W.N., Madow W.G., (1953), *Sample Survey Methods and Theory*, Vol. I i II, Wiley, New York.
[64] Kalton G., Citro C.F., (1993), Panel surveys: adding the fourth dimension, *Survey Methodology,* 19, 205-215.
[65] Kalton G., Kasprzyk D., (1986), The treatment of missing survey data, *Survey Methodology,* 12, 1-16.
[66] Kalton G., Kordos J., Platek, R., (1993), *Small Area Statistics and Survey Designs,* Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion. Central Statistical Office, Warsaw.
[67] Kiaer A., (1897), The representative method of statistical surveys (angielskie tłumaczenie z norweskiego oryginału z 1976 r., *Central Bureau of Statistics of Norway*, Oslo.
[68] Kish L., (1965), *Survey Sampling*, New York.
[69] Kish L., (1987), *Statistical Design for Research,* John Wiley and Sons, New York.
[70] Kish L., (1996), Stuletnie zmagania o badania reprezentacyjne, *Wiadomości Statystyczne*, 8, 3-16.
[71] Kordos J., (1959), Szacunek rozkładu ludności według grup zamożności, *Wiadomości Statystyczne*, 3, 4-8.
[72] Kordos J., (1963), Rozkład ludności pozarolniczej według wysokości dochodów na osobę w 1960 r., *Biuletyn Komitetu Przestrzennego Zagospodarowania Kraju PAN*, 8.
[73] Kordos J., (1967), Metoda rotacyjna w badaniach reprezentacyjnych, *Przegląd Statystyczny*, 4, 373-394.
[74] Kordos J., (1968), Zastosowanie metody reprezentacyjnej w statystyce rolniczej, *Przegląd Statystyczny,* 3, 251-264.
[75] Kordos J., (1971), Eksperytentalne badania budżetów rodzinnych metodą rotacyjną, *Przegląd Statystyczny,* 3-4, 227-246.
[76] Kordos J., (1973), On Analysis of Sampling and Non-sampling Errors in Official Statistics in Poland. *Proceeding of the 40th Session of the International Statistical Institute*, XLV, Vienna, 609-616.
[77] Kordos J., (1974), Reprezentacyjne badania warunków bytu ludności, *Przegląd Statystyczny*, 2, 195-209.
[78] Kordos J., (1975), 25 lat działalności Komisji Matematycznej GUS, *Przegląd Statystyczny*, 1, 171-173.
[79] Kordos J., (1982), Metoda rotacyjna w badaniach budżetów rodzinnych w Polsce. *Wiadomości Statystyczne,* 9, 1-6.
[80] Kordos J., (1985), Towards an Integrated System of Household Surveys in Poland, *Bulletin of the International Statistical Institute, (invited paper), vol. 51, Amsterdam,* Book 2, 1.3.1-18.
[81] Kordos J., (1987), Dokładność danych w badaniach społecznych, *Biblioteka Wiadomości Statystycznych, GUS,* 35, Warszawa.
[82] Kordos J., (1988), *Jakość danych statystycznych,* PWE, Warszawa.

[83] Kordos J., (1991), Statystyka małych obszarów a badania reprezentacyjne, *Wiadomości Statystyczne,* 4, 1-5.
[84] Kordos J., (1993), Refleksje nad rozwojem metod badania budżetów gospodarstw domowych, *Wiadomości Statystyczne,* 7, 15-20.
[85] Kordos J., (1995), Metodologiczne problemy kontroli badań statystycznych, W: *Rozwój metodologii badań statystycznych w Polsce, Biblioteka Wiadomości Statystycznych*, 44, 45-56, Warszawa.
[86] Kordos J., (1997), 40 lat badań budżetów gospodarstw domowych w Polsce, *Wiadomości Statystyczne*, 7, 27-42.
[87] Kordos J., (1998), Social Statistics in Poland and its Harmonisation with the European Union Standards, *Statistics in Transition,* 3 (4), 617-639.
[88] Kordos J., (1999), Problemy estymacji danych dla małych obszarów, *Wiadomości Statystyczne*, 1, 85-101.
[89] Kordos J., (2000a), Teoria i sztuka badań reprezentacyjnych, *Wiadomości Statystyczne*, 1, 1-13.
[90] Kordos J., (2000b), Nowy projekt zastosowania estymacji dla małych obszarów, *Wiadomości Statystyczne*, 8, 1-10.
[91] Kordos J., (2001a), Some Data Quality Issues in Statistical Publications in Poland, *Statistics in Transition*, 5 (3), 475-489.
[92] Kordos J., (2001b), Globalne zarządzanie jakością wkracza do statystyki, *Kwartalnik Statystyczny,* nr 2, 9-12.
[93] Kordos J., (2004), Niektóre aspekty jakości w statystyce małych obszarów, W: A. Zeliaś (red.), *Tradycje i obecne zadania statystyki w Polsce*, 95-124, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
[94] Kordos J., (2007), Some Aspects of Post-Enumeration Surveys in Poland, *Statistics in Transition-new series,* 8 (3), 563-576.
[95] Kordos J., (2009), Impact of J. Neyman and W.E. Deming on Sample Survey Practice in Poland, The third Conference of the European Survey Research Association, Warsaw, 2009, http://www.europeansurveyresearch.org/sites/default/files/abstracts.pdf, Chapter II.
[96] Kordos J., (2011), Professor Jerzy Neyman – some reflections, *Lithuanian Journal of Statistic*, 2011, No. 1, s. 112-122. *www.statisticsjournal.lt.*
[97] Kordos J., (2012a), Działalność Komisji Matematycznej GUS w latach 1950-1993, *Wiadomości Statystyczne*, 9, 11-25.
[98] Kordos J., (2012b), Application of Rotation Methods in Sample Surveys in Poland, *Statistics in Transition-new series,* 13 (2), 47-64.
[99] Kordos J., Szarkowski A., (1974), Metoda oceny precyzji podstawowych estymatorów stosowanych w badaniach budżetów rodzinnych, *Wiadomości Statystyczne*, 5, 18-20.
[100] Kordos J., Kursa L., (1997), Metoda reprezentacyjna w rolniczych badaniach statystycznych w Polsce, *Wiadomości Statystyczne*, 9, 12-34.
[101] Kordos J., Paradysz J., (2000), Some Experiments in Small Area Estimation in Poland, *Statistics in Transition*, 4 (4), 963-977.
[102] Kordos J., Zięba-Pietrzak A., (2010), Development of Standard Error Estimation Methods in Complex Household Sample Surveys in Poland, *Statistics in Transition-new series,* 11(2), 231-252.
[103] Kowalczyk B., (2003), Estimation of the population total on the current occasion under second stage unit rotation pattern, *Statistics in Transition,* 6 (4), 503-513.
[104] Kowalski J., (2006), Rotation in sampling patterns, in review for *Journal of Statistical Planning and Inference,* 2006.
[105] Kowalski J., (2009), Optimal Estimation in Rotation Pattern, *J. Statist. Plann. Infer.,* 139 (4), 2429-2436.
[106] Kozak M., (2004), Multivariate Sample Allocation Problem in Two Schemes of Two-Stage Sampling, *Statistics in Transition*, 6 (7), 1047-1054.

[107] Kozak M., (2006), Multivariate Sample Allocation: Application of Random Search Method, *Statistics in Transition*, 7 (4), 889-900.
[108] Kruskal W.H., Mosteller F., (1979-1980), Representative Sampling I, II, III i IV. *International Statistical Review,* 1895-1939.
[109] Krzysztofik W., (1939), Gospodarstwa karłowate w świetle ankiety losowej, Warszawa.
[110] Kubacki J., (2004), Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey, *Statistics in Transition,* 6 (5), 785-796.
[111] Kubacki J., (2006), Remarks on Using the Polish LFS Data for Unemployment Estimation by County, *Statistics in Transition,* 7 (4), 901-916.
[112] Kursa L., Lednicki, B., (2006), The agricultural sample surveys in Poland in transition period, *Statistics in Transition,* 7(5), 981-1008.
[113] Lednicki B., (1973), Zastosowanie metody reprezentacyjnej w badaniu składników wynagrodzeń w przemyśle, *Wiadomości Statystyczne*, 12, 25-26.
[114] Lednicki B., (1974), Badanie efektywności losowania zespołowego na podstawie Narodowego Spisu Powszechnego 1970 roku, *Z prac Zakładu Badań Statystyczno-Ekonomicznych GUS,* 78.
[115] Lednicki B., (1979), Zastosowanie metody reprezentacyjnej w kwartalnych spisach pogłowia zwierząt gospodarskich, W: Metodologia badań reprezentacyjnych w GUS. Prace Komisji Matematycznej, *Biblioteka Wiadomości Statystycznych,* 29, 35-38.
[116] Lednicki B., (1982), Schemat losowania i metoda estymacji w rotacyjnym badaniu budżetów gospodarstw domowych, *Wiadomości Statystyczne*, 9, 7-15.
[117] Lednicki B., (1987), Schemat losowania „próby-matki" do Zintegrowanego Systemu Badań Gospodarstw Domowych w Polsce. W: Problemy integracji badań gospodarstw domowych. *Biblioteka Wiadomości Statystycznych*, 34, 269-279.
[118] Lednicki B., (1989), Badanie efektywności różnych schematów losowania przy wykorzystaniu informacji o cechach dodatkowych, *Wiadomości Statystyczne,* 1, 18-22.
[119] Lednicki B., Wesołowski J., (1994), Lokalizacja próby pomiędzy subpopulacje, *Wiadomości Statystyczne*, 9, 2-4.
[120] Lednicki B., Wieczorkowski R., (2003), Optimal Stratification and Sample Allocation Between Sub–population and Strata, *Statistics in Transition*, 6 (2), 287-305.
[121] Mahalanobis P.C., (1944), On large-scale sample surveys, *Philosophical Transactions of the Royal Society of London,* 231, 329-451.
[122] Mahalanobis P.C., (1952), Some aspects of the design of sample surveys. *Sankhya,* 12, 1-7.
[123] Neyman J., (1925), Contributions to the Theory of Small Samples Drawn from a Finite Population, *Biometrika*, 17.
[124] Neyman J., (1933), Zarys teorii i praktyki badania struktury ludności metoda reprezentacyjna, *Instytut Spraw Społecznych,* Warszawa.
[125] Neyman J., (1934), On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, 97, 558-625.
[126] Neyman J., (1935), On the Problem of Confidence Intervals, *The Annals of Mathematical Statistics*, 6, 111-116.
[127] Neyman J., (1938), Contribution to the Theory of Sampling Human Population, *Journal of American Statistical Association*, 33, 101-116.
[128] Niemiro W., Wesołowski J., (2001), Fixed Precision Optimal Allocation in Two-Stage Sampling, *Applicationes Mathematicae*, 28 (1), 73-82.
[129] Niemiro W., Popiński W., Wesołowski J., Wieczorkowski R., (2002), Optymalna alokacja próby w warunkach niekompletnej realizacji badania, *Wiadomości Statystyczne*, 9, 1-9.
[130] Niemiro W., Wesołowski J., (2012), Linear estimation and prediction under model-design approach with small area efects, *Statistics: A Journal of Theoretical and Applied Statistics*, 46 (4), 523-547.

[131] O'Muircheartaigh C., Wong S. T., (1981), The impact of sampling theory on survey sampling practice: a review, *Bulletin of International Statistical Institute Invited Paper*, 49 (I), 437-446.
[132] Paradysz J., (1998), Small Area Statistics in Poland – First Experiences and Application Possibilities, *Statistics in Transition,* 3 (5), 1003-1015.
[133] Patterson H.D., (1950), Sampling on successive occasions with partial replacements of units, *Journal of the Royal Statistical Society*, ser. B, 12, 241-255.
[134] Pawłowska J., (1969), Efektywność metod estymacji i schematów losowania w badaniach pogłowia zwierząt gospodarskich, *Z prac Zakładu Badań Statystyczno-Ekonomicznych,* GUS, 15, Warszawa.
[135] Pawłowski Z., (1972), *Wstęp do statystycznej metody reprezentacyjnej*, PWN, Warszawa
[136] Pekasiewicz D., Pruska K., (2002), Analysis of Distribution of Some Estimators in Small Area Statistics, *Folia Oeconomica*, 156, 91-111.
[137] Piekałkiewicz J., (1934), Sprawozdanie z badań składu ludności robotniczej w Polsce metodą reprezentacyjną, *Instytut Spraw Społecznych,* Warszawa.
[138] Platek R., Rao J.N.K., Särndal C.E., Singh, M.P., (1987), *Small Area Statistics – An International Symposium*, John Wiley & Sons, New York.
[139] Popiński W., (2006), Development of the Polish Labour Force Survey, *Statistics in Transition,* 7 (5), 1009-1030.
[140] Rao J.N.K., (2005), Interplay between sample survey theory and practice: An appraisal, *Survey Methodology*, 31(2), 117-138.
[141] Rao, J.N.K., Graham, J.E. (1964), Rotation designs for sampling on repeated occasions. *Ann. Math. Statist*. 35, 492-509.
[142] Riga (1999), *Small Area Estimation – Conference Proceedings*, Riga, Latvia.
[143] Särndal C.-E., Swensson B., Wretman, J., (1992), *Model Assisted Survey Sampling*, Springer, New York.
[144] Smith T.M.F., (1976), The foundations of Survey Sampling: a review, *Journal of Royal Statistical Society,* A. 139, 183-204.
[145] Smith T.M.F., (1994), Sample Surveys 1975-1990; an age of reconciliation? *International Statistical Review,* 62, 3-34.
[146] Steczkowski J., (1988), *Zastosowanie metody reprezentacyjnej w badaniach społeczno-ekonomicznych,* PWN, Warszawa.
[147] Steinhaus H., (1956), Liczby złote i żelazne, *Zastosowania Matematyki*, 3, 51-65.
[148] Steinhaus H., (2000), Między duchem a materią pośredniczy matematyka, PWN, Warszawa.
[149] Sukhatme P.V., Sukhatme, B.V. (1970), *Sampling Theory of Surveys with Applications*. Asia Publishing House, London.
[150] Szabłowski P.J., Wesołowski J., Wieczorkowski R., (1996), Indeks zgodności jako miara jakości danych, *Wiadomości Statystyczne,* 4 , 43-49.
[151] Szarkowski A., Witkowski J., (1994), The Polish Labour Force Survey, *Statistics in Transition*, 1 (4), 467-483.
[152] Szreder M., (2004), *Metody i techniki sondażowych badań opinii*, PWE, Warszawa.
[153] Szulc S. (1967), *Metody statystyczne*, PWE, Warszawa.
[154] Szutkowska J., (2012), Zarządzanie jakością w statystyce publicznej: modele, standardy, metody i narzędzia, *Wiadomości Statystyczne*, 11, s. 38-51.
[155] UN Statistics Division (2010), *Post Enumeration Surveys*, Operational guidelines, Technical Report, New York, April 2010.
[156] Verma J. i Betti G. (2006), EU Statistics on Income and Living Conditions (EU-SILC): Choosing the Survey Structure and Sample Design, *Statistics in Transition*, 7 (5), 935-970.
[157] Wesołowski J. (2004), Problemy estymacji dla małych obszarów, *Wiadomości Statystyczne,* 3, 9-14.
[158] Wesołowski J. (2010), Recursive optimal estimation in Szarkowski rotation scheme, *Statistics in Transition-new series,* 11(2), 267-285.

[159] Wierzchosławski S., (1995), Węzłowe problemy metodologii badań statystycznych zjawisk społeczno-ekonomicznych w Polsce, W: *Rozwój metodologii badań statystycznych w Polsce, Biblioteka Wiadomości Statystycznych*, 44, 18-44, Warszawa.
[160] Wywiał J., (1995), *Wielowymiarowe aspekty metody reprezentacyjnej*. Ossolineum, Wrocław–Warszawa–Kraków.
[161] Wywiał J., (1996), On two-phase sampling for stratification, *Statistics in Transition*, 2 (6), 971-977.
[162] Wywiał J., (1998), Estimation of population average on the basis of strata formed by means of discrimination functions. *Statistics in Transition,* 3 (5), 903-912.
[163] Wywiał J., (1999), Generalisation of Singh and Srivastava's schemes providing unbiased regression estimatiors. *Statistics in Transition,* 4 (2), 259-281.
[164] Wywiał J., (2000), On precision of Horvitz-Thompson strategies, *Statistics in Transition,* 4 (5), 779-798.
[165] Wywiał J., (2001), Estimation of population mean on the basis of non-simple sample when non-response error is present. *Statistics in Transition,* 5(3), 443-450.
[166] Wywiał J., (2007), Simulation analysis of accuracy estimation of population mean on the basis of strategy dependent on sampling design proportionate to the order statistic of an auxiliary variable. *Statistics in Transition-new series,* 8(1), 125-137.
[167] Wywiał J., (2010), *Wprowadzenie do metody reprezentacyjnej*, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice.
[168] Wywiał J., Gamrot W., (red.), (2010), *Survey Sampling Methods in Economic and Social Research,* Katowice University of Economics, Katowice.
[169] Wywiał J., Żądło T., (red.), (2009), *Survey Sampling Methods in Economic and Social Research,* Katowice University of Economics, Katowice.
[170] Yates F., (1980), *Sampling Methods for Censuses and Surveys*,(4th ed.), London.
[171] Zarkovich S.S., (1966). *Quality of Statistical Data*. FAO, Rome.
[172] Zasępa R., (1958), Problematyka badań reprezentacyjnych GUS w świetle konsultacji z prof. J. Neymanem, *Wiadomości Statystyczne,* 6, 7-12.
[173] Zasępa R., (1962), *Badania statystyczne metodą reprezentacyjną*, PWN, Warszawa.
[174] Zasępa R., (1968), Zastosowanie metody reprezentacyjnej przy spisach rolnych, *Studia i Prace Statystyczne*, GUS, 10, Warszawa.
[175] Zasępa R., (1972), *Metoda reprezentacyjna*, PWE, Warszawa.
[176] Zasępa R., (1993a). Precyzja wyników badań budżetów rodzinnych, „*Wiadomości Statystyczne*", 3, 30-32.
[177] Zasepa R., (1993b), Use of Sampling Methods in Population Censuses in Poland, *Statistics in Transition,* 1(1), 69-78.

WSPÓŁZALEŻNOŚĆ POMIĘDZY ROZWOJEM TEORII I PRAKTYKI BADAŃ REPREZENTACYJNYCH W POLSCE

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

Autor bada współzależności między rozwojem teorii i praktyki badań reprezentacyjnych w Polsce w ciągu ponad 60 lat. Rozpoczyna od klasycznej pracy Neymana (Neyman, 1934), która dała teoretyczne podstawy probabilistycznego podejścia wyboru próbki umożliwiające wnioskowanie z badań reprezentacyjnych. Główne idee tej pracy były najpierw opublikowane po polsku w 1933 r. (Neyman, 1933) i miały istotny wpływ na a praktykę badań reprezentacyjnych Polsce przed i po II Wojnie Światowej. Badania reprezentacyjne prowadzone w latach 1950-tych i 1960-tych były konsultowane z J. Neymanem w czasie jego wiz w Polsce w latach 1950 i 1958 (Fisz, 1950; Zasępa, 1958).

Praktyczne problemy występujące przy planowaniu i analizie badań reprezentacyjnych Polsce były częściowo rozwiązywane przez Komisję Matematyczną GUS powołaną w końcu 1949 r. jako organ doradczy i opiniodawczy Prezesa GUS. Komisja ta składała się ze specjalistów zarówno z GUS jak i ośrodków naukowo-badawczych w kraju (Kordos, 2012a). Komisja działała do 1993 r. i wpływała w istotny sposób na praktykę badań reprezentacyjnych Polsce.

Specjalną uwagę poświęcono wpływowi teorii badań próbkowych na świecie i w Polsce na praktykę badań reprezentacyjnych w Polsce, a w szczególności na plany i schematy losowania i metody estymacji w badaniach prowadzonych *w czasie,* a głównie *metodzie rotacyjnej* (Greń, 1969; Kordos, 1967, 2012b; Kowalczyk, 2004; Kowalski, 2006; Lednicki, 1982; Popiński, 2006; Wesołowski, 2010); *lokalizacji próby* (Bracha et al.,2004b, Greń, 1964, 1966; Lednicki, 1979, 1989; Lednicki i Wesołowski, 1994;), *metodom estymacji* (Bracha, 1996, 1998; Greń, 1970; Kordos, 1982; Lednicki, 1979, 1987, Wesołowski, 2004, Zasępa, 1962, 1972), *jakości danych* (Kordos, 1973, 1988; Zasępa, 1993) oraz *metodom estymacji dla małych obszarów* (Bracha, 1994, 2003; Bracha et al., 2004b; Dehnel, 2010; Domański, Pruska, 2001; Golata, 2004ab, 2012; Kalton et al., 1993; Kordos, 1991, 2000b; Kordos, Paradysz, 2000; Niemiro, Wesołowski, 2012; Paradysz, 1998; Wesołowski, 2004). W zakończeniu prowadzone są rozważania na temat przyszłych zapotrzebowań na informacje z badań reprezentacyjnych.

Słowa kluczowe: badanie reprezentacyjne, metoda estymacji, złożone badanie reprezentacyjne, jakość danych, statystyka małych obszarów, metoda rotacyjna, metoda alokacji próby

THE INTERPLAY BETWEEN SAMPLE SURVEY THEORY AND PRACTICE IN POLAND

Abstract

The author examines interplay between sample survey theory and practice in Poland over the past 60 years or so. He begins with the Neymans's (1934) classic landmark paper which laid theoretical foundations to the probability sampling (or design-based) approach to inference from survey samples. Main ideas of that paper were first published in Polish in 1933 (Neyman, 1933) and had a significant impact on sampling practice in Poland before and after the World War II. Sample surveys conducted in 1950s and 1960s were consulted with J. Neyman during his visits in Poland in 1950 and 1958 (Fisz, 1950a; Zasepa, 1958).

Some practical problems encountered in the design and analysis of sample surveys were partly solved by the Mathematical Commission of the CSO which was established in 1949, as an advisory and opinion-making body to the CSO President in the field of sample surveys. The Commission concentrated specialists in the sampling methods both from the CSO and research centres in the country (Kordos, 2012a). The Commission had a significant impact on sampling practice in Poland and was active till 1993.

Special attention is devoted to the impact of sampling theory on sampling practice in Poland, and particularly on sample designs and estimation methods in: *sampling in time and rotation methods* (Kordos, 1967, 2012b; Kowalczyk, 2004; Kowalski, 2006; Wesołowski, 2010); *sample allocation and estimation methods* (Bracha, 1994, 1996, 2003; Greń, 1964, 1966, 1969, 1970; Kordos, 1969, 1973, 1982; Wesołowski, 2004, Zasępa, 1962, 1972, 1993); *data quality* (Kordos, 1973, 1988); *estimation methods for small areas* (Dehnel, 2010; Domański, Pruska, 2001; Golata, 2004ab, 2012; Kalton et al., 1993; Kordos, 1991, 2000b, 2004; Kordos, Paradysz, 2000; Niemiro, Wesołowski, 2012; Paradysz, 1998; Wesołowski, 2004). Concluding remarks are given at the end.

Key words: sample survey, estimation methods, complex sample surveys, data quality, estimation for small area, rotation sampling, sample allocation methods

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

DANIEL KOSIOROWSKI[1]

GŁĘBIA POŁOŻENIA-ROZRZUTU W STRUMIENIOWEJ ANALIZIE DANYCH EKONOMICZNYCH

1. WPROWADZENIE

Współczesna gospodarka w sposób ciągły generuje gigantyczne zbiory danych. Analiza, monitorowanie, decydowanie w oparciu o wielkie zbiory danych stanową bez wątpienia wieloaspektowe wyzwanie dla współczesnej statystyki (por. Aggerwal, 2007).

Przykładem tego typu wyzwań jest tzw. analiza strumienia danych (strumieniowe przetwarzanie danych). Analiza taka przykładowo może obejmować monitorowanie setek tysięcy finansowych szeregów czasowych w celu znalezienia użytecznych inwestycyjnie zależności pomiędzy nimi, analizę danych generowanych przez stacje pogodowe w pewnym obszarze oceanu, monitorowanie centrum miasta za pomocą systemu kamer, decydowanie co do podjęcia interwencji na rynku zbóż w oparciu o dane dostarczane przez giełdy towarowe.

Ujmując zagadnienie nieprecyzyjnie możemy określić strumień danych jako „*nieokreślonej długości ciąg z reguły wielowymiarowych obserwacji*" (por. Szewczyk 2010). Należy zwrócić uwagę, że w przypadku tradycyjnie rozumianej analizy procesu stochastycznego, powiedzmy $\{X_t\}$, zakładamy ustalony przedział czasowy, powiedzmy $[0, T]$. Nasze obliczenia dotyczą tego przedziału a więc wnioskujemy na podstawie informacji uzyskanej do chwili T. W przypadku analizy strumienia danych nie ustalamy przedziału badania $[0, T]$. Każda kolejna chwila oznacza nową analizę procesu stochastycznego. Strumieniowe przetwarzanie danych, analizę strumienia danych można określić, jako sekwencję analiz procesu stochastycznego. Terminologia wywodzi się z informatyki, gdzie tego typu zagadnienia były rozważane po raz pierwszy. Oczywiście strumieniowe przetwarzanie danych można rozpatrywać na gruncie teorii procesów stochastycznych i w szczególności na gruncie teorii szeregów czasowych. W nawiązaniu do uwag jednego z Recenzentów dotyczących związków pomiędzy analizą procesów stochastycznych a analizą strumieni danych – autor uważa, że w obrębie znanych mu prac z zakresu zastosowań procesów stochastycznych w ekonomii najbliższe praktyki strumieniowego przetwarzania danych są prace duetu Aït-Sahalia i Jacod (por. Aït-Sahalia, Jacod, 2012 i odniesienia do literatury tamże) dotyczące badań procesów dyfuzji ze skokami. Prace te jednakże dotyczą jednej analizy procesu obserwowanego na pewnym przedziale

[1] Artykuł powstał w części dzięki wsparciu Narodowego Centrum Nauki w postaci grantu DEC-011/03/B/HS4/01138

czasu [0,T] – nie dotyczą rodziny takich analiz. Wspomniani autorzy skupiają swą uwagę na jednowymiarowych modelach parametrycznych o jednym reżimie. Przyjmują raczej mocne założenia odnośnie tychże modeli oraz odnośnie sposobu pobierania danych. Nie rozważają obserwacji odstających. Elegancja prezentacji zagadnień przez wspomniany duet uczonych stanowi punkt odniesienia i cel dla autora niniejszej pracy w przyszłości.

W literaturze dotyczącej analizy strumieni danych, strumieniowego przetwarzania danych w zasadzie nie podaje się wprost odwołania do probabilistycznego modelu danych. Jednakże wczytując się w tę literaturę można pokusić się o stwierdzenie, że analiza taka jest w istocie rodziną analiz procesu stochastycznego odznaczających się następującymi cechami:

1. Obserwacje generowane są przez proces, w którym ma miejsce nieliniowa zależność teraźniejszości od przeszłości.
2. Obserwacje modeluje się na ogół przez proces niestacjonarny, którego nie da się sprowadzić do procesu stacjonarnego za pomocą różnicowania, usunięcia deterministycznego trendu. Proces na ogół odznacza się występowaniem pewnej ilości reżimów. Typ niestacjonarności, liczba i charakterystyki reżimów mogą zmieniać się w czasie.
3. Analizę strumienia prowadzimy opierając się na stale aktualizowanej próbie – na podstawie ustalonej długości ruchomego okna (można rozważać okna różnej długości dla różnych skal czasu – sekund, minut, dni itd.). Na podstawie takiej stale aktualizowanej próby podejmujemy decyzje, na jej podstawie monitorujemy położenie, rozrzut strumienia.
4. Strumienie na ogół liczą setki tysięcy wielowymiarowych obserwacji. Z reguły dane z racji swej wielkości nie są magazynowane w pamięci komputera – muszą być przetwarzane na bieżąco (ang. *on–line processiong)*.
5. Dane napływają do obserwatora z reguły w nierównych odstępach czasu, w pakietach nierównej wielkości. Można założyć, że modelem strumienia jest proces stochastyczny z czasem ciągłym. Wówczas mamy na uwadze sytuację, gdy częstość próbkowania obserwacji ze strumienia jest zmienną losową. Można założyć stosownie skonstruowany proces dyskretny odwołując się np. do teorii procesów podporządkowanych, warunkowych procesów trwania, bądź tak jak w niniejszej pracy wyjść od takiego procesu, który losowo generuje sygnał (odpowiednio zdefiniowany) w chwilach równo od siebie odległych.
6. Do analizy strumieni stosuje się są na ogół procedury nieparametryczne, które muszą spełniać wysokie wymagania w zakresie złożoności obliczeniowej, które muszą radzić sobie z problemem „rzadkości danych" w wielu wymiarach (por. Hastie i in. 2009).

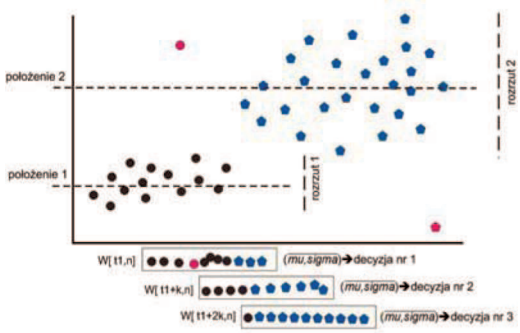
W niniejszej pracy zakładamy, że strumień generowany jest przez pewną konkretną postać ogólnego modelu określanego mianem CHARME (por. Stockis i in. 2010). Rozważamy tym samym proces stochastyczny z czasem dyskretnym o ustalonej liczbie reżimów. Zakładamy, że w obserwowanych przez nas danych występują obserwacje od-

stające. Mamy tutaj na uwadze sytuację, gdy na badany proces działa tzw. *addytywny proces odstawania* (AO) (ang. additive outliers process) – przyjmujemy ramy pojęciowe zaproponowane w klasycznym podręczniku Marona i in. (2006). Niech x_t oznacza proces warunkowo stacjonarny[2], niech v_t oznacza stacjonarny proces odstawania. Niech $P(v_t = 0) = 1 - \varepsilon$, co oznacza, że „niezerowa" część procesu v_t pojawia się z prawdopodobieństwem ε. W modelu AO, zamiast x_t obserwujemy $y_t = x_t + v_t$ przy czym zakłada się, że procesy x_t i v_t są wzajemnie niezależne. AO można określić, jako *proces błędów grubych*, obserwacje odstające na ogół są izolowane. W niniejszym artykule skupiamy naszą uwagę na procesie podejmowania decyzji na podstawie stale uaktualnianej niewielkiej próby ze strumienia. Decyzje dotyczą m.in. prognozowania kolejnych wartości strumienia, prognozowania i monitorowania charakterystyk rozrzutu, położenia i skośności, (bezwarunkowych i warunkowych względem obserwowanej próby w przeszłości), monitorowania zależności pomiędzy teraźniejszością i przeszłością strumienia. Naszym zadaniem jest stworzenie stosownych narzędzi umożliwiających nam odczytanie sygnału zawartego w strumieniu w sytuacji występowania obserwacji odstających. Należy jednakże podkreślić, że w przeciwieństwie do nauk inżynieryjnych (dane = deterministyczny sygnał+ losowy szum) przez sygnał rozumiemy relację pomiędzy charakterystykami liczbowymi probabilistycznego modelu danych[3]. W zasadzie w przyjętych przez nas dalej ramach pojęciowych odczytanie sygnału wiążemy ze wskazaniem reżimu procesu generującego strumień. Zagadnienie schematycznie przedstawiają rys. 1-2. W niniejszej pracy nie nawiązujemy bezpośrednio do teoretycznych trudności związanych z pomiarem odporności statystyki w przypadku analizy procesów stochastycznych. *Odporność naszych propozycji rozumiemy w duchu jednolitego i ogólnego podejścia Gentona i Lucasa (2003) jako odporność reguły decyzyjnej określonej na stale uaktualnianej próbie ze strumienia (za punkt odniesienie bierzemy np. medianę w przestrzeni decyzji, rozważamy różne funkcje straty np. LINEX).* Według Gentona i Lucasa (2003) krytyczna cecha estymatora sprowadza się do tego, że ten przyjmuje różne wartości dla różnych realizacji próby. Jeżeli możliwe jest kontinuum prób a estymator jest ciągły, to oczekujemy kontinuum jego wartości. Załamanie estymatora polega na tym, że ta jego własność zanika, estymator przyjmuje jedynie skończoną liczbę różnych wartości pomimo kontinuum możliwych prób. Jeżeli dla przykładu rozważany przez nas model strumienia dopuszcza powiedzmy 10 reżimów, a procedura mająca wskazać te reżimy wskazuje jedynie jeden z nich, to powiemy, że procedura łamie się.

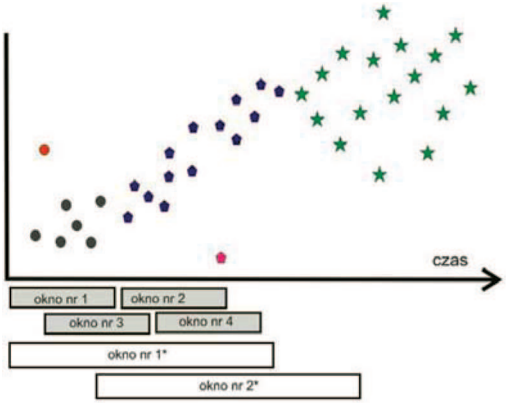
Można umownie wyróżnić dwa nurty podejść do analizy strumieni danych – nurt związany z metodami eksploracyjnej analizy bardzo wielkich zbiorów danych (ang. ve-

[2] Mówimy, że jednowymiarowy proces jest warunkowo stacjonarny, jeżeli jego rozkładu warunkowe są niezmiennicze względem przesunięć w czasie (por. Shalizi, Kantorovich, 2007 def. 51 str. 35)

[3] W niniejszym artykule zakładamy, że dane generuje pewien niestacjonarny proces stochastyczny. Sygnał utożsamiamy z charakterystykami liczbowymi jego modelu(li). Jednakże o ile zmienimy rozumienie sygnału – można rozważać strumienie generowane przez procesy stacjonarne bądź układy stricte deterministyczne. W kontekście zastosowań tematyki w ekonomii – przyjęte ramy wydają się być najwłaściwsze.



Rysunek 1. Ilustracja zagadnienia decydowania co do zmiany położenia – rozrzutu na podstawie ruchomego okna



Rysunek 2. Trzy reżimy strumienia danych. Dane zawierają obserwacje odstające, rozważamy ruchome okna różnej długości

ry big high-dimensional data mining) oraz nurt związany z klasyczną nieparametryczną analizą szeregów czasowych (por. Fan, Yao, 2005). W obrębie pierwszego nurtu (por. Aggerwal, 2007) wyróżnić można m.in.: dynamiczną redukcję wymiaru zagadnienia za pomocą tzw. mikro-skupisk, badanie dynamicznych klasyfikacji, stosowanie adaptacyjnej metody najbliższych sąsiadów, wykorzystywanie drzew regresyjnych i klasyfikacyjnych, wykorzystanie sieci neuronowych, sieci bayesowskich. Drugi nurt wiąże się z adaptacjami metod nieparametrycznej analizy szeregów czasowych. Mamy tutaj na uwadze adaptacje lokalnej liniowej, lokalnej wielomianowej regresji w tym szereg wariantów nieparametrycznej *regresji Nadaraya-Watsona* (patrz Hall i in., 1999), metody wykorzystujące wielomiany ortogonalne, regresję nieliniową z ograniczeniami (np. metody LOESS, LASSO por. Hastie i in., 2009), sklejki itd. Należy podkreślić, że w przypadku analizy strumieni danych na ogół wielowymiarowych niezmiernie istotne jest, aby procedura radziła sobie z tzw. *„przekleństwem wielowymiarowości" – rzadkość danych* (ang. sparse data) *w wielu wymiarach*. Owo przekleństwo sprawia m.in., że dla przykładu dobre statystyczne własności jednowymiarowej regresji Nadaraya-

Watsona zanikają w wielu wymiarach, istotność statystyczna oszacowań wielowymiarowych modeli stosowanych w empirycznych finansach budzi poważne wątpliwości (por. Kosiorowski, Snarska, 2012).

W literaturze jak dotychczas nie jest znanych wiele odpornych metod analizy strumieni danych. Wiąże się to między innymi z trudnościami z rozumieniem odstawania w przypadku strumieni generowanych przez model o wielu reżimach. Pojawia się dla przykładu pytanie *czy rozumienie odstawania powinno się w takim przypadku wiązać z konkretnym reżimem procesu*? Co ciekawe w przypadku analizy strumienia z jednostkami odstającymi stosowana procedura powinna być odporna, jednak nie bardzo odporna (tzn. jej punkt załamania nie powinien osiągać maksymalnej możliwej wartości) – tak, aby pomijała wpływ obserwacji odstających, lecz jednocześnie była wrażliwa na zmianę reżimu modelu.

W pracy proponujemy proste i przyjazne dla użytkownika metody analizy strumienia danych ekonomicznych odwołujące się do tzw. koncepcji głębi danych (por. Kosiorowski, 2012). W kolejnych częściach artykułu prezentujemy odpowiednio: w drugiej części wprowadzamy model strumienia danych, w trzeciej przedstawiamy wybrane zagadnienia związane z głębią położenia-rozrzutu Mizery i Müller, w czwartej przedstawiamy propozycje procedur wykorzystujących tę głębię, w piątej wyniki badań własności propozycji za pomocą symulacji, artykuł kończymy konkluzjami i literaturą.

2. MODEL STRUMIENIA DANYCH EKONOMICZNYCH

W literaturze nie jest znanych wiele modeli strumienia danych, do nielicznych należy zaliczyć propozycję Hahsler i Dunhamr (2010), w której rozważa się zmienny w czasie łańcuch Markowa dla mikro skupisk. Wydaje się jednak, że model strumienia danych można skonstruować na podstawie jednego z wykorzystywanych w ekonometrii modeli dla zjawisk o zmiennym reżimie np. model TAR (ang. treshold auroregressive model) bądź jego nieliniową wersję FAR (ang. functional auroregressive model) (por. Fan, Yao, 2005).

Wybór modelu strumienia danych wykorzystywanego w niniejszej pracy podyktowany jest względami wygody oraz jego elastycznością w zakresie opisu szerokiego spektrum możliwych zjawisk. Zdecydowano się budować model strumienia danych na bazie *warunkowego heteroskedastycznego nieparametrycznego modelu* CHARN postaci

$$X_t = m(X_{t-1}, ., X_{t-p}) + \sigma(X_{t-1}, ., X_{t-p})\epsilon_t, \tag{1}$$

o dowolnych lecz ustalonych funkcjach $m(\cdot)$ oraz $\sigma(\cdot)$ (np. $m(\mathbf{x}) = E(X_t|[X_{t-1}, ..., X_{t-p}] = \mathbf{x})$, $\sigma^2(\mathbf{x}) = Var(X_t|[X_{t-1}, ..., X_{t-p}] = \mathbf{x})$, gdzie $\mathbf{x} = (x_{t-1}, ..., x_{t-p})$ oraz o niezależnych o tym samym rozkładzie innowacjach ε_t o wartości oczekiwanej zero (zobacz Fan, Yao, 2005).

Model (1) stanowi punkt wyjścia dla procesu budowy modelu strumienia danych ekonomicznych. Podkreślmy jednakże, że w kontekście analizy strumieni danych ekonomicznych z zasady nie zakładamy, że obserwowany proces ma tę samą funkcję trendu m oraz tę samą zmienność σ w każdej chwili. Nie zakładamy też, że funkcje

te zmieniają się powoli, stopniowo w czasie. W takim oto kontekście skupiamy naszą uwagę na modelu CHARME (ang. *conditional heteroscedastic autoregressive mixture of experts*) (zobacz Stockis i in., 2010). Model CHARME stanowi ogólne podejście do modelowania szeregów czasowych o zmiennym reżimie. Układ ekonomiczny oscyluje pomiędzy pojedynczymi stanami, których same dynamiką rządzi model CHARN (1). CHARME w przypadkach szczególnych obejmuje wiele nieliniowych szeregów czasowych jak np. modele dwuliniowe, modele progowe TAR. W modelu CHARME dynamiką strumienia $\{X_t\}$ rządzi ukryty łańcuch Markowa $\{Q_t\}$ na skończonym zbiorze stanów $\{1, 2, ., K\}$, sam model zdefiniowany jest w następujący sposób:

$$X_t = \sum_{k=1}^{K} S_{tk}(m_k(X_{t-1}, ., X_{t-p}) + \sigma_k(X_{t-1}, ., X_{t-p})\epsilon_t) + b_t\theta_t, \tag{2}$$

gdzie $S_{tk} = 1$ dla $Q_t = k$ oraz $S_{tk} = 0$ w przeciwnym przypadku, m_k, σ_k, $k = 1, ..., K$, są nieznanymi funkcjami, ϵ_t są niezależnymi zmiennym losowymi o średniej zero, człon $b_t\theta_t$ reprezentuje obserwacje odstające typu AO (por. Maronna i in., 2006), b_t jest nieobserwowalną binarną zmienną losową wskazującą pojawienie się obserwacji odstającej w chwili t, natomiast θ_t oznacza losową wielkość obserwacji odstającej. Aby uniknąć „*przekleństwa wielowymiarowości*", postulujemy przyjąć $p = 1$ bądź $p = 2$.

Wprowadzając ukryty łańcuch Markowa rządzący zmianami reżimów dopuszczamy występowanie nagłych wartości strumienia. Dodatkowo przyjmujemy następujące założenia odnośnie strumienia i wykorzystywanego do jego opisu modelu CHARME:

1. Losowa liczba obserwacji odstających w strumieniu, pojawiających się do chwili t dana jest za pomocą $N_t = \sum_{i=1}^{t} b_i$ oraz jest ograniczona według prawdopodobieństwa warunkowo względem N_{t^*}, $t^* < t$. Oznacza to, że zamiast ustalać z góry liczbę obserwacji odstających stosujemy ograniczenie na prawdopodobieństwo ich pojawienia się. Umożliwiamy tym samym rozróżnienie pomiędzy częstymi, zwykłymi szokami oraz rzadkimi zdarzeniami odstającymi.
2. Nie zakładamy jakiejkolwiek wiedzy, co do liczby i położenia obserwacji odstających, nie nakładamy też ograniczeń na strukturę zależności $\{b_t\}$.
3. Strumień, który jest modelowany za pomocą modelu CHARME jest *warunkowo stacjonarny* (por. Shalizi, Kantorovich, 2007)
4. Zakładamy, że ukryty łańcuch Markowa będzie zmieniał swą wartość rzadko, tzn. obserwowany proces będzie podlegał temu samemu reżimowi przez względnie długi okres czasu zanim nastąpi zmiana. Stawiamy tym samym ograniczenia, co do postaci macierzy przejścia $\mathbf{P} = [p_{rs}]$, $r, s = 1, ..., k$, dla łańcucha Q_t postaci $p_{rr} >> p_{rs}$.

Niech $x_1, x_2, ...$ oznacza strumień danych generowany przez model (2). Przez okno $W_{i,n}$ rozumiemy ciąg punktów kończący się w punkcie x_i i o długości n: $W_{i,n} = (x_{i-n+1}, ..., x_i)$. Czasem wygodnie jest rozważać okno $W_{[i,j]}$ – podciąg strumienia danych pomiędzy i-tą oraz j-tą obserwacją. Spora część technik analizy strumieni danych opiera się na monitorowaniu różnych odległości pomiędzy rozkładami empirycznymi

wyznaczanymi na podstawie dwóch bądź więcej okien $W_{i,n}$ i $W_{j,n}$. Można przy tym rozważać ustalone okna, ruchome okna itd. Rozważając wielowymiarowy strumień danych $\mathbf{x}_1, \mathbf{x}_2, ...$ podobnie badamy zachowanie się wielowymiarowych okien $\mathbf{W}_{i,n}$ (por. Kosiorowski, Snarska, 2012).

PROBLEM 1: Rozważmy sytuację, gdy w oparciu o stale uaktualniane okna $W_{i,n}, W_{i+1,n}, ...$ przewidujemy odpowiednio kolejne obserwacje $\hat{x}_{i+1}, \hat{x}_{i+2}, ...$ bądź kolejne okna $\hat{W}_{i+1,k}$,$\hat{W}_{i+2,k}$,..., $k \ll n$. Chcielibyśmy znaleźć optymalną procedurę prognostyczną, tzn. minimalizującą pewną funkcję straty w sytuacji, gdy strumień danych zawiera jednostki odstające.

PROBLEM 2: Na podstawie monitorowania ruchomego okna $W_{i,n}$, $i = 1, 2, ...$ zamierzamy wykryć bezwarunkowe zmiany w modelu generującym strumień danych. Jeżeli założymy pewien model postaci (2), to naszym celem jest wykrycie stanu Q_k ukrytego łańcucha Markowa, a w konsekwencji funkcji m_k oraz σ_k występujących w (2).

PROBLEM 3: Monitorujemy strumień danych $x_1, x_2, ...$ oraz naszym celem jest wykrycie zmiany rozkładu warunkowego okna $W_{i+1,n}$, pod warunkiem zaobserwowanego okna $W_{i,n}$, $i = 1, 2, ...$, tzn. zmiany $P(W_{i+1,n} \in A | W_{i,n} = \mathbf{x})$, $A \subset \mathrm{R}$, dla $i = 1, 2, ...$ W ramach pojęciowych, wyznaczonych przez model (2) naszym celem jest wykrycie zmian macierzy przejścia ukrytego łańcucha Q_k, bądź zmiany funkcji m_k oraz σ_k oznaczających przykładowo warunkowe położenie i warunkowy rozrzut.

PROBLEM 4: Monitorujemy d-wielowymiarowy strumień danych $\mathbf{x}_1 = (x_{11}, ..., x_{1d})$, $\mathbf{x}_2 = (x_{21}, ..., x_{2d})$,..., a naszym celem jest wykrycie zmian łącznego (warunkowego) rozkładu $\mathbf{x}_i$ na podstawie $\mathbf{W}_{i-1,n}$, $i = 1, 2, ...$ W szczególności jesteśmy zainteresowani wykrywaniem zmian postaci związku liniowego pomiędzy współrzędnymi wektorów $\mathbf{x}_i$.

Stosowanie w zagadnieniu predykcji (problem 1) lokalnego liniowego, lokalnego wielomianowego modelowania wymaga, aby zmiany pomiędzy reżimami były gładkie (a w konsekwencji dawały się lokalnie aproksymować za pomocą funkcji liniowej bądź wielomianowej) oraz aby strumień nie zawierał obserwacji odstających. Stosowanie globalnych sklejek z jednej strony wymaga „zatrzymania" analizy, aby można było taki model oszacować, z drugiej strony napotykamy problemy związane ze złożonością obliczeniową oraz przekleństwem wielowymiarowości. Stosowanie wielomianów ortogonalnych z kolei zmusza nas od ograniczenia się do kowariancji jako miary zależności obserwacji w czasie – co nie jest właściwe w przypadku strumieni o na ogół nieliniowej strukturze zależności w czasie.

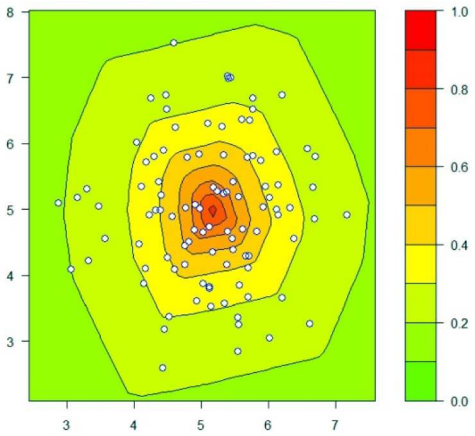
Jest powszechnie wiadomo, że oszacowania momentów procesów stochastycznych są użyteczne jedynie o ile poczynimy bardzo mocne założenia odnośnie tychże procesów. Mamy tutaj na uwadze m.in. istnienie momentów odpowiednich rzędów, jednoznaczność opisu rozkładu za pomocą momentów, postać funkcji autokowariancji itd. (por. Jacod, Shiryaev, 2003). Podobnie rzecz się przedstawia z prognozowalnością procesów. Strumienie na ogół nie spełniają takich założeń. Jednakże zamiast opisywać proces za pomocą miar konstruowanych na podstawie momentów możemy go opisywać

za pomocą miar wykorzystujących statystyki pozycyjne i porządkowe. Opis procesu w kategoriach indukowanych przez statystyki pozycyjne i porządkowe jest możliwy nawet w sytuacji, gdy nie da się opisać procesu za pomocą statystyk wykorzystujących momenty. W takim oto kontekście proponujemy wykorzystać narzędzia koncepcji głębi danych do odpornej analizy strumienia danych. W celu oszacowania niepewności związanej z analizą strumienia danych proponujemy wykorzystać metody Monte Carlo.

3. GŁĘBIA STUDENTA

Dla danego rozkładu prawdopodobieństwa F na R^d, $d \geq 2$ statystyczna funkcja głębi $D(\mathbf{x}, F)$ przyporządkowuje $\mathbf{x} \in \mathrm{R}^d$ liczbę z przedziału [0,1] będącą miarą centralności tej obserwacji względem rozkładu F. Statystyczne funkcje głębi kompensują brak naturalnego porządku w R^d, $d \geq 2$, poprzez orientowanie punktów względem centrum – względem d-wymiarowej mediany indukowanej przez konkretną funkcje głębi. Wyższe wartości głębi reprezentują wyższy stopień centralności. Wprowadzenie do koncepcji głębi danych znajdziemy w pracach Serflinga (2006) oraz Kosiorowskiego (2012).

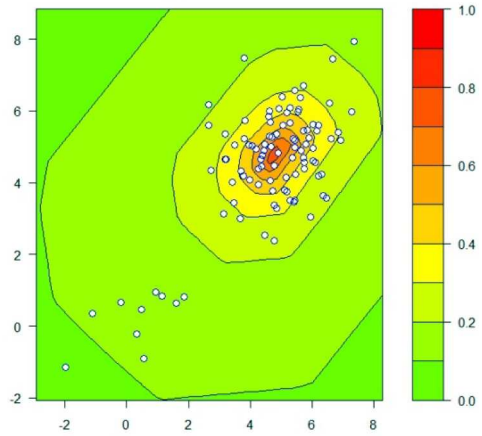
Na rysunku 3 przedstawiono empiryczną funkcję głębi projekcyjnej dla próby złożonej ze stu obserwacji wygenerowanych z dwuwymiarowego rozkładu normalnego. Rysunek 4 przedstawia empiryczną funkcję głębi projekcyjnej dla mieszaniny dwóch dwuwymiarowych rozkładów normalnych. Dwuwymiarowe mediany projekcyjne znajdują się wewnątrz najbardziej centralnych obszarów. W niniejszym artykule skupiamy naszą uwagę na szczególnym przypadku statystycznej funkcji głębi – na *głębi Studenta*. Głębi określonej dla pary: *miara położenia i miara rozrzutu,* odnoszącej się do jednowymiarowego zbioru danych.



Rysunek 3. Wykres konturowy głębi projekcyjnej z próby – 100 obserwacji z rozkładu normalnego 2d

Źródło: Obliczenia własne - pakiet środowiska R {depthproc 0.1}

Wychodząc od jednowymiarowego modelu położenia i rozrzutu Mizera i Müller 2004 wprowadzili pojęcie jednowymiarowej głębi położenia-rozrzutu oraz pokazali wybrane jej zastosowania. Ważny przypadek szczególny ich koncepcji to głębia Studenta oraz estymator maksymalnej głębi Studenta – mediana Studenta.



Rysunek 4. Wykres konturowy głębi projekcyjnej z próby – 100 obserwacji z mieszaniny rozkładów normalnych 2d

Źródło: Obliczenia własne – pakiet środowiska R {depthproc 0.1}

Mizera (2002) rozpoczyna swe rozważania od obserwacji z_i, $i = 1, ..., n$, następnie wprowadza funkcję kryterium $F_i = F_i(z_i)$ – dla danego dopasowania reprezentowanego przez θ, funkcja kryterium F_i wyraża brak dopasowania θ do konkretnego punktu z_i. Oznacza to, że θ^* odzwierciedla (pasuje do) z_i lepiej niż θ, jeżeli $F_i(\theta^*) < F_i(\theta)$.

Według propozycji Mizery ogólna głębia Tukey'a może zostać zdefiniowana jako miara dopuszczalności dopasowania zważywszy na zaobserwowane dane. *Możemy zdefiniować głębię dopasowania θ jako frakcję danych, której pominięcie sprawia, że θ staje się brakiem dopasowania, dopasowaniem, które może zostać zdominowane jednostajnie przez każde inne*. W oparciu o tę zasadę Mizera (2002) definiuje globalną głębię oraz bardziej operacyjną wersję tej głębi – głębię styczną – wynik przejścia od ogólnego kryterium optymalności do jego wersji różniczkowej. Biorąc pochodne w zagadnieniu optymalizacji z wykorzystaniem funkcji kryterium F_i Mizera definiuje głębię styczną dopasowania θ jako:

$$d(\theta) = \inf_{\mathbf{u} \neq 0} \#\{i : \mathbf{u}^t \nabla_\theta F_i(\theta) \geq 0\}, \tag{3}$$

gdzie # oznacza względną proporcję zbioru indeksów – ich liczbę podzieloną przez n, $\nabla_\theta F_i(\theta)$ oznacza gradient funkcji kryterium F w punkcie θ dla obserwacji i.

Teoretyczna innowacja Mizery i Müller (2004) polega na wykorzystaniu funkcji wiarygodności w charakterze funkcji kryterium. Niech y_i oznaczają zmienne losowe o gęstości f.

DEFINICJA 1: Głębia położenia–rozrzutu Mizery i Müller $(\mu, \sigma) \in \mathrm{R} \times [0, \infty)$ względem próby $\mathrm{Y}^n = \{y_1, ..., y_n\}$ definiowana jest jako

$$D((\mu, \sigma), \mathrm{Y}^n) = \begin{cases} \inf\limits_{\mathbf{u} \neq 0} \# \left\{ i : (u_1, u_2) \begin{pmatrix} \psi(\tau_i) \\ \chi(\tau_i) - 1 \end{pmatrix} \geq 0 \right\} & dla, \quad \sigma > 0 \\ \#\{i : y_i = \mu\} & dla, \quad \sigma = 0 \end{cases} \tag{4}$$

gdzie znak mnożenia interpretujemy jako iloczyn skalarny, τ_i jest skrótem dla oraz funkcje ψ, χ zależą od ustalonej gęstości f, $\psi(\tau) = (-\log f(\tau))' = -f'(\tau)/f(\tau)$, oraz $\chi(\tau) = \tau\psi(\tau)$.

Definicja 1, wywodząc się z metody największej wiarogodności, wprowadza rodzinę głębi zależnych od przyjętej gęstości.

DEFINICJA 2: Głębia Studenta $(\mu, \sigma) \in \mathrm{R} \times [0, \infty)$ względem rozkładu prawdopodobieństwa P na R dany jest jako

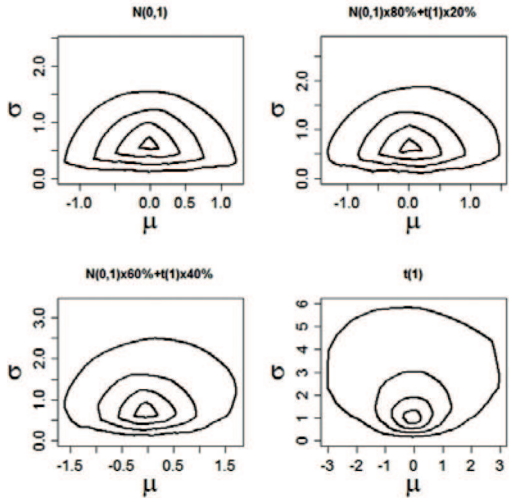
$$D(\mu, \sigma, P) = \inf_{(u_1, u_2) \neq 0} P\{y : u_1(y - \mu) + u_2((y - \mu)^2 - \sigma^2) \geq 0\} \tag{5}$$

Głębię Studenta z próby $\mathrm{Y}^n = \{y_1, ..., y_n\}$ otrzymujemy poprzez podstawienie w definicji 2 rozkładu empirycznego P_n wyznaczonego na podstawie tej próby.

Głębia położenia-rozrzutu jest ekwiwariantna względem położenia i rozrzutu, mediana Studenta ma tę samą własność. Mediana Studenta jest bardzo dobrym estymatorem centrum symetrii dla małych zbiorów danych, rzędu 30-100 obserwacji. Rysunki 5-6 przedstawiają wykresy konturowe głębi Studenta dla mieszanin standardowego rozkładu normalnego, rozkładu Studenta o jednym stopniu swobody i dla skośnego rozkładu Studenta o jednym stopniu swobody i parametrze skośności -2 – t(1,-2) (por. pakiet {skewt}programu R i literatura tamże). Rysunki te sugerują możliwość dyskryminacji pomiędzy tymi rozkładami na podstawie wykresu konturowego sporządzonego na podstawie próby. Zaznaczmy, że wykresy konturowe głębi Studenta można potraktować jako uogólnienie jednowymiarowego wykresu kwantyl – kwantyl.

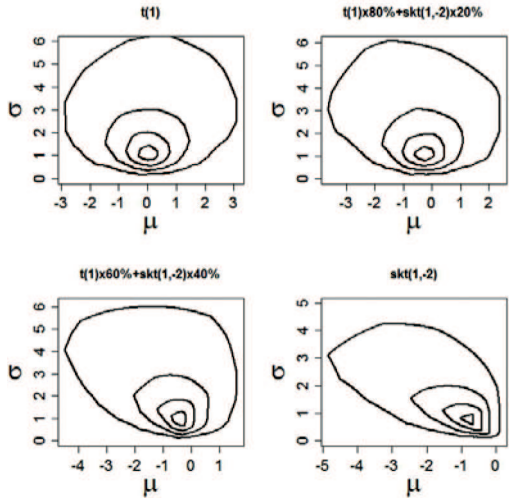
Mizera i Müller (2004), zakładając próbę losową prostą, pokazują jednostajną względem (μ, σ) zbieżność prawie na pewno estymatorów maksymalnej głębi Studenta oraz jej zadowalającą efektywność dla modelu normalnego. Pokazują też, że punkt załamania próby skończonej mediany Studenta jest w przybliżeniu równy 33% oraz mediana Studenta ma ograniczoną funkcję wpływu. Symulacje prowadzone przez autora niniejszego artykułu sugerują bardzo dobre własności mediany Studenta w sytuacji, gdy modele generujące dane odznaczały się heteroskedastycznością, nieliniową zależność pomiędzy obserwacjami oraz skośnością rozkładu. *Należy podkreślić, że w przeciwieństwie do estymatorów największej wiarygodności oraz ich uogólnień w postaci M-estymatorów Hubera – przy estymacji za pomocą maksymalnej głębi położenia – rozrzutu nie korzystamy wprost z założenia niezależności obserwacji w próbie. Korzystamy jedynie z „rankingu" dopasowania zważywszy na obserwowany zbiór danych. To*

istotna cecha estymatora w kontekście jego zastosowania do analizy strumienia danych (gdzie mamy do czynienia z zależnością obserwacji w czasie). Fakty te skłaniają autora do wykorzystania mediany Studenta w analizie strumienia danych ekonomicznych. W dalszej części wykorzystujemy algorytm {lsdepth}zaproponowany przez Ch. Müller dla jednoczesnego obliczania konturów głębi Studenta oraz mediany Studenta.



Rysunek 5. Wykresy konturowe głębi studenta dla mieszanin N(0,1) i t(1)

Źródło: Obliczenia własne, lsdepth.



Rysunek 6. Wykresy konturowe głębi studenta dla mieszanin N(0,1) i t(1,-2)

Źródło: Obliczenia własne, lsdepth.

4. PROPOZYCJE

Praktyka wymaga, aby odporna analiza strumieni danych (estymacja, predykcja i podejmowanie decyzji) prowadzona była w tempie odpowiadającym napływowi nowych danych, pojawianiu się istotnych merytorycznie zdarzeń. Taki postulat eliminuje wiele dobrych procedur rozważanych w statystyce odpornej. Napływ nowych informacji powinien poprawiać precyzję takiej analizy. Znaczna część prezentowanych w literaturze podejść do analizy strumieni danych da się zakwalifikować do jednej z dwóch kategorii: *analizy wsadowej* (ang. batch-incremental) oraz *analizy odtworzeniowej* (ang. regenerative approach). W przypadku analizy wsadowej analizujemy strumień partiami – wykorzystujemy uaktualniany model predykcyjny do momentu, gdy nie wykryjemy istotnej zmiany (trendu). W przypadku analizy odtworzeniowej tworzymy nowy model predykcyjny z każdego nowego okna (por. Aggerwal, 2007).

Bardzo popularną techniką analizy strumieni danych jest monitorowanie okna uczącego o ustalonej wielkości, zazwyczaj wyznaczonej wcześniej a priori przez użytkownika. Zwróćmy uwagę na dylemat przed jakim stoi użytkownik w takiej sytuacji: *wybrać krótkie okno tak aby to okno odpowiadało rozkładowi wyznaczonemu na podstawie obecnego stanu strumienia czy też wybrać większe okno, tak aby model był niejako bardziej reprezentatywny w okresach stabilności*. Warto zaznaczyć, że wykorzystując pewne klasyczne algorytmy zazwyczaj zakładamy, że okno uczące (przeszłość) i okno testowe (przyszłość) pochodzą z tego samego rozkładu. W kontekście analizy strumienia danych ekonomicznych rozkład z zasady zmienia się w czasie. Pojawia się pytanie, jeżeli użytkowany model wydaje się być niewłaściwy, to czy powinniśmy go modyfikować czy też odrzucić całkowicie. Zmiany modelu mogą pojawiać się stopniowo bądź nagle, dokładny punkt zmiany może być niewykrywalny.

W niniejszej pracy skupiamy naszą uwagę na statystycznych funkcjach głębi. Procedury indukowane przez głębie wykazują bardzo dobre własności w zakresie odporności, efektywności i łatwości interpretacyjnej. Stosowane procedury reprezentują tendencję wyrażoną przez większość obserwacji. Nasz sposób decydowania odznacza się konserwatyzmem jednakże zabezpieczamy się przed wpływem obserwacji odstających na nasze decyzje.

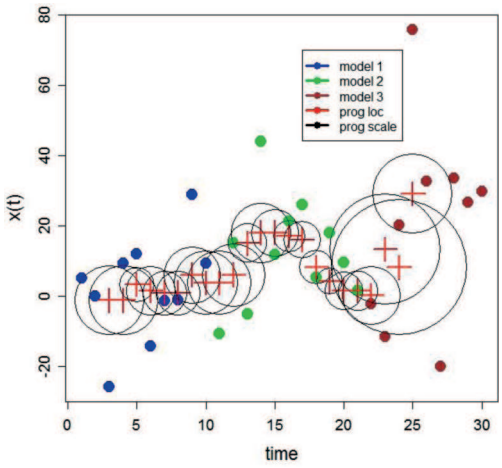
Oznaczmy przez $(\hat{\mu}_{i,n}; \hat{\sigma}_{i,n})$ medianę Studenta obliczoną na podstawie okna $W_{i,n}$. Niech $W_{Q(k),n}$ oznacza próbę wygenerowaną z $k-$tego reżimu $Q(k)$ modelu (2) o długości n.

PROPOZYCJA 1: W celu rozwiązania pierwszego problemu proponujemy przyjąć

$$\hat{x}_{t+1} = \hat{\mu}_{i,n}, i = 1, 2, ... \tag{6}$$

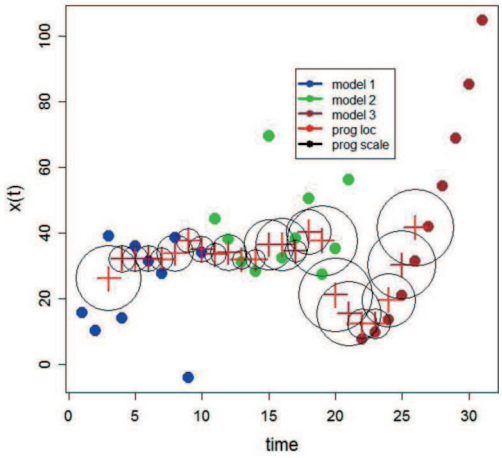
Jako przewidywanie następnej wartości strumienia danych bierzemy współrzędną położenia mediany Studena policzonej dla okna $W_{i,n}$. Jako „błąd standardowy" takiego przewidywania proponujemy wziąć $\hat{\sigma}_{i,n}$, tzn. współrzędną rozrzutu mediany Studenta policzonej dla okna $W_{i,n}$ (zobacz rys. 7-8).

PROPOZYCJA 2: W celu rozwiązania drugiego problem proponujemy ustalić próby referencyjne $W_{Q(1),n}, ..., W_{Q(k),n}$ wygenerowane z reżimów $Q(1), ..., Q(k)$ roz-



Rysunek 7. Ilustracja zastosowania pierwszej propozycji – heteroskedastyczność

Źródło: Obliczenia własne, lsdepth.



Rysunek 8. Ilustracja zastosowania pierwszej propozycji – lokalny liniowy trend

Źródło: Obliczenia własne, lsdepth.

patrywanego modelu (2), bądź wyznaczonych w związku z pewnym celem merytorycznym. Aby wykryć zmiany w bezwarunkowej wartości oczekiwanej procesu oraz w poziomie zmienności procesu, proponujemy monitorować zachowanie się następujących statystyk:

$$D_1 = \frac{\hat{\mu}_{i,n} - \hat{\mu}_{Q(j),n}}{\hat{\sigma}_{Q(j),n}}, \quad j = 1, ..., k, \quad i = 1, 2, ... \tag{7}$$

dla wykrycia zmian *poziomu charakterystyki położenia*

$$D_2 = \frac{min(\hat{\sigma}_{i,n}, \hat{\sigma}_{Q(j),n})}{max(\hat{\sigma}_{i,n}, \hat{\sigma}_{Q(j),n})}, \quad j = 1, ..., k, \quad i = 1, 2, ... \tag{8}$$

dla wykrycia *zmian w zmienności procesu*.

Dla wykrycia *zmian skośności procesu* proponujemy porównać wykres konturowy głębi Studenta z wybranymi wykresami referencyjnymi związanymi z interesującymi nas zagadnieniami.

PROPOZYCJA 3: W celu rozwiązania trzeciego problem proponujemy monitorować relacje pomiędzy współrzędnymi ruchomej mediany Studenta $(\hat{\mu}_{i,n}; \hat{\sigma}_{i,n})$ rozważanej dla okien różnej długości i porównywać te wykresy z wykresami referencyjnymi sporządzonymi dla danych wygenerowanych ze znanych modeli, bądź modeli ważnych ze względów merytorycznych

$$\hat{\sigma}_{i,n} vs. \hat{\sigma}_{i-l,n}, \quad l = 2, ..., k, \quad i = 1, 2, ... \tag{9}$$

$$\hat{\mu}_{i,n} vs. \hat{\sigma}_{i-l,n}, \quad l = 2, ..., k, \quad i = 1,2,... \tag{10}$$

Zauważmy, że stosując okna różnej długości możemy wykrywać zmiany następujące w różnych skalach czasowych (sekundy, godziny, dni), możemy jednocześnie wykorzystywać układ okien sporządzanych dla różnych skal czasowych. Istnieją, co najmniej dwa typy okien wykorzystywanych a analizie strumieni danych. W modelu *z dołączonymi oknami* (ang. adjacent windows) monitorujemy różnicę pomiędzy dwoma ruchomymi $W_{t,n}$ oraz $W_{t-n,n}$, gdzie t oznacza aktualny czas. W modelu *z ustalonym oknem* mierzymy różnice pomiędzy ustalonym oknem W_n i ruchomym oknem $W_{t,n}$. Pierwszy model lepiej wychwytuje „*intensywność zmian*" w danej chwili, podczas gdy drugi model jest lepszy do wykrycia stopniowych zmian, które mogą kumulować się w czasie. W praktyce zalecamy stosowanie co najmniej dwóch typów okien (dwóch częstości pobierania obserwacji).

5. WŁASNOŚCI PROPOZYCJI

W celu wykazania statystycznych własności propozycji w przypadku małej i umiarkowanej wielkości próby wykonano szereg symulacji w tym m. in. z następujących modeli CHARME o dwóch reżimach:

Posługując się symulatorem wchodzącym w skład pakietu {fGarch}środowiska R autorstwa Diethelma Wuertza i Rmetrics Core Team wykorzystano powszechnie stosowane w ekonometrii modele AR(1)-GARCH(1,1) o specyfikacji $X_t = \mu + \theta X_{t-1} + \varepsilon_t$ dla członu AR oraz $Z_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = c_0 + \alpha Z_{t-1}^2 + \beta \sigma_{t-1}^2$ dla członu GARCH i przy standardowych założeniach odnoście wartości parametrów (por. Fan, Yao, 2005).

MODEL 1: PIERWSZY REŻIM: składający się z dwóch modeli AR(1)-GARCH(1,1) z lokalnym liniowym trendem stochastycznym, pierwszy z parametrami AR($\mu = 5, \theta = 0.5$), GARCH($c_0 = 10^{-6}, \alpha = 0.1, \beta = 0.75$) i rozkładem warunkowym t-Studenta

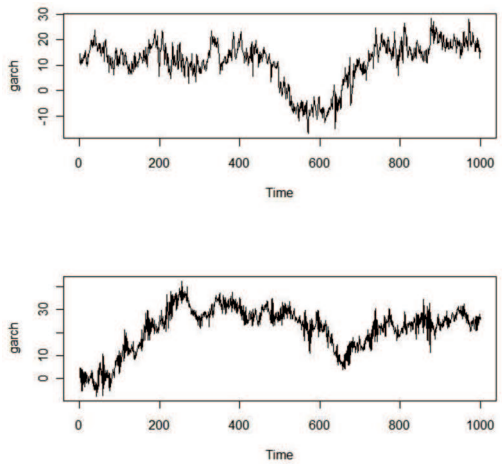
o czterech stopniach swobody; DRUGI REŻIM z parametrami AR($\mu = -5, \theta = -0.5$), GARCH($c_0 = 10^{-6}$,$\alpha = 0.6, \beta = 0.1$) i rozkładem warunkowym t-Studenta o czterech stopniach swobody (por. rys. 9-10).

MODEL2: PIERWSZY REŻIM: składający się z dwóch modeli AR(1)-GARCH(1,1) z lokalnym kwadratowym trendem stochastycznym, pierwszy z parametrami AR($\mu = 5, \theta = 0.5$), GARCH($c_0 = 10^{-6}, \alpha = 0.1, \beta = 0.75$) i rozkładem warunkowym t-Studenta o czterech stopniach swobody; DRUGI REŻIM z parameterami AR($\mu = -0.5, \theta = -0.8$), GARCH($c_0 = 10^{-6}$,$\alpha = 0.5, \beta = 0.1$) i rozkład warunkowy t-Studenta o czterech stopniach swobody.

MODEL 3: składający się z dwóch błądzeń przypadkowych z trendem $dB = mdt + sdX$, obserwowanych w równoodległych dyskretnych chwilach tzn. $B_t - B_{t-1} = m + \varepsilon_t$, $\varepsilon_t \sim N(0, s^2)$. Pierwszy z parametrami $m = 1, s = 1$, drugi z parametrami $m = -1, s = 2$.

Zmianą reżimów rządziła macierz przejścia $\boldsymbol{P}$ o kolumnach postaci $(0{,}99;0{,}01)^T$ i $(0{,}03;\ 0{,}97)^T$. Rozważano strumienie zawierające do 5% obserwacji odstających oraz strumienie bez jednostek odstających.

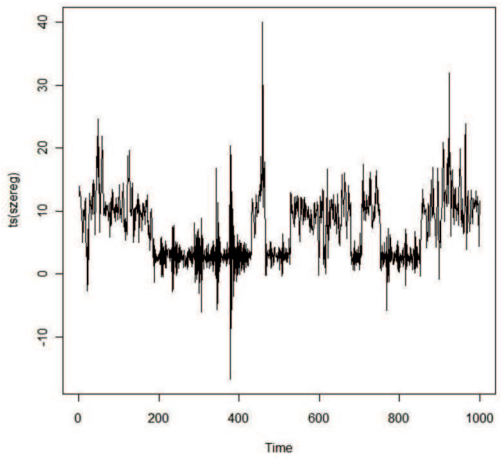
Dla każdego modelu generowano po 500 trajektorii składających się z 1000 obserwacji. Wykorzystując pierwszą propozycję obliczano jednookresową predykcję na podstawie ruchomego okna składającego się z 30 obserwacji. Przykład takiej predykcji zamieszczono na rysunkach 7-8. Za pomocą krzyży zaznaczono przewidywania na podstawie 5-elemntowego ruchomego okna, powierzchnie tarcz reprezentują błędy przewidywań. Własności propozycji porównywano przewidywaniami wykonywanymi za pomocą lokalnej[4] regresji najmniejszych kwadratów, lokalnego estymatora maksymalnej głębi regresyjnej, nieparametrycznej regresji Nadaraya-Watsona, algorytmu



Rysunek 9. Składowe modelu CHARME wykorzystywanego w symulacjach

Źródło: Obliczenia własne, {fGarch}.

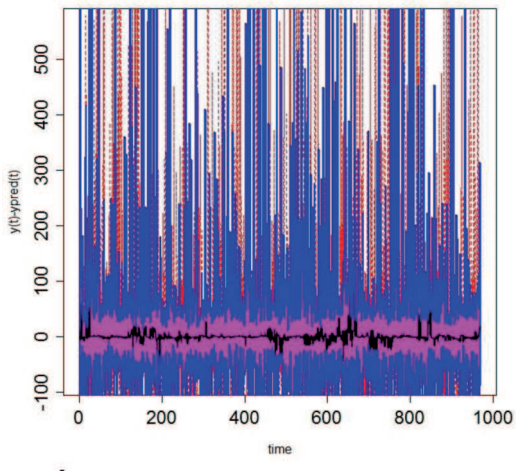
[4] Przez „lokalny" rozumiemy obliczany dla każdego okna.



Rysunek 10. Przykładowa trajektoria modelu CHARME o dwóch reżimach

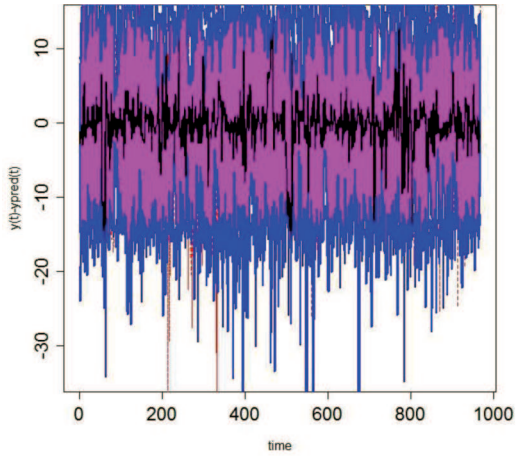
Źródło: Obliczenia własne, {fGarch}.

LOESS (najlepszej alternatywy). W tabelach 1-3 zawarto wyniki naszych symulacji w porównaniu z algorytmem LOESS. Rysunki 11-12 przedstawiają funkcjonalne wykresy ramka wąsy naszych symulacji (por. Ramsay i in., 2010). Obliczano także średniokwadratowy pierwiastek błędu prognozy (RMSEF), maksymalną z pięciuset średnich liczonych dla każdego z 1000 rozpatrywanych w symulacjach chwil $\max_i(\bar{x}_j)$ i średnią z pięciuset średnich liczonych dla każdego z 1000 rozpatrywanych w symulacjach chwil $\textit{średnia}_i(\bar{x}_j)$. Wyniki przeprowadzonych symulacji (tab. 1-3) przemawiają na korzyść propozycji, zwłaszcza w sytuacji obecności obserwacji odstających.



Rysunek 11. Funkcjonalny rysunek ramka-wąsy – wyniki symulacji dla modelu 1 i algorytmu LOESS i 5% AO

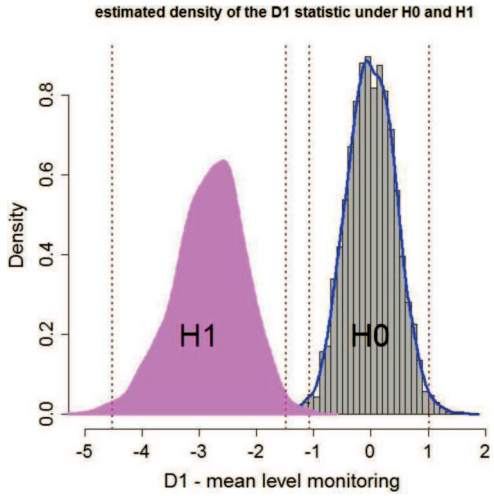
Źródło: Obliczenia własne, {fda}.



Rysunek 12. Funkcjonalny rysunek ramka-wąsy – wyniki symulacji dla modelu 1, pierwszej propozycji i 5% AO

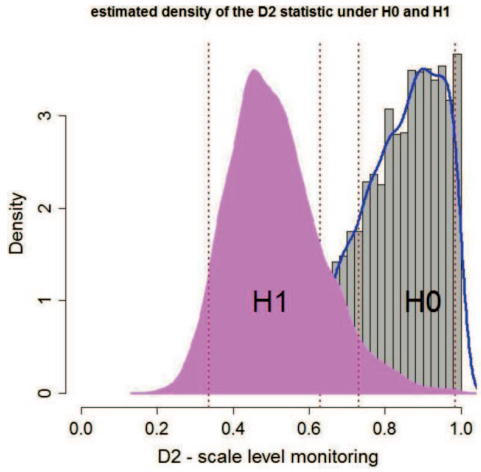
Źródło: Obliczenia własne, {fda}.

W celu sprawdzenia własności naszej drugiej propozycji – za pomocą symulacji oszacowano rozkład statystyk D1 i D2 pod warunkiem hipotezy zerowej głoszącej, że nie następuje zmiana charakterystyki położenia strumienia (w przypadku statystyki D1) oraz nie następuje zmiana charakterystyki rozrzutu strumienia (w przypadku statystyki D2) i kilku hipotez alternatywnych (następuje zmiana charakterystyki położenia albo charakterystyki rozrzutu strumienia) przy założeniu wyszczególnionych powyżej modeli 1, 2, i 3.



Rysunek 13. Propozycja druga – rozkład statystyki przy prawdziwości H_0 lub H_1, odpowiednio – położenie

Źródło: Obliczenia własne, {lsdepth}.



Rysunek 14. Propozycja druga – rozkład statystyki przy prawdziwości H_0 lub H_1 odpowiednio – rozrzut

Źródło: Obliczenia własne, {lsdepth}.

Rysunki 13-14 pokazują bardzo dobre własności naszych propozycji co do wykrywania zmian położenia centrum strumienia oraz zmian jego rozrzutu.

Tabela 1.

Wyniki symulacji własności pierwszej propozycji dla modelu 1

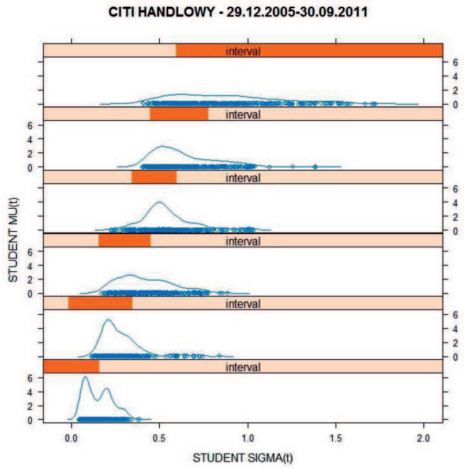
MODEL 1	**RMS**	$\max_i(\bar{x}_j)$	$średnia_i(\bar{x}_j)$
LOESS	4,223	0,591	0,105
LOESS+5\%	6507	34,067	0,204
STUDENT MED	36,158	1,77	0,554
STUDENT MED+5\% AO	2573	0,755	-0,032

Rysunki 15-16 przedstawiają zastosowanie trzeciej propozycji do empirycznego szeregu cen akcji spółki Citi Handlowy notowanych na GPW pomiędzy 29.12.2005 a 30.09.2011 roku. Łatwo możemy zauważyć zależność pomiędzy poziomem położenia centrum procesu (zysk) a zmiennością procesu (ryzyko). W kontekście analizy skośności procesu zalecamy stosowanie wykresu konturowego głębi Studenta.

Tabela 2.

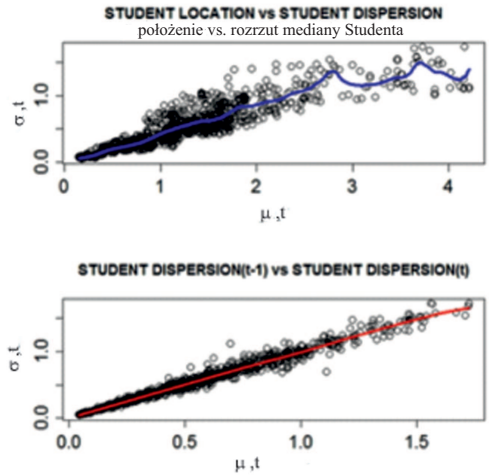
Wyniki symulacji własności pierwszej propozycji dla modelu 2

MODEL 2	**RMS**	$\max_i(\bar{x}_j)$	$średnia_i(\bar{x}_j)$
LOESS	7,95	0,52	0,027
LOESS+5\%	3991	17,84	-0,12
STUDENT MED	20,47	0,83	0,12
STUDENT MED+5\% AO	1166,35	5,85	0,32



Rysunek 15. Przykład zastosowania trzeciej propozycji

Źródło: Obliczenia własne, {lsdepth}.



Rysunek 16. Przykład zastosowania trzeciej propozycji

Źródło: Obliczenia własne, {lsdepth.}

Tabela 3.

Wyniki symulacji własności pierwszej propozycji dla modelu 2

MODEL 3	**RMS**	$\max_i(\bar{x}_j)$	$średnia_i(\bar{x}_j)$
LOESS	1,389	0,203	-0,016
LOESS+5\%	108528	230,2	-0,473
STUDENT MED	15,775	0,693	-0,227
STUDENT MED+5\% AO	14,809	0,24	-0,222

6. PODSUMOWANIE

Warto podkreślić, że w przypadku ekonomicznych strumieni danych z racji tego, że dane muszą być przetwarzane na bieżąco oraz ich napływ nie ma końca – typowy sposób analizy danych statystycznych nie ma zastosowania. Nie możemy takich danych analizować za pomocą znanych procedur klasycznej statystyki wywodzących się z postulatów Fishera z lat dwudziestych ubiegłego wieku (por. Huber, 2011). Nie mamy bowiem do czynienia z dobrze zdefiniowanym eksperymentem, z danymi generowanymi przez precyzyjnie zdefiniowany model. Strumień danych niesie sygnałpojawiający się w losowych chwilach. Dodatkowo strumienie danych generowane są przez procesy niestacjonarne o zmiennym typie niestacjonarności.

W pracy przedstawiono trzy propozycje narzędzi przeznaczonych do odpornej analizy strumieni danych mogących zawierać obserwacje odstające. Badania symulacyjne wskazują na dobre własności statystyczne propozycji. Zdaniem autora odporna analiza strumieni danych ekonomicznych może przyczynić się do lepszego rozumienia zachowań uczestników rynku. Przedstawione w pracy propozycje stanowią punkt wyjścia dla dalszych studiów zagadnień analizy strumieni danych[5].

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] Aggerwal Ch.C. (ed.), (2007), *Data Streams – Models and Algorithms*, Springer, New York.
[2] Ait-Sahalia Y., Jacod J., Li J., (2012), Testing for jumps in noisy high frequency data, *Journal of Econometrics*, 168, 207-222.
[3] Bocian, Kosiorowski, Węgrzynkiewicz, Zawadzki (2012), pakiet środowiska R {depthproc 1.0} https://r-forge.r-project.org/projects/depthproc/.
[4] Das T., Krishnan S., Venkatasubramanian S., Yi K., (2006), An Information-Theoretic Approach to Detecting Changes in Multi-Dimensional Data Streams. Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications (Interface '06)}, Pasadena, CA.
[5] Fan, J. Yao, Q. (2005), *Nonlinear time series: nonparametric and parametric methods*, Springer, New York.
[6] Genton M. G., Lucas A. (2003), Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations, *Journal of the Royal Statistical Society Series* B 65(1), 81-84.
[7] Hall, P., Rodney, C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94, (445), 154-163.
[8] Hastie T., Tibshiriani R., Friedman J., (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer.
[9] Hahsler M., Dunhamr H. M., (2010), EMM: Extensible Markov Model for Data Stream Clustering in R, *Journal of Statistical Software*, 35(5), 2-31.
[10] Huber P., Ronchettii E. M., (2009), Robust Statistics. John Wiley & Sons. New York.
[11] Huber P., (2011) *Data Analysis: What Can Be Learned From the Past 50 Years*, John Wiley & Sons. New York.

[5] Autor uprzejmie dziękuje za szereg sugestii anonimowych Recenzentów, które w znaczącym stopniu poprawiły jakość niniejszego artykułu.

[12] Jacod J., Shiryaev A.N., 2003, *Limit Theorems for Stochastic Processes*, second ed., Springer-Verlag, New York.
[13] Kong L., Zuo Y., (2010), Smooth Depth Contours Characterize the Underlying Distribution, *Journal of Multivariate Analysis* 101, 2222-2226.
[14] Kosiorowski D., (2010), Depth Based Procedures for Estimation ARMA and GARCH Models, Y. Lechevallier, G. Saporta (ed.) Proceedings of COMPSTAT'2010 19th International Conference on Computational Statistics, Physica–Verlag, 1207-1214.
[15] Kosiorowski D., (2012), *Statystyczne funkcje głębi w odpornej analizie ekonomicznej*, Wydawnictwo UEK w Krakowie, Kraków.
[16] Kosiorowski D., (2012), Student depth in robust economic data stream analysis, Colubi A.(Ed.) Proceedings COMPSTAT'2012, The International Statistical Institute/International Association for Statistical Computing.
[17] Kosiorowski D., Snarska M., (2012), Robust monitoring of a multivariate data stream, LINSTAT 2012, artykuł złożony do *Communications in Statistics.*
[18] Maronna R.A., Martin R.D., Yohai V.J., (2006), *Robust Statistics - Theory and Methods.* Chichester: John Wiley & Sons Ltd.
[19] Mizera I., (2002), On Depth and Depth Poins: a Calculus. *The Annals of Statistics* (30), 1681-1736.
[20] Mizera I., C.H. Müller (2004), Location-scale Depth (with discussion), *Journal of the American Statistical Association* 99, 949-966.
[21] Ramsay J.O., Hooker G., Graves S., (2010), *Functional Data Analysis with R and Matlab*, Springer, New York.
[22] Shalizi C.R., Kontorovich A., (2007), *Almost None of the Theory of Stochastic Processes A Course on Random Processes, for Students of Measure-Theoretic Probability, with a View to Applications in Dynamics and Statistics*, http://www.stat.cmu.edu/˜cshalizi/almost-none/
[23] Serfling R., (2006). Depth Functions in Nonparametric Multivariate Inference, In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): Series in Discrete Mathematics and Theoretical Computer Science, AMS, vol. 72, 1-15.
[24] Stockis J.-P., Franke J., Kamgaing J.T., (2010). On geometric ergodicity of CHARME models, *Journal of the Time Series Analysis* 31, 141-152.
[25] Szewczyk W., (2010), Streaming data, *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(1), (on-line journal).

GŁĘBIA POŁOŻENIA-ROZRZUTU W STRUMIENIOWEJ ANALIZIE DANYCH EKONOMICZNYCH

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

Z praktycznego punktu widzenia priorytetowym celem analizy ekonomicznego szeregu czasowego jest uzyskanie wglądu na podstawie stale uaktualnianej próby umiarkowanej długości w krótkookresowe właściwości probabilistyczne procesu generującego dane. Na podstawie takiej w ogólności nieprecyzyjnej wiedzy dokonywanych jest szereg decyzji ekonomicznych oraz prognoz. W praktyce bardzo ważną kwestią jest odpowiedź na pytanie co mówi nam większość danych o przyszłym zachowaniu większości uczestników pewnego rynku. Szczególnie trudno odpowiedzieć na takie pytanie w przypadku wielkich zbiorów danych generowanych przez zmieniający się wielowymiarowy model. W artykule prezentujemy wybrane zastosowania procedur indukowanych przez głębię położenia-rozrzutu Mizery i Müller w odpornej analizie strumienia danych ekonomicznych.

Słowa kluczowe: strumień danych, statystyczna funkcja głębi, wielowymiarowa mediana

LOCATION-SCALE DEPTH IN ECONOMIC DATA STREAM ANALYSIS

A b s t r a c t

In this paper we study the properties of the location-scale depth procedures introduced by Mizera & Müller and look into the probabilistic information of the underlying time series model carried by them. We focus our attention on short term multivariate quantile based description of the possible time series model. We study robustness and utility of such the description in a decision making process. In particular we investigate properties of the moving Student median (two dimensional Tukey median in a location–scale problem).

Key words: Data Stream, Statistical depth function, multivariate median

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

JAN W. OWSIŃSKI

ON THE OPTIMAL DIVISION OF AN EMPIRICAL DISTRIBUTION (AND SOME RELATED PROBLEMS)[1]

1. THE BACKGROUND

It is quite frequent that a set of entities observed and analysed is divided up into subsets. This is done so as to treat the entities within the subsets as similar, or even "equivalent" in terms of the feature(s) considered, and to be able to correctly distinguish between the subsets, i.e. entities put in different subsets being sufficiently dissimilar with respect to these features.

There may be various reasons for such divisions, examples being: differentiation of policy measures with respect to regions featuring various levels of GDP per capita, membership fees for countries featuring various economic power, social policy measures for households characterised by different levels of poverty, etc. There may also simply be the reasons of pragmatism in addressing intensities of respective phenomena ("highly developed", "developed",..., countries, or "marginally poor", "poor", "very poor", etc. households). These reasons are usually closely coupled with a cognitive interest, in that we would like to make the divisions possibly "objective" and that we would like to know whether there exist any mechanisms that drive the divisions – and what they are, if they do exist.

This problem is very closely associated with what is called "categorisation" – i.e. representation of a set of values, taken by some variable, characterising the entities analysed, by a limited number of "categorical" values, to be then used instead of the original (measured) values. Another aspect is associated with the possibility of representing these categories through some natural language expressions.

In very general terms one might classify the approaches to determination of the distribution division here meant in the following manner:

– *statistical*, most often based on some quantile-type criteria, with the respective levels selected on the basis of either tests, or, more often, expert-based assessments;
– *political*, when expert-based assessment is directly used to set the dividing lines, over which a political decision process takes place;

[1] The work, whose results are reported here, was partly financed from the projects: TIROLS, N N516 195237, financed by the Polish Ministry of Science and Higher Education, and the modeling project in the framework of the "Development Trends of Masovia", MBPR/W/TRM-0712-08-2011, co-financed from the EU funds.

– *substantive*, when analysis is carried out of the phenomenon, giving rise to the observed values, and the dividing lines are determined as meaningful from the point of view of the phenomenon itself.

In reality, of course, the three often appear in conjunction, with varying emphasis. We shall illustrate these approaches and the related questions with the academic example of Fig. 1[2]. This figure shows monthly remuneration in Polish zlotys in a small company.

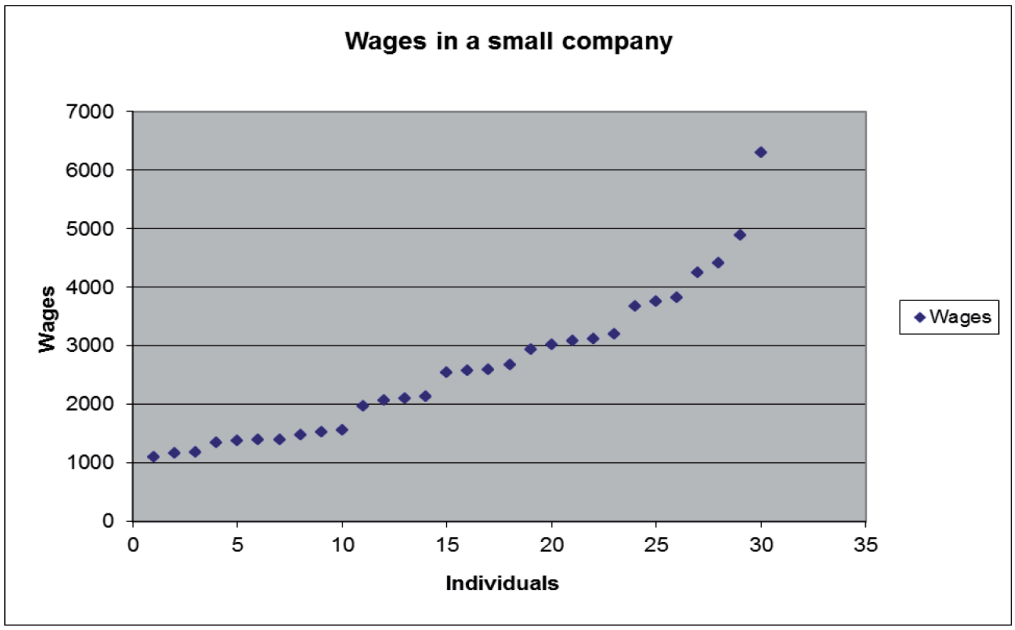


Figure 1. An academic example of a distribution: wages in a small company

If we wanted to split the distribution like the one of Fig. 1 into few meaningful categories, the premises could be as follows:

– statistical (or, actually, in this case, "statistical"): for instance, given that the descriptive statistics for this case are min = 1 100 PLN per month, max = 6 300 PLN, mean = 2621.67 PLN, and standard deviation = 1273.64 PLN, it would seem reasonable to divide the distribution into "low wages" (up to mean – standard deviation), "middle wages" (within the standard deviation away from the mean), and "high wages" (above mean + standard deviation); the 30 observations would then be split up into three subsets, composed of, respectively, 4, 20 and 6 observations; otherwise, this could be done on the basis of quantiles; the problem with this division is that it takes into account the shape of the distribution to a very limited degree – actually, it cuts through one of the distinct plateaus, visible in Fig. 1, while associating in one category a lot of different wage levels;

[2] All figures in his text result from the research of the author.

– political: a division could be based on the "minimum wage" definition, along with other criteria, linked, e.g., with taxation, social security etc.; such criteria do not, of course, account for the shape of the distribution;

– substantive: wages may depend upon certain regulations, involving education, length of work experience, position in the company etc.; this division would seem to correspond the best to the respective reality, but, especially in this particular case, the multiplicity of criteria and the corresponding partial subdivisions would yield either too narrow categories or would require some additional aggregating operator to be applied.

All in all – we would like to have, therefore, a methodology that could lead to reasonable division into categories on the basis of actual reality in the form of the given empirical distribution.

2. NOTATION AND INITIAL FORMULATION

We are given a univariate empirical distribution of a quantity x taking values from $\boldsymbol{R}_+$. The distribution consists of n observations, indexed i, $i = 1,\ldots,n$. This set of indices shall be denoted I. Without any loss to any sort of reasoning, we assume that the values taken by x in this distribution, denoted x_i, are ordered in a non-decreasing sequence, i.e. $x_{i+1} \geq x_i$ for all i.

Next, assume we consider, instead of the sequence $\{x_i\}_I$, the corresponding sequence, formed by the corresponding cumulative distribution, i.e. the values z_i defined as $z_i = \sum_{i'=1,\ldots,i} x_i$. We deal, therefore, with the sequence $\{z_i\}_I$ that is increasing, and, moreover, also convex. This means that a straight line, joining any two points of the sequence $\{z_i\}_I$, say z_i and $z_{i+\Delta I}$, where Δi is any integer number contained in the interval $[2,n\text{-}i]$, has values above those of the corresponding z_i, i.e. $z_{i+1},\ldots,z_{i+\Delta i-1}$.

Having such data, very much like a Lorenz curve, we would like to construct a sort of piece-wise linear approximation that is exactly based upon the segments of straight lines, which is in some sense "optimal". Namely, we would like to determine a set of segments such that the resulting error (sum of absolute differences between the actual values of z_i and the corresponding values of the approximating function) is possibly low, while the number of segments we distinguish is also kept reasonably low.

If we could obtain such a "more objective" division, based only on the shape of the sequence of z_i, without predefining the number of classes, then the assignment of labels, i.e. class names, such as "highly developed" etc., would be done a posteriori on the basis of characteristics of the classes obtained, rather than as the result of an arbitrary and largely subjective perspective on how the classes "should" be defined or named.

The present analysis was motivated by exactly such a proposal, forwarded by Nielsen, (Nielsen, 2011), with respect to the indicators of country development levels. Another domain of interest with similar features is the one of distribution of wealth within a society, with the i's corresponding to the successive, somehow defined, wealth

classes. The proposal of Nielsen (2011), is analysed here and significantly extended on the basis of the "bi-partial" approach, as exemplified, for instance, in the papers by the present author, Owsiński (1990, 2011, 2012).

3. SOME PROPERTIES OF THE PROBLEM AND AN OBJECTIVE FUNCTION

A way to solve the problem as formulated before might consist in minimising the error for subsequent numbers of classes (segments) and finding the "most appropriate" solution in terms of the error value and the number of classes. The weak point of such a procedure consists in the necessity of finding a "proper" trade-off between the differences in error and the count of the number of segments.

Namely, obviously, the error for the optimum approximation decreases monotonically as a function of the number of classes. Thus, as the number of classes increases by one, the (optimum, i.e. minimum) error decreases by a certain number, in general – different for each addition of one class. Hence, it is easy to see that the essential problem persists – namely that of indicating the "most appropriate" number of classes.

It would therefore appear as "natural" to look for a different form of the objective function we try to optimise (minimise), rather than, in very general terms, the previously discussed, "total error + number of classes".

For this purpose we shall introduce some further notation. Thus, let us denote with q the index of the subsequent classes, ranging from 1 to p, i.e. p is the overall number of classes distinguished. Denote with A_q the set of indices i of observations x_i (and therefore also z_i), classified in class q. Note that if, as proposed before, the piece-wise linear approximation is based on straight line segments, defined by values of z_i, i.e. the lines always go through two points z_i, then the sets A_q may either overlap on one observation (the maximum index in A_q being the same as the minimum index in A_{q+1}) or be disjoint (the maximum index in A_q being equal the minimum index in A_{q+1} minus 1[3]). We shall denote by $z^{q\,\min}$ and $z^{q\,\max}$, respectively, the minimum and maximum values of cumulative distribution, corresponding to the set A_q. These values, in turn, correspond to the indices $i^{q\,\min}$ and $i^{q\,\max}$, respectively. Additionally, we denote with $q(i)$ the index of the class, to which i-th observation belongs. (Note that in the case the classes overlap on "border" observations, $q(i)$ are the sets of two consecutive values.) Let us denote the set of i index values, defining the partition of the sequence $1,\ldots,n$ into the subsets A_q, i.e. the sequence composed of the values $1 = i^{1\,\min}, i^{1\,\max}, i^{2\,\min}, i^{2\,\max}, i^{3\,\min}, \ldots, i^{p\,\max} = n$, by $\boldsymbol{iq}$. Of course, by specifying $\boldsymbol{iq}$, for a given sequence $\{z_i\}$, we define $\{A_q\}$ and the entire solution. When referring to the explicit set of subsets $\{A_q\}$ we may also use the notation P, for partition of the set of observations.

[3] We shall stick to the second option in what follows. This assumption is, of course, just a choice with no deeper consequences: in the examples further on we deal with both cases.

Given that we assume the piece-wise linear form of the approximation, we can write down the general expression for the q-th piece as

$$z^q(i) = a^q i + b^q, \tag{1}$$

where, in fact, we can no longer care whether i is discrete or continuous, but we observe the respective values only for natural i. This is an essential assumption for the further calculations. The values of the coefficients a^q and b^q are determined in a natural manner from the standard formulae, where we assume, formally, that each segment is composed of at least two consecutive observations, i.e. $i^{q\,\max} > i^{q\,\min}$:

$$a^q = \frac{z^{q\,\max} - z^{q\,\min}}{i^{q\,\max} - i^{q\,\min}}, \tag{2}$$

and

$$b^q = z^{q\,\min} - \frac{z^{q\,\max} - z^{q\,\min}}{i^{q\,\max} - i^{q\,\min}} i^{q\,\min}. \tag{3}$$

Note that after differentiating $z^q(i)$ as in (1) we obtain the increasing sequence of levels a^q, corresponding to classes in terms of values of x_i.

In view of the convexity of the sequence of z_i, the sequence of a^q is non-decreasing (the differences $i^{q\,\max}$-$i^{q\,\min}$ are accompanied by appropriately increasing differences $z^{q\,\max}$-$z^{q\,\min}$), while the sequence of b^q is non-increasing.

We can now formulate the "minimum approximation error" problem, with the respective objective function, denoted $C_D(\{A_q\})$, as follows:

$$\min_{\boldsymbol{iq}} \left(C_D\left(\{A_q\}\right) = \sum\nolimits_q \sum\nolimits_{i \in Aq} \left(z^q(i) - z_i\right) \right), \tag{4}$$

where minimisation is performed with respect to the sequence $\boldsymbol{iq}$, defining the approximation. We shall denote the optimum sequence, corresponding to the minimum in (4), by $\boldsymbol{iq}^*$.

Since in conditions of convexity there is just one optimum sequence $\boldsymbol{iq}^*$ for each consecutive value of p (quite in line with Nielsen, 2011), we can denote the minimum value of $C_D(\{A_q\})$ for a given p by $C_D^*(p)$. Thus, we have $C_D^*(p) \geq C_D^*(p{+}1)$. Obviously, equality can only occur, when there exist sequences of $x_i = x_{i+1} = \ldots$, so that corresponding $z_i, z_{i+1}, \ldots$, are indeed situated on a straight line. Otherwise, any increase of p leads to the decrease of $C_D^*(p)$. Actually, one could go, with the above formulation of the problem, to the extreme of $p = n$, when $C_D^*(n) = 0$, an "ideal approximation"! Each observation would then constitute a separate "class" with just one representative. Of course, formulae (2) and (3) would become obsolete (division by zero). (Note, though, that we do not deal here with the proper "approximation", as it is understood in numerical analysis, where problem formulation is altogether different.)

Obviously, when the previously mentioned sequences of $x_i = x_{i+1} = \ldots$, occur, so that the corresponding $z_i, z_{i+1}, \ldots$, are situated on a straight line, $C_D^*(p)$ shall remain

at the absolute minimum of 0 also for $p < n$, down to the value, determined by the total length of such uniform sequences.

While construction of approximating segments for such sequences is not a question, the issue that we address here is related to the possibility of finding a way to tell "how different the successive observations have to be in order to assign them to different segments (classes)".

4. CONSTRUCTION OF A BI-PARTIAL OBJECTIVE FUNCTION

If the problem were formulated as "minimise the error with as low number of segments as possible", we would face the following minimisation problem:

$$\min(C_D(\{A_q\}) + w(p)), \tag{5}$$

where $w(p)$ expresses the weight we attach to the number of segments in the potential solution. For consecutive values of p the minimum of $C_D(\{A_q\})$ would be found, and then the minimum of the function from (5) determined. Although this procedure might seem altogether cumbersome, but, since we expect not too many segments to correspond to optimum, it would still be numerically feasible. We shall not go into the technical details of such an approach, for reasons given below.

Namely, now, the essence of the problem is transferred to the determination (or knowledge) of the weight function $w(.)$. In some cases, when p has a concrete interpretation, carrying with it, for instance, some cost, while error minimisation leads to definite benefits, like in technical applications, or in operational research, then determination of $w(.)$ is quite feasible, even if still charged with difficulties. This is not, however, the case with our problem, where we look for some possibly "natural" division of the cumulative distribution, and no cost / benefit, except for the facility of use of appropriate linguistic labels ("very highly developed", "highly developed",...), is involved.

As we try to find the "natural" division of the cumulative distribution (that is – provided it exists, and the method we aim at ought to tell us whether it does), therefore, we should refer to some "counterweight", analogous to that of $w(p)$ in (5), but having the same sort of meaning and kind of measurement as $C_D(\{A_q\})$. In this way we might be able to try to define the proper p and at the same time the ***iq***, or, otherwise, the $\{A_q\} = P$.

Thus, similarly as in (5), we would like to add to $C_D(\{A_q\})$ a component that would penalize, in this case, for the division into segments that are in some way "too similar", especially in terms of subsequent a^q. In general terms the respective bi-partial objective function and the corresponding problem would look like

$$\min(C_D(\{A_q\}) + C^S(\{A_q\})), \tag{6}$$

where $C^S(\{A_q\})$ is exactly the component corresponding to the similarity between the consecutive segments, based primarily on the differences of consecutive a^q. The more detailed form of $C^S(\{A_q\})$ might be constructed as follows:

– first, the value of a kind of difference between the two consecutive segments, q-1 and q, could be measured, from the point of view of the succeeding segment, q, as, for instance,

$$z_{iq\,\min} - a^{q-1} i^{q\,\min} - b^{q-1} \tag{7}$$

expressing the difference between the actual value of z_i at the beginning of the next, q^{th} subset of observations, and the "approximation" of the same, resulting from the previous segment. This difference is always non-negative, due to convexity of $\{z_i\}$, and can be interpreted as a "distance" between the two consecutive segments in the approximation;

– now, as we wish to penalize (minimize) with $C^S(.)$ the *similarity* rather than distance (or difference), in order to convert (7) into similarity, we should subtract it from some kind of upper bound; this upper bound might be constituted by the maximum of a similar difference for a given data set, namely as defined by the biggest difference of tangents along the curve of z_i, i.e. between the beginning and the end of the curve; we shall denote the two extreme tangents by $a^{(1)}$ and $a^{(n)}$, and they are defined as:

$$a^{(1)} = z_1/i_1; \quad \text{and} \quad a^{(n)} = (z_n - z_{n-1})/(i_n - i_{n-1});^4 \tag{8}$$

in order, however, to calculate the proper difference, we must have the complete expressions for the lines, corresponding to $a^{(1)}$ and $a^{(n)}$, satisfying the conditions that allow for their use with respect to consecutive subsets A_q; we assume, namely, that all four lines involved here, corresponding to the tangents a^q, a^{q-1}, $a^{(1)}$ and $a^{(n)}$ cross at the point, defined otherwise by the crossing of the lines, corresponding to subsets A_{q-1} and A_q, i.e. at the point, denoted i^{q*} in Fig. 2; from this condition we can derive the values of b, to be used in conjunction with $a^{(1)}$ and $a^{(n)}$ (denoted, respectively, $b^{*(1)}$ and $b^{*(n)}$) in the appropriate expression, namely:

$$b^{*(1)} = b^q - (a^{(1)} - a^q)(b^{q-1} - b^q)/(a^q - a^{q-1}), \tag{9a}$$

$$b^{*(n)} = b^q - (a^{(n)} - a^q)(b^{q-1} - b^q)/(a^q - a^{q-1}). \tag{9b}$$

Now, the expression for $C^S(.)$ for a single q, following the reasoning here presented, can be written down as

$$a^{(n)} i^{q\,\min} + b^{*(n)} - (a^{(1)} i^{q\,\min} + b^{*(1)}) - (z_{iq\,\min} - a^{q-1} i^{q\,\min} - b^{q-1}), \tag{10}$$

where the second term in brackets is equivalent to the difference, given by (7), while the preceding terms define the reference for the given q. Now, then, the proposed $C^S(P)$ is the sum over q of (10). Altogether, the minimised objective function we wished to have, takes on the form:

---

[4] This notation with respect to i is meant to emphasise the generality of formulae; actually, if i are just the natural numbers, starting with $i = 1$, and ending with $i = n$, these formulae get much simpler: $a^{(1)} = z_1$; and $a^{(n)} = z_n - z_{n-1}$.

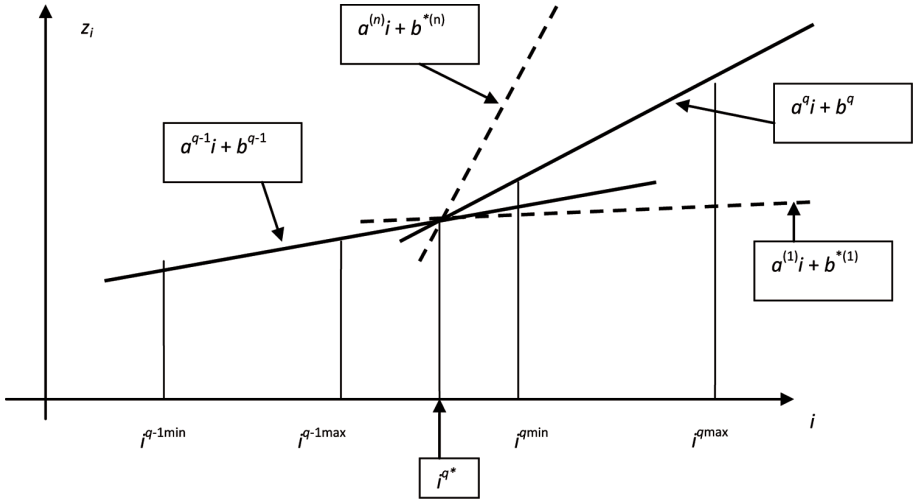


Figure 2. Schematic presentation of the way, in which parameters of $C^S(.)$ are determined

$$C_D^S(\{A_q\}) =$$
$$C_D(\{A_q\}) + C^S(\{A_q\})) =$$
$$\Sigma_q \Sigma_{i \in Aq}(a^q i + b^q - z_i) + \Sigma_q(a^{(n)} i^{q\min} + b^{*(n)} - (a^{(1)} i^{q\min} + b^{*(1)}) - (z_{iq\min} - a^{q-1} i^{q\min} - b^{q-1})). \quad (11)$$

in which we take $a^0 = 0$ (which is quite natural), and $b^0 = 0$ (which is somewhat artificial).

Let us note at this point that the objective function defined above implies for the example of Fig. 1 a solution composed of five segments, the divisions into less and more segments, minimizing $C_D(.)$, yielding higher values of $C_D^S(.)$.

At the end of this section let us also indicate that the approach, in which fit or precision of some kind of rendition or representation is counterbalanced by a sort of "cost" incurred by it, e.g. in the form of the number of parameters (like in the polynomial approximation, for instance), is quite common to statistics, especially empirical statistics, taking the concrete forms in such criteria as, say, AIC or BIC. Yet, in these criteria we deal with two quite different 'parts', each one expressing a completely different aspect of the problem at hand, while in the objective function here proposed we try to reflect possibly faithfully the actual duality of the respective problem, and to express the two aspects in the same kind of terms.

5. SOME GENERAL PROPERTIES AND POTENTIAL ALGORITHMS

Generally, the construction of the bi-partial objective function follows only the prerequisites of "global' rationality – namely that we oppose two measures that individually represent a one-sided rationality (here, in particular, error minimization),

and that together imply some sort of compromise, based on their joint minimization or maximization (see Owsiński, 2011, 2012). In this, we do not enforce neither the concrete structure (value of p), nor any weight – although weights can, of course, be applied, and even may be effectively used, as this shall be seen later on. Actually, the methodology based on the bi-partial objective function applies equally to the multidimensional distributions.

In this particular case, we constructed the bi-partial objective function out of two components, one, $C_D(\{A_q\})$, corresponding to the error, resulting from the "approximation" of the sequence of z_i with a limited number of line segments, and the second, $C^S(\{A_q\})$, corresponding to the penalty for the too small change of the angle of the line between two consecutive segments. Although we have not shown this with respect to the second component, the two components display opposite monotonicity along the number of segments, p, that is – minimum $C_D(\{A_q\})$, or $C^*_D(p)$, decreases along p, while $C^S_D(\{A_q\})$ increases (we refer here only to the sequence of $\boldsymbol{iq}$ minimizing $C_D(\{A_q\})$).

The above remark indicates one of the fundamental principles of construction of the bi-partial objective function, namely the *opposite monotonicity* of the two components.

One might indicate, in this context, that the two components in this example do not quite correspond to each other (are not "symmetric"): there is only one element per segment in $C^S(\{A_q\})$, while there are cardA_q-2 elements per segment in $C_D(\{A_q\})$, which, definitely, introduces a bias (in this realisation of the bi-partial objective function for the distribution division problem the segments obtained cannot be too big, i.e. card A_q too high). Indeed, we have provided here only an example – actually, the entire formulation of the problem, also involving the "error function", is arbitrary (we could use, e.g. the sum of squares formulation). The primary purpose of the exercise presented was to show the way the bi-partial function can be constructed and how it functions.

Concerning the optimisation algorithms, the off-the-shelf choice for a single-dimensional problem is the dynamic programming approach, similarly as in the classical categorisation problem (see, e.g., Gan, Ma and Wu, 2007), or that proposed by Thierry Gafner (1991). For the example at hand, though, we can also consider the approach by the present author, which is closely associated with the idea and the properties of the bi-partial objective function. We shall provide here only the basic precepts of this approach.

Actually, we shall illustrate the approach with the concrete form of $C^S_D(\{A_q\})$, considered here. Thus, assume we consider, instead of (11), a parameterised form:

$$C^S_D(\{A_q\}, r) = (1 - r)C_D(\{A_q\}) + rC^S(\{A_q\})) \tag{12}$$

with $r \in [0,1]$, and we look for the minimum of $C^S_D(\{A_q\},r)$ over $\boldsymbol{iq}$, i.e. $C^{S*}_D(r)$.

Then, assume we start the procedure from $r = 0$. Thus, we have

$$C^S_D(\{A_q\}, 0) = C_D(\{A_q\}), \tag{13}$$

and, of course, the "optimum" partition for this situation is the one with $p = n$, or, at most (according to our form of the "error function"), $p = \text{int}[n/2]+1$, where int[v] is the highest integer number lower than v (this fact resulting from the zeroing of elements of $C_D(\{A_q\})$ at the endpoints of each segment). Yet, we are not, in general, interested in such a trivial solution. As we increase r from 0, non-zero weight starts to be assigned to $C^S(\{A_q\})$, and in order to obtain $C_D^{S*}(r)$ for such r, it will "pay" at some definite value of r, say r^1, to join two segments, for which the difference of angle is the smallest, and hence the penalty in $C^S(\{A_q\})$ is the biggest. This value of r^1 can indeed be easily determined on the basis of the formulae here provided.

As we increase the parameter r further, we find its next value, r^2, for which merging of another pair of consecutive segments "pays" in terms of $C_D^S(\{A_q\},r)$. And so on. The proper solution is found for the last r^t that is not bigger than $^1/_2$ – i.e. for the equal weights of the two components of $C_D^S(\{A_q\},r)$. This, of course, is a sub-optimisation algorithm, as it does not guarantee reaching of the proper minimum of $C_D^S(\{A_q\})$, since the sole operations involved are the aggregations of segments. Yet, experience shows that it either actually reaches the minimum, or is very close to it.

No matter which method is used (the dynamic programming or the one proposed by the author), the basic rationale consists in forming the segments, A_q, composed of sequences of possibly similar x_i, i.e. z_i forming a curve possibly close to a straight line. An adequately pronounced "jump" over one or several consecutive i's would then correspond to a change from q to $q+1$ in the optimum solution.

6. PROBLEMS WITH REAL-LIFE EMPIRICAL DISTRIBUTIONS

The here outlined treatment of the problem of dividing the empirical distributions, though, encounters in many instances an essential hindrance in the form of the actual form of such empirical distributions. The essential question relates to two, often closely interconnected issues:

(a) a very regular – or very close to very regular – shape of the empirical distribution; examples of such situations are provided in Figs. 3 and 4, showing, first, the annual value of own revenue per capita of the more than 300 municipalities of the Polish province of Masovia, based on the data from the Central Statistical Office (GUS), ordered in decreasing sequence; although the data are just for one year, and not for a very big "sample", the regularity is, indeed, striking; then, Fig. 4 shows – for the same set of municipalities – the shares of population of definite age segment, and, again, the regularity defies any attempt of "natural" division of the distribution; these figures show also the second of the issues that ought to be considered, namely

(b) the shape of the empirical distribution that runs "counter" to the intuitive precepts we have outlined before, i.e. that a "category" ought to be formed by the values of x_i that are possibly similar; the difficulty in the case of Fig. 3 is with the upper end of the distribution – how many categories one would (intuitively) assign there? The really striking cases of distributions where similar problems appear at their

both ends were considered in Owsiński (2012) on the basis of QOL (2005) and QOL (2007).

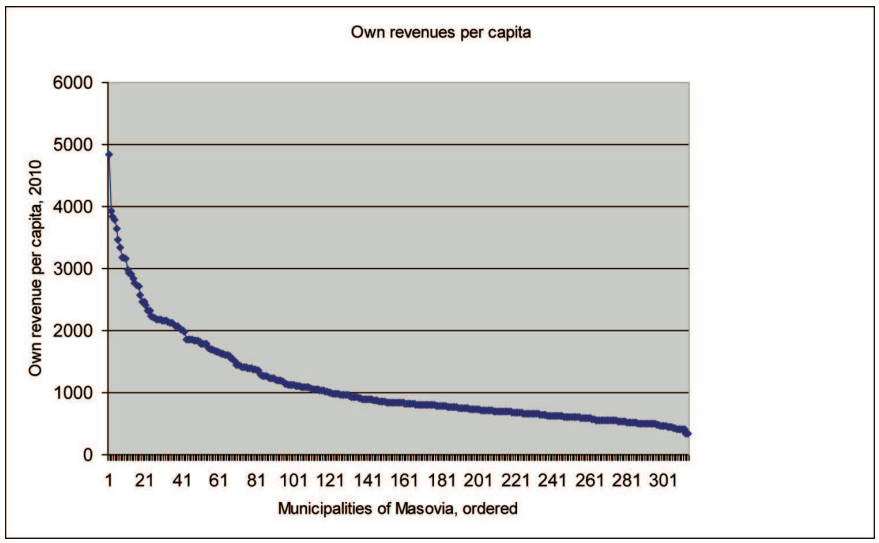


Figure 3. An example of an empirical distribution: own revenues per capita of the municipalities of Masovia in Poland in 2010 (in Polish zlotys per annum)

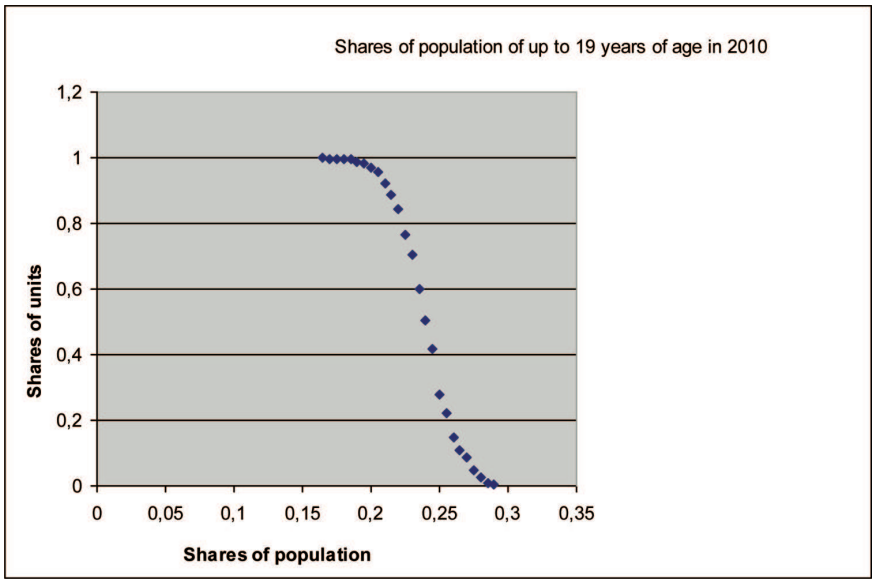


Figure 4. An example of an empirical distribution: shares of municipalities of Masovia in Poland in 2010 with definite shares of population below 19 years of age

7. WHAT CAN BE DONE TO "SPITEFUL" DISTRIBUTIONS?

With respect to such regular (and yet – characteristically shaped) distributions the following questions appear to be justified:

– for a (very) regular distribution, which could relatively easily be approximated by a certain functional shape (notwithstanding its potentially established "theoretical-statistical" foundations) is any division at all, of statistical or any other character, reasonable?
– if, however, a methodology, like the one outlined here, were applied to such a regular distribution, what should be the interpretation of the results obtained?

The first of these questions could be answered through the following procedure:

i. definitely, it must be established what functional form, and, if possible, what theoretical statistical distribution fits the given empirical distribution with a sufficient degree of exactness and/or significance;
ii. this ought to be analysed against the background of the substantive knowledge of the processes behind the empirical distribution; it is feasible that in some cases an explanation for the shape obtained could be provided, or at least an interpretation of the values, represented in the distribution, and that possibly in intuitive terms;
iii. unless the functional and/or theoretical fitted shape provides an obvious basis for a division, there is, indeed no well-founded "computational" ground for performing it;
iv. the sole basis for any potential division remains, therefore, with the substantive analysis of the respective processes; in many cases, though, given the frequent purposes of such an analysis, the risk appears of a political influence on the results (like in case of classification of regions, countries, or setting of poverty lines); in any case, though, point **ii** above preserves its validity.

Now, regarding the second question, it might also be answered through a procedure like the one outlined above, in its variant, consisting of the following steps:

i'. assume it is established what functional form, and, if possible, what theoretical statistical distribution fits the given empirical distribution with a sufficient degree of exactness and/or significance;
ii'. this form ought to be analysed in the perspective of the substantive knowledge of the processes behind the empirical distribution; possibly, in some cases an explanation for the shape obtained could be provided, or at least an interpretation of the values, represented in the distribution, and that possibly in intuitive terms;
iii'. division is performed with the methodology outlined, based on the functional form, identified under step i', meaning that the *approximation is carried out not with respect to the piece-wise linear function, like in the illustrative case here considered, but with respect to the concrete form identified, with its parameters shifting between the segments of the division*;[5]

[5] The very first step might consist in approximation with the simplest quadratic polynomial segments, forming a continuous approximating function. Thereby, some of the segments – representing, e.g. the middle part of the distribution, could be reduced to linear ones.

iv'. the segments, forming the division obtained, ought to be analysed against the background of substantive knowledge, in analogy to (or: in the framework of) the analysis, carried out in point ii' above; with this, it ought to be remembered that the segments obtained are founded directly on the shape of distribution.

In the case of distribution, shown in Fig. 3, the results obtained indicated the optimum number of segments, p, as equal between 8 and 15, with relatively small differences of value of $C_D^S(.)$ for these p. This confirms the hesitation, concerning such forms of distributions, as concerns application of the dividing approaches. Yet, analysis of the segments obtained for this series of similarly good solutions, ought to be quite educative also in deeper substantive sense. In particular, one might try to assess the "quality" of the empirical distribution against both the substantive precepts and the degree of fitting to the functional shape identified (like in the case of the QoL rankings, analysed in Owsiński, 2012).

Instytut Badań Systemowych PAN

LITERATURA

[1] Gafner Th., (1991), *Mathematical programming approach to classification*, Ph. D. dissertation, Institute of Statistics, Faculty of Economics and Business, University of Neuchatel.

[2] Gan G., Ma Ch., Wu J., (2007), *Data Clustering*, Theory, Algorithms and Applications, SIAM & ASA, Philadelphia.

[3] QOL, (2005), http://www.economist.com/media/pdf/QUALITY_OF_LIFE.PDF, The Economist Intelligence Unit's quality-of-life index (as seen on September 25th, 2012).

[4] QOL, (2007), http://www.il-ireland.com/il/qofl07/2007 Quality of Life Index (as seen on September 25th, 2012).

[5] Nielsen L., (2011), *Classification of Countries Based on Their Level of Development: How it is Done and How it Could be Done*, IMF Working Paper, WP/11/31, IMF.

[6] Owsiński J.W., (1990), *On a new naturally indexed quick clustering method with a global objective function*, Applied Stochastic Models and Data Analysis, 6, 157-171.

[7] Owsiński J.W., (2011), *The bi-partial approach in clustering and ordering: the model and the algorithms*, Statistica & Applicazioni, Special Issue, 43-59.

[8] Owsiński J.W., (2012), *On dividing an empirical distribution into optimal segments*, SIS (Italian Statistical Society) Scientific Meeting, Rome, June 2012, http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/viewFile/2368/229

OPTYMALNY PODZIAŁ ROZKŁADU EMPIRYCZNEGO (I KILKA PROBLEMÓW Z TYM ZWIĄZANYCH)

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

Praca zajmuje się podziałem empirycznego rozkładu wielkości x_i, gdzie i jest indeksem jednostki, dla której obserwujemy tę wielkość (np. x_i to PKB na mieszkańca w kraju i-tym). Wartości x_i uporządkowano niemalejąco. Analizujemy dystrybuantę rozkładu, tj. wartości $z_i = \Sigma_{i'=1,...,i} x_i$, które tworzą ciąg

wypukły. Chcemy otrzymać taki podział dystrybuanty na podzbiory, by przybliżyć kształt rozkładu $\{z_i\}$ z możliwie małym błędem przy pomocy odcinków linii prostej, odpowiadających podzbiorom, a zarazem – by tych odcinków było możliwie mało. Odpowiada to kategoryzacji podobnych rozkładów (np. kraje „rozwinięte", „rozwijające się", ...), gdzie zwykle nie stosuje się metod statystycznych, tylko przesłanki „merytoryczne", bądź stosowanie metod statystycznych ogranicza się do ustalenia, np., kwantyli rozkładu, bez uwzględniania kształtu i innych przesłanek dla rozwiązania, optymalizującego wspomniane kryterium. Zaproponowano ogólną metodykę optymalizacji podziału takich rozkładów w duchu wspomnianego kryterium, funkcję celu i jej konkretną realizację, wraz z algorytmami. Na podstawie przykładów konkretnych rozkładów, zarysowano także problemy, wynikające z faktu, że rozkłady empiryczne mają często charakter, stawiający pod znakiem zapytania podstawy przyjętej metodyki i w ogóle sens podobnych zadań. Przeanalizowano możliwe pochodzenie tych rozkładów oraz skutki dla ewentualnej kategoryzacji. Zaproponowana metodyka daje podstawy do kategoryzacji empirycznych dystrybuant i narzędzie do oceny racjonalności sposobu ich otrzymywania.

Słowa kluczowe: rozkład empiryczny, kategoryzacja, optymalizacja, funkcja kryterium, dwustronna funkcja kryterium

ON THE OPTIMAL DIVISION OF AN EMPIRICAL DISTRIBUTION (AND SOME RELATED PROBLEMS)

A b s t r a c t

We consider division of an empirical distribution of x_i, i being the index of a unit, for which we observe x_i (e.g., province i, for which x_i is the GDP per capita). Values x_i are ordered non-decreasingly. We analyse the cumulative distribution, $z_i = \Sigma_{i'=1,...,i} x_i$. The sequence z_i is convex. We want to divide the distribution of z_i into subsets of i, with the shape of the distribution $\{z_i\}$ possibly well approximated by the segments of the straight line, determined for the subsets, forming a piecewise linear contour, the number of segments being possibly small. This corresponds to the frequently used categorisations for similar distributions (e.g., "developed", "developing",... countries). For such categorisations, usually no formal methods are applied but "substantive" prerequisites, or the methods applied are limited to establishing quantiles of the distribution, without considering its shape and the objective premises for determination of a different number of segments, including optimisation of the criterion mentioned before. A general approach is proposed for optimising division of such distribution conform to the criterion mentioned. A general objective function is proposed and its concrete realisation, as well as algorithms. The methodology proposed allows for obtaining the optimum divisions into categories for arbitrary distributions. Yet, on the basis of concrete empirical distributions, problems are outlined, due to the fact that the distributions obtained often display the features, leading to questioning of the foundations of the methodology proposed, and of the very sense of such categorisations. Examples of distributions of this kind, and consequences for the potential categorisations, are discussed. In summary, the methodology proposed, including the criterion function, constitutes a basis for the categorisation with respect to the cumulative distribution, and a tool for evaluating the rationality of the way, in which the distributions are obtained.

Key words: empirical distribution, optimum division, classes, objective function, bi-partial approach

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

WIOLETTA GRZENDA

BADANIE DETERMINANT POZOSTAWANIA BEZ PRACY OSÓB MŁODYCH Z WYKORZYSTANIEM SEMIPARAMETRYCZNEGO MODELU COXA[1]

1. WSTĘP

W Polsce w ostatnich latach wskaźniki zatrudnienia dla ludzi młodych, w porównaniu z ogółem społeczeństwa, są bardzo niskie. Podobną sytuację obserwuje się w całej Europie. Według danych BAEL w czwartym kwartale 2011 roku stopa bezrobocia w Polsce wynosiła 9,7%, natomiast wśród osób młodych w wieku 15-34 lata 36,6%. W roku 2009 w tym samym kwartale stopa bezrobocia wynosiła odpowiednio 8,5% i 31%.

Trudności w znalezieniu pracy w znacznym stopniu ograniczają usamodzielnienie ekonomiczne ludzi młodych, a z tym często wiąże się zależność ekonomiczna od rodziców lub opiekunów oraz odkładnie decyzji o zakładaniu własnej rodziny. W celu zwiększenia swoich szans na rynku pracy młode osoby często kontynuują edukację, jednak otrzymanie dyplomu nie zawsze poprawia ich sytuację. Ponadto młode osoby kończąc edukację często nie posiadają doświadczenia zawodowego. Aby to zmienić, w trakcie procesu kształcenia, powinny one podejmować różne formy praktyk zawodowych. Jednocześnie w dynamicznej gospodarce można obserwować niedopasowania strukturalne popytu na pracę i podaży pracy pod względem zawodów, wykształcenia, kwalifikacji oraz miejsca zamieszkania i pracy (Kwiatkowski, 2007). Istotne znaczenie mają badania dotyczące wpływu indywidualnych cech osób bezrobotnych na długość czasu poszukiwania pracy, bowiem otrzymywane oferty pracy zależą głównie od tych cech oraz sytuacji na rynku pracy. Wśród badanych cech najczęściej wyróżnia się: płeć, stan cywilny, wykształcenie, wiek oraz rodzaj wcześniej wykonywanej pracy (Collier, 2003). Celem niniejszego opracowania jest identyfikacja czynników demograficznych oraz społeczno-ekonomicznych wpływających na długość czasu pozostawania bez pracy osób młodych w wieku od 15 do 35 lat.

Analizę determinant bezrobocia najczęściej prowadzi się na podstawie standaryzowanych stóp bezrobocia (Socha, Sztanderska, 2000). W analizie przyczyn zróżnicowania długości okresu pozostawania bez pracy poszczególnych osób wykorzystuje się również metody ekonometryczne (Lancaster, 1979). Modelowe ujęcie bezrobocia oparte na modelu ekonometrycznym, z uwzględnieniem między innymi zmiennych opi-

[1] Badanie zostało zrealizowane w ramach badania statutowego „Badanie rynku pracy z wykorzystaniem bayesowskich modeli analizy historii zdarzeń" nr 03/S/0044/12, Szkoła Główna Handlowa.

sujących zatrudnienie w małych i średnich przedsiębiorstwach, nakłady na działalność innowacyjną oraz transport, zawarte jest w pracy Balcerowicz-Szkutnik i inni (2010). Przy badaniu cech demograficznych oraz społeczno-ekonomicznych, takich jak płeć, poziom wykształcenia, stan cywilny, wiek, czy miejsce zamieszkania wykorzystywane są głównie modele logitowe (Grzenda, 2011) oraz modele analizy historii zdarzeń, w tym model wykładniczy przedziałami stały (Drobnič, Frątczak, 2001), jak również wielopoziomowe modele przeżycia (Biggeri, Bini, Grilli, 2001). W ostatniej pracy analizowane są nie tylko wybrane cechy różnicujące absolwentów uczelni wyższych, ale również efektywność kształcenia na uczelniach w kontekście szans znalezienia pracy.

W niniejszej pracy do badania determinant długości czasu pozostawania bezrobotnym wykorzystano bayesowski semiparametryczny model Coxa dla danych indywidualnych. Model ten został zaproponowany przez Coxa (Cox, 1972; Cox, 1975) i najczęściej jest wykorzystywany w medycynie do badania przeżycia. W pracy (Merrick, Soyer, Mazzuchi, 2002) zaproponowano również wykorzystanie tego modelu w badaniu niezawodności, tj. analizy długości czasu właściwego działania maszyny. Obecnie często spotyka się wykorzystanie tego modelu w ekonomii i naukach społecznych do badania czasu trwania określonego zjawiska. Stosowanie tego modelu jest zalecane w badaniu kierunku i siły wpływu zmiennych na czas trwania danego zjawiska (Blossfeld, Rohwer, 2002). Ponadto model ten daje możliwość uwzględnienia w badaniu zmiennych, których stopień wpływu na czas trwania danego zjawiska jest różny w badanym okresie.

2. ZAKRES BADAŃ

Podejście bayesowskie, za pomocą rozkładów *a priori*, daje możliwość włączenia do wnioskowania wstępnych informacji spoza próby. W niniejszym opracowaniu zaproponowano wykorzystanie charakterystyk rozkładów *a posteriori* otrzymanych w roku poprzednim, jako informacji *a priori* dla modelowania danych pochodzących z roku następnego.

Dwa zbiory danych wykorzystane w badaniu pochodzą z badań Głównego Urzędu Statystycznego „Budżety Gospodarstw Domowych 2008" oraz „Budżety Gospodarstw Domowych 2009". W pierwszym badaniu zbadano 37358 gospodarstw domowych obejmujących łączną sumę 109819 respondentów, w drugim odpowiednio 37302 gospodarstw domowych i 108038 respondentów. Ze względu na cel badania w niniejszej analizie uwzględniono tylko osoby bezrobotne w wieku od 15 do 35 lat, które poszukiwały pracy i były gotowe podjąć pracę (por. zalecenia EUROSTAT). W ten sposób wybrano do analizy 1258 osób z badania w roku 2008 oraz 1747 osób z badania w roku 2009. Wśród osób wybranych do badania, 128 osób z badania w roku 2008 oraz 130 osób z badania w roku 2009 znalazło już pracę i czekało na jej rozpoczęcie. Takie osoby to jednostki, dla których wystąpiło zdarzenie. Pozostałe zaś osoby, to jednostki ocenzurowane. Uwzględniając to, że różne czynniki mogą determinować bezrobocie, w zależności od czasu jego trwania, badano tylko te osoby, które pozostawały bez pracy maksymalnie 12 miesięcy.

W badaniu zmienną zależną jest *czas* wyrażony liczbą miesięcy pozostawania bez pracy, natomiast jako zmienne objaśniające wybrano: płeć, stan cywilny, poziom wykształcenia, informację o tym, czy respondent nadal się dokształca, grupę społeczno-ekonomiczną, region Polski (według podziału przyjętego przez EUROSTAT), który zamieszkuje respondent, klasę miejscowości zamieszkania oraz wiek. Poniżej podajemy charakterystykę tych zmiennych oraz problemy badawcze z nimi związane.

Wiele opracowań dotyczy równości szans na rynku pracy kobiet i mężczyzn, zatem jako pierwszą potencjalną zmienną wybrano *płeć*: 1 – mężczyzna (55,7%), 2 – kobieta (44,3%). Przypuszcza się, że mężczyźni mają potencjalnie większe szanse na znalezienie pracy niż kobiety, z uwagi na rolę, jaką pełnią kobiety w rodzinie.

Osoby młode, szczególnie kobiety, są mniej skłonne do zakładania rodziny, ponieważ obawiają się trudności związanych ze znalezieniem pracy. Przypuszcza się jednocześnie, że pracodawcy chętniej zatrudniają osoby stanu wolnego. Zatem warto zweryfikować to przypuszczenie. Zmienna *stan cywilny*: 1 – kawaler, panna, w separacji, rozwiedziony(a), wdowiec, wdowa (77,96%), 2 – żonaty, zamężna (22,04%).

Można przypuszczać, że poziom wykształcenia jest jednym z najsilniejszych czynników wpływających na szanse znalezienia pracy. Inne opracowania wskazują, że największe szanse na znalezienie pracy mają osoby z wykształceniem wyższym oraz średnim zawodowym. Zmienna *wykształcenie*: 1 – wyższe (14,54%), 2 – policealne i średnie zawodowe (27,19%), 3 – średnie ogólnokształcące (17,46%), 4 – zasadnicze zawodowe (26,96%), 5 – gimnazjalne lub podstawowe (13,85%).

Badamy osoby młode, które często dokształcają się, zatem zweryfikujemy, czy znaczenie ma fakt, że respondent nadal dokształca się. Informację o tym zawiera zmienna *kontynuacja nauki*: 1 – tak (14,6%), 2 – nie (85,4%).

Przypuszcza się również, że większe szanse na znalezienie pracy mają osoby zamieszkujące duże miasta niż mieszkańcy wsi. Zmienna *klasa miejsca zamieszkania*: 1 – miasto powyżej 100 tys. mieszkańców (21,47%), 2 – miasto poniżej 100 tys. mieszkańców (32,05%), 3 – wieś (46,48%).

Poszczególne obszary Polski cechuje różny rozwój gospodarczy; dlatego zbadamy, czy i w jakim stopniu podział terytorialny wpływa na szanse znalezienia pracy. Zmienna *region*: 1 – centralny (województwa: łódzkie, mazowieckie) (17,52%), 2 – południowo-zachodni (województwa: dolnośląskie, opolskie) (17,57%), 3 – południowy (województwa: małopolskie, śląskie) (20,89%), 4 – północno-zachodni (województwa: wielkopolskie, zachodniopomorskie, lubuskie) (15,34%), 5 – północny (województwa: kujawsko-pomorskie, warmińsko-mazurskie, pomorskie) (10,65%), 6 – wschodni (województwa: lubelskie, podkarpackie, świętokrzyskie, podlaskie) (18,03%).

Kolejna determinanta, którą uwzględniono w badaniu, to *grupa społeczno-ekonomiczna*: 1– pracownicy najemni (60,78%), 2 – rolnicy (5,55%), 3 – pracujący na własny rachunek (5,84%), 4 – emeryci, renciści, utrzymujący się z niezarobkowych źródeł (27,82%). Ponadto w przypadku tej cechy warto zweryfikować sytuację osób samo zatrudniających się.

W badaniu uwzględniamy tylko osoby młode w wieku 15-35 lat. Uwzględniając potencjalną możliwość uzyskania odpowiedniego poziomu wykształcenia, podzielono respondentów na grupy według wieku. Zmienna *grupa wieku*: 1 – do 19 lat (11,56%), 2 – 20-24 lata (51,57%), 3 – powyżej 25 lat (36, 86%).

3. METODA BADANIA I ESTYMACJA MODELU

3.1. MODEL COXA I JEGO ZAŁOŻENIA

W niniejszej pracy do analizy determinant wpływających na długość czasu pozostawania bez pracy wybrano model proporcjonalnego hazardu Coxa (Cox, 1972). Model ten nie wymaga założeń dotyczących kształtu rozkładu przeżyć. W wersji klasycznej ma on następującą postać:

$$h(t;\, \mathbf{x}) = h_0(t) \exp(\beta \mathbf{x}),$$

lub równoważnie:

$$h\left(t;\, x_1,\, x_2,\, \ldots,\, x_p\right) = h_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right),$$

gdzie

$\mathbf{x} = \left(x_1,\, x_2,\, \ldots,\, x_p\right)'$ – wektor zmiennych objaśniających,

$\beta = \left(\beta_1,\, \beta_2,\, \ldots,\, \beta_p\right)$ – wektor współczynników regresji,

$h\left(t;\, x_1,\, x_2,\, \ldots,\, x_p\right)$ – oznacza ryzyko (hazard) w czasie t dla p zmiennych objaśniających,

$h_0(t)$, $h_0(t) > 0$ – niewyspecyfikowany parametrycznie hazard odniesienia (*baseline hazard*).

Po obustronnym zlogarytmowaniu otrzymujemy model liniowy, którego parametry są szacowane:

$$\ln(h(t)) = \ln(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

Funkcją zależną od t jest tylko hazard odniesienia, zatem w takim modelu zakłada się, że pozostałe zmienne w badanym okresie nie zmieniają się w czasie.

Przed przystąpieniem do estymacji bayesowskiej sprawdzono metodami niebayesowskimi założenia klasycznego modelu Coxa. Pierwsze mówi, że istnieje log-liniowa zależność pomiędzy zmiennymi niezależnymi a funkcją hazardu. Drugie, zwane założeniem proporcjonalności, mówi, że iloraz hazardu dla dwóch wektorów zmiennych niezależnych jest niezależny od czasu.

W modelu uwzględniamy tylko zmienne dyskretne, zatem weryfikujemy założenie proporcjonalności hazardów. W tym celu oszacowano model Coxa uwzględniający interakcję poszczególnych zmiennych ze zmienną czasową. Dodatkowe zmienne zostały określone jako iloczyny zmiennych objaśniających i zmiennej oznaczającej czas. W wyniku estymacji otrzymano, że współczynniki przy tych iloczynach są statystycznie istotne, zatem założenie proporcjonalności hazardów nie jest spełnione. Inną metodą

weryfikacji tego założenia jest metoda graficzna. Wyznaczamy wykresy funkcji hazardu dla grup wyznaczonych przez poszczególne poziomy niezależnych zmiennych jakościowych i weryfikujemy, czy są równoległe względem czasu. Zamiennie zamiast funkcji hazardu można rozważać funkcje dożycia:

$$\frac{\ln(S_1(t))}{\ln(S_2(t))} = const.$$

Otrzymane wykresy potwierdziły nasze wcześniejsze przypuszczenia, że dla rozważanych danych należy rozważać model Coxa ze zmiennymi objaśniającymi zależnymi od czasu. Wynika to z faktu, iż zmienne objaśniające w różnym stopniu wpływają na ryzyko zależnie od czasu trwania badanego zjawiska. Wówczas model ten można zapisać w następującej postaci:

$$h(t;\, \mathbf{x}) = h_0(t) \exp(\beta \mathbf{x}(t)).$$

Na podstawie otrzymanych wykresów funkcji przeżycia zostały do modelu wprowadzone zmienne wskaźnikowe zależne od czasu. W tym celu oszacowano dla każdej zmiennej objaśniającej kilka modeli i wybrano ten najlepszy opierając się o kryteria AIC (*Akaike'e Information Criterion*) i SBC (*Shwarz's Bayesian Criterion*).

W niniejszej pracy przeprowadzono analizę determinant bezrobocia z wykorzystaniem modelu Coxa w ujęciu bayesowskim. Rozważania dotyczące modeli przeżycia w ujęciu bayesowskim można znaleźć w wielu pracach (Sinha, Dey, 1997; Ibrahim, Chen, Sinha, 2001; Sinha, Ibrahim, Chen, 2003; Congdon, 2007).

W modelach przeżycia obserwowane dane można zapisać w następującej postaci: $D = (n, \mathbf{t}, \mathbf{X}, \mathbf{v})$. Wektor $\mathbf{t} = (t_1, t_2, \ldots, t_n)'$ oznacza czasy przeżycia, $\mathbf{v} = (v_1, v_2, \ldots, v_n)'$ wektor zmiennych cenzurujących, przy czym $v_i = 0$, jeśli t_i jest prawostronnie ocenzurowany oraz $v_i = 1$, jeśli dla t_i wystąpiło zdarzenie, $i = 1, ..., n$, a $\mathbf{X}$ oznacza macierz zmiennych objaśniających o wymiarze $n \times p$, której i-ty wiersz oznaczamy $\mathbf{x}'_i$.

Podstawową metodą estymacji parametrów modelu Coxa jest metoda częściowej wiarygodności (*partial likelihood method*). Jeśli nie mamy wyspecyfikowanej parametrycznie funkcji dla hazardu bazowego, to nie można stosować tradycyjnej metody największej wiarygodności. Wówczas wektor $\boldsymbol{\beta}$ współczynników regresji jest szacowany poprzez maksymalizację tzw. funkcji częściowej wiarygodności. Oszacowania współczynników modelu uzyskane w wyniku maksymalizacji funkcji częściowej wiarygodności mają właściwości podobne do tych uzyskanych metodą maksymalizacji standardowej funkcji wiarygodności (Cox, 1972; Cox, 1975; Blossfeld, Rohwer, 2002).

W rozważanym modelu parametrami są współczynniki regresji $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$. Podejście bayesowskie wymaga określenia rozkładów *a priori* dla wszystkich parametrów modelu; łączny rozkład *a priori* oznaczamy $p(\boldsymbol{\beta})$. Najczęściej dla wektora parametrów $\boldsymbol{\beta}$ wybierany jest p-wymiarowy rozkład normalny *a priori* $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, gdzie $\boldsymbol{\mu}_0$ oznacza wektor średnich, a $\boldsymbol{\Sigma}_0$ macierz wariancji-kowariancji. Wnioskowanie w podejściu bayesowskim opiera się na rozkładach *a posteriori* $p(\boldsymbol{\beta}|D)$ wyznaczonych

z wykorzystaniem twierdzenia Bayesa. W przypadku wykorzystania w estymacji funkcji częściowej wiarygodności otrzymuje się pewne przybliżenie rozkładu *a posteriori* dla wektora parametrów $\boldsymbol{\beta}$.

W podejściu bayesowskim wnioskowanie o dowolnej współrzędnej β_i wektora parametrów $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ odbywa się na podstawie brzegowego rozkładu *a posteriori*, otrzymanego poprzez całkowanie po pozostałych współrzędnych łącznego rozkładu *a posteriori* (Osiewalski, 2001):

$$p(\beta_i|D) = \int \ldots \int p(\beta_1, \ldots, \beta_p|D)\, d\beta_1 \ldots d\beta_{i-1} d\beta_{i+1} \ldots d\beta_p$$

W praktyce często wykorzystujemy metody symulacyjne, które umożliwiają generowanie prób losowych[2] z dowolnej gęstości *a posteriori*, dla których można wyznaczać różne charakterystyki. Obecnie dużą popularnością cieszą się metody Monte Carlo oparte na łańcuchach Markowa (*Markov Chain Monte Carlo Method* – MCMC).

Podstawą metod MCMC jest generowanie ergodycznego łańcucha Markowa, który po upływie odpowiednio długiego czasu osiąga rozkład stacjonarny, którym jest rozkład *a posteriori*. Najbardziej znanym algorytmem tych metod obliczeniowych jest algorytm Metropolisa, w niniejszej pracy zostało wykorzystane jego uogólnienie – algorytm ARMS (*Adaptive Rejection Metropolis Sampling Algorithm*). Opis użytego algorytmu zawarty jest w pracy Gilks, Best i Tan (1995).

3.2. SPECYFIKACJA I ESTYMACJA MODELU

Z uwagi na to, że zmienne objaśniające w różnym stopniu wpływają na ryzyko zależnie od czasu, do modelu oprócz zmiennych objaśniających przedstawionych w części 2 niniejszej pracy zostaną włączone zmienne wskaźnikowe zależne od czasu (Tabela 1.). Zmienne te zostały wyznaczone na podstawie wykresów funkcji dożycia.

Przed przystąpieniem do estymacji głównego modelu dla danych z roku 2009 oszacowano analogiczny model w podejściu bayesowskim dla danych z roku 2008. Chcąc osiągnąć wyniki obiektywnie poprawne, w pierwszym modelu użyto nieinformacyjnych rozkładów[3] *a priori*.

---

[2] W literaturze spotyka się to pojęcie, należy jednak pamiętać, że ciągi liczb uzyskane w wyniku działania deterministycznych algorytmów są pseudolosowe, ponieważ żaden komputer nie jest w stanie wygenerować prawdziwie losowego ciągu liczb.

[3] W podejściu bayesowskim szczególną rolę odgrywają tzw. nieinformacyjne rozkłady *a priori* „pozwalające danym mówić za siebie". Rozkłady te mogą służyć do wyrażenia zupełnego braku wiedzy o szacowanych parametrach albo mogą służyć jako punkt odniesienia przy porównywaniu wyników uzyskiwanych z wykorzystaniem informacyjnych rozkładów *a priori*, wówczas zwane są one rozkładami referencyjnymi lub wzorcowymi (Osiewalski, 2001, Szreder, 1994). Jeśli otrzymane rezultaty w głównej mierze mają odzwierciedlać wpływ zaobserwowanych danych, to jako rozkłady *a priori* wybiera się rozkłady rozproszone (Pipień, 2006). Wówczas w praktyce zamiast rozkładów nieinformacyjnych, stosujemy tzw. rozkłady wystarczająco mało informacyjne. W przypadku rozkładów normalnych najczęściej dla średniej przyjmujemy wartość zero, a dla wariancji pewną dużą liczbę.

Tabela 1.

Zmienne wskaźnikowe

Zmienna	**Poziom**	**Przedział czasu**
płeć 1	1	<= 9 miesięcy
płeć 2	1	>9 miesięcy
stan cywilny 1	1	<= 3 miesięcy
stan cywilny 2	1	>3 miesięcy
kontynuacja nauki 1	1	<= 6 miesięcy
kontynuacja nauki 2	1	>6 miesięcy
klasa miejsca zamieszkania 1	1	<= 9 miesięcy
klasa miejsca zamieszkania 1	2	<= 9 miesięcy
klasa miejsca zamieszkania 2	1	>9 miesięcy
klasa miejsca zamieszkania 2	2	>9 miesięcy
grupa wieku 1	1	<= 3 miesięcy
grupa wieku 1	2	<= 3 miesięcy
grupa wieku 2	1	>3 miesięcy
grupa wieku 2	2	>3 miesięcy
grupa społeczno-ekonomiczna 1	2	<= 3 miesięcy
grupa społeczno-ekonomiczna 1	3	<= 3 miesięcy
grupa społeczno-ekonomiczna 1	4	<= 3 miesięcy
grupa społeczno-ekonomiczna 2	2	>3 miesięcy
grupa społeczno-ekonomiczna 2	3	>3 miesięcy
grupa społeczno-ekonomiczna 2	4	>3 miesięcy
wykształcenie 1	1	<= 3 miesięcy
wykształcenie 1	2	<= 3 miesięcy
wykształcenie 1	3	<= 3 miesięcy
wykształcenie 1	4	<= 3 miesięcy
wykształcenie 2	1	>3 miesięcy
wykształcenie 2	2	>3 miesięcy
wykształcenie 2	3	>3 miesięcy
wykształcenie 2	4	>3 miesięcy
region 1	1	<= 3 miesięcy
region 1	2	<= 3 miesięcy
region 1	3	<= 3 miesięcy
region 1	4	<= 3 miesięcy
region 1	5	<= 3 miesięcy
region 2	1	>3 miesięcy
region 2	2	>3 miesięcy
region 2	3	>3 miesięcy
region 2	4	>3 miesięcy
region 2	5	>3 miesięcy

Źródło: obliczenia własne[4] na podstawie danych GUS 2009.

[4] Wszystkie obliczenia zostały wykonane w systemie SAS 9.2.

W estymacji przyjęto nieinformacyjne rozkłady normalne *a priori* ze średnią równą 0 i wariancją 10^6 dla wszystkich parametrów regresji: $\beta \sim N\left(\mathbf{0},\ 10^6\mathbf{I}\right)$. W obu modelach przyjęto, że liczba cykli spalonych ma wynosić 2000, a liczba właściwych realizacji przyjętych jako próba z rozkładu *a posteriori* 10000.

Podejście bayesowskie daje możliwość uwzględnienia w badaniu, za pomocą rozkładów *a priori*, dodatkowej informacji spoza próby. Szczególne znaczenie w estymacji mają rozkłady *a priori*, które zostały skonstruowane na podstawie poprzednich empirycznych oszacowań parametrów dokonanych na bazie wcześniej posiadanych informacji. Uzyskane na postawie wcześniej pobranej próby rozkłady *a posteriori* stają się wiedzą *a priori*, dla nowych danych. W ten sposób można dokonywać ciągłej weryfikacji ocen na podstawie nowych informacji (Szreder, 1994). W niniejszym opracowaniu charakterystyki rozkładu *a posteriori* dla roku 2008 zostały wykorzystane jako informacja *a priori* w modelu dla danych z roku 2009.

Przed przystąpieniem do wnioskowań z wykorzystaniem prób z rozkładu *a posteriori* zweryfikowano zbieżność wygenerowanych łańcuchów. Analogiczne postępowanie zostało przeprowadzone dla danych z roku 2008 i 2009. Natomiast wyniki przedstawiamy tylko dla roku 2009. Często wykorzystywanym testem do sprawdzenia, czy wygenerowany łańcuch osiągnął rozkład stacjonarny jest test Geweke'a (Geweke, 1992). Na podstawie wyników przedstawionych w tabeli 2, dla danych z roku 2009, stwierdzono brak podstaw do odrzucenia hipotezy, że wygenerowane łańcuchy dla poszczególnych parametrów modelu są zbieżne, przy poziomie istotności 0,01.

Tabela 2.

Wyniki dla testu Geweke'a

Parametr		**Wartość statystyki *z***	***p - value***
płeć 1	1	-0,5852	0,5584
płeć 2	1	1,1985	0,2307
stan cywilny 1	1	-0,6410	0,5215
stan cywilny 2	1	0,4992	0,6177
kontynuacja nauki 1	1	-0,5718	0,5674
kontynuacja nauki 2	1	-0,9181	0,3586
klasa miejsca zamieszkania 1	1	-2,3914	0,0168
klasa miejsca zamieszkania 1	2	-2,0012	0,0454
klasa miejsca zamieszkania 2	1	0,0278	0,9778
klasa miejsca zamieszkania 2	2	0,5047	0,6138
grupa wieku 1	1	0,1116	0,9111
grupa wieku 1	2	0,1408	0,8880
grupa wieku 2	1	-1,3907	0,1643
grupa wieku 2	2	-1,0742	0,2827
grupa społeczno-ekonomiczna 1	2	-1,0430	0,2969
grupa społeczno-ekonomiczna 1	3	-1,6176	0,1058
grupa społeczno-ekonomiczna 1	4	-0,7212	0,4708

cd. Tabela 2.

grupa społeczno-ekonomiczna 2	2	0,5396	0,5894
grupa społeczno-ekonomiczna 2	3	-0,0881	0,9298
grupa społeczno-ekonomiczna 2	4	-0,5912	0,5544
wykształcenie 1	1	-0,3599	0,7189
wykształcenie 1	2	-0,5383	0,5904
wykształcenie 1	3	-0,4091	0,6825
wykształcenie 1	4	-0,6179	0,5366
wykształcenie 2	1	-0,0952	0,9242
wykształcenie 2	2	-0,5559	0,5783
wykształcenie 2	3	-0,4902	0,6240
wykształcenie 2	4	-0,2891	0,7725
region 1	1	-0,5509	0,5817
region 1	2	0,0109	0,9913
region 1	3	-1,3753	0,1690
region 1	4	-0,5463	0,5849
region 1	5	-0,8963	0,3701
region 2	1	-0,7386	0,4602
region 2	2	-0,5522	0,5808
region 2	3	-1,4720	0,1410
region 2	4	-0,9901	0,3221
region 2	5	-1,6833	0,0923

Źródło: obliczenia własne na podstawie danych GUS 2009

W badaniu stacjonarności równie powszechna jest metoda graficzna, która potwierdziła otrzymane wnioski z testu Geweke'a. Zatem po odrzuceniu 2000 cykli wstępnych, 10000 stanów łańcucha zostało przyjętych jako próba z rozkładu *a posteriori*.

Wartości oczekiwane *a posteriori* parametrów modelu dla roku 2008 zostały przedstawione w tabeli 3, a dla roku 2009 w tabeli 4.

Tabela 3.

Charakterystyki *a posteriori* dla danych z 2008 roku

Parametr		**Wartości oczekiwane *a posteriori***	**Odchylania standardowe *a posteriori***	**Obszary największej wartości funkcji gęstości *a posteriori*[5]**	
płeć 1	1	0,1988	0,00543	0,1881	0,2091
płeć 2	1	0,4452	0,0177	0,4103	0,4795
stan cywilny 1	1	0,4516	0,00925	0,4334	0,4695
stan cywilny 2	1	-0,0440	0,0103	-0,0634	-0,0232
kontynuacja nauki 1	1	-0,2418	0,00888	-0,2593	-0,2246

[5] Obszary największej wartości funkcji gęstości *a posteriori* (HPD) zostały wyznaczone przy $\alpha = 0,05$ dla wszystkich parametrów modelu.

cd. Tabela 3.

kontynuacja nauki 2	1	0,4371	0,0180	0,4022	0,4721
klasa miejsca zamieszkania 1	1	0,00384	0,00667	-0,00944	0,0166
klasa miejsca zamieszkania 1	2	-0,0774	0,00615	-0,0892	-0,0652
klasa miejsca zamieszkania 2	1	0,9785	0,0268	0,9256	1,0298
klasa miejsca zamieszkania 2	2	1,3041	0,0248	1,2551	1,3531
grupa wieku 1	1	-0,2338	0,0117	-0,2570	-0,2115
grupa wieku 1	2	-0,5760	0,00738	-0,5903	-0,5616
grupa wieku 2	1	-0,0174	0,0157	-0,0484	0,0129
grupa wieku 2	2	0,0603	0,00896	0,0427	0,0776
grupa społeczno-ekonomiczna 1	2	0,4488	0,0152	0,4193	0,4788
grupa społeczno-ekonomiczna 1	3	-0,7212	0,0165	-0,7538	-0,6892
grupa społeczno-ekonomiczna 1	4	-0,1038	0,00723	-0,1184	-0,0900
grupa społeczno-ekonomiczna 2	2	-0,3534	0,0219	-0,3967	-0,3112
grupa społeczno-ekonomiczna 2	3	-0,1373	0,0146	-0,1662	-0,1094
grupa społeczno-ekonomiczna 2	4	-0,3249	0,00929	-0,3426	-0,3066
wykształcenie 1	1	0,9720	0,0140	0,9445	0,9991
wykształcenie 1	2	0,9126	0,0132	0,8865	0,9382
wykształcenie 1	3	0,7549	0,0144	0,7261	0,7821
wykształcenie 1	4	0,8834	0,0130	0,8578	0,9087
wykształcenie 2	1	0,4980	0,0124	0,4736	0,5222
wykształcenie 2	2	-0,0686	0,0115	-0,0913	-0,0462
wykształcenie 2	3	-0,0560	0,0129	-0,0823	-0,0316
wykształcenie 2	4	-0,5485	0,0124	-0,5714	-0,5229
region 1	1	0,6351	0,00997	0,6161	0,6544
region 1	2	0,3016	0,0108	0,2809	0,3231
region 1	3	-0,2968	0,0115	-0,3192	-0,2736
region 1	4	-0,8575	0,0152	-0,8872	-0,8276
region 1	5	0,0409	0,0129	0,0142	0,0653
region 2	1	0,1681	0,0139	0,1417	0,1956
region 2	2	0,2160	0,0134	0,1902	0,2435
region 2	3	-0,0654	0,0132	-0,0912	-0,0392
region 2	4	0,7914	0,0127	0,7661	0,8156
region 2	5	-0,8038	0,0220	-0,8454	-0,7595

Źródło: obliczenia własne na podstawie danych GUS 2008

Tabela 4.

Charakterystyki *a posteriori* dla danych z 2009 roku

Parametr		**Wartości oczekiwane** ***a posteriori***	**Odchylania standardowe** ***a posteriori***	**Obszary największej wartości funkcji gęstości** ***a posteriori***	
płeć 1	1	0,0457	0,00514	0,0356	0,0558
płeć 2	1	-0,2666	0,0168	-0,2990	-0,2335
stan cywilny 1	1	-0,3218	0,00817	-0,3370	-0,3050
stan cywilny 2	1	-0,0891	0,0101	-0,1080	-0,0687
kontynuacja nauki 1	1	0,3332	0,00665	0,3204	0,3465
kontynuacja nauki 2	1	-0,6685	0,0269	-0,7214	-0,6173
klasa miejsca zamieszkania 1	1	0,0162	0,00671	0,00339	0,0298
klasa miejsca zamieszkania 1	2	0,0899	0,00579	0,0788	0,1015
klasa miejsca zamieszkania 2	1	-0,6098	0,0301	-0,6654	-0,5473
klasa miejsca zamieszkania 2	2	0,8915	0,0196	0,8543	0,9307
grupa wieku 1	1	0,7466	0,0114	0,7235	0,7678
grupa wieku 1	2	0,3490	0,00758	0,3343	0,3638
grupa wieku 2	1	-0,0611	0,0160	-0,0924	-0,0301
grupa wieku 2	2	0,0905	0,00874	0,0738	0,1079
grupa społeczno-ekonomiczna 1	2	-0,5546	0,0211	-0,5987	-0,5159
grupa społeczno-ekonomiczna 1	3	0,1872	0,0120	0,1642	0,2114
grupa społeczno-ekonomiczna 1	4	0,3389	0,00678	0,3259	0,3523
grupa społeczno-ekonomiczna 2	2	-1,1247	0,0304	-1,1825	-1,0629
grupa społeczno-ekonomiczna 2	3	1,1158	0,00940	1,0968	1,1337
grupa społeczno-ekonomiczna 2	4	-0,2941	0,00963	-0,3127	-0,2753
wykształcenie 1	1	1,1645	0,0123	1,1400	1,1879
wykształcenie 1	2	0,3591	0,0122	0,3349	0,3829
wykształcenie 1	3	0,7404	0,0119	0,7180	0,7646
wykształcenie 1	4	0,1673	0,0122	0,1428	0,1906
wykształcenie 2	1	1,8094	0,0133	1,7840	1,8335
wykształcenie 2	2	0,9653	0,0127	0,9434	0,9885
wykształcenie 2	3	0,8577	0,0129	0,8359	0,8812
wykształcenie 2	4	1,0065	0,0132	0,9797	1,0308
region 1	1	-0,1589	0,00933	-0,1782	-0,1415
region 1	2	-0,8124	0,0114	-0,8348	-0,7903
region 1	3	-0,5792	0,00991	-0,5983	-0,5595
region 1	4	-0,2035	0,00972	-0,2224	-0,1847
region 1	5	0,1402	0,0102	0,1196	0,1597
region 2	1	-0,3793	0,0131	-0,4045	-0,3535
region 2	2	0,1241	0,0108	0,1034	0,1454
region 2	3	-0,8076	0,0129	-0,8316	-0,7811
region 2	4	0,5861	0,0105	0,5660	0,6069
region 2	5	-0,4445	0,0153	-0,4746	-0,4149

Źródło: obliczenia własne na podstawie danych GUS 2009

Na podstawie otrzymanych obszarów największej wartości funkcji gęstości *a posteriori* można stwierdzić (Bolstad, 2007), że wszystkie zmienne przedstawione w tabeli 4 są statystycznie istotne.

Interpretacji poddajemy współczynniki ryzyka (*hazard ratio*), który definiujemy jako stosunek wartości ryzyka dla jednego respondenta do wartości ryzyka dla innego respondenta (tabela 5.)

Tabela 5.

Oszacowania współczynników ryzyka

Parametr		**Współczynnik ryzyka**
płeć 1	1	1,0468
płeć 2	1	0,7660
stan cywilny 1	1	0,7248
stan cywilny 2	1	0,9148
kontynuacja nauki 1	1	1,3954
kontynuacja nauki 2	1	0,5125
klasa miejsca zamieszkania 1	1	1,0163
klasa miejsca zamieszkania 1	2	1,0941
klasa miejsca zamieszkania 2	1	0,5435
klasa miejsca zamieszkania 2	2	2,4388
grupa wieku 1	1	2,1098
grupa wieku 1	2	1,4176
grupa wieku 2	1	0,9407
grupa wieku 2	2	1,0947
grupa społeczno-ekonomiczna 1	2	0,5743
grupa społeczno-ekonomiczna 1	3	1,2059
grupa społeczno-ekonomiczna 1	4	1,4034
grupa społeczno-ekonomiczna 2	2	0,3247
grupa społeczno-ekonomiczna 2	3	3,0520
grupa społeczno-ekonomiczna 2	4	0,7452
wykształcenie 1	1	3,2043
wykształcenie 1	2	1,4320
wykształcenie 1	3	2,0968
wykształcenie 1	4	1,1821
wykształcenie 2	1	6,1068
wykształcenie 2	2	2,6256
wykształcenie 2	3	2,3577
wykształcenie 2	4	2,7360
region 1	1	0,8531
region 1	2	0,4438
region 1	3	0,5603

cd. Tabela 5.

region 1	4	0,8159
region 1	5	1,1505
region 2	1	0,6843
region 2	2	1,1321
region 2	3	0,4459
region 2	4	1,7970
region 2	5	0,6411

Źródło: obliczenia własne

4. WYNIKI I WNIOSKI

Otrzymane wyniki nie potwierdziły jednoznacznie wcześniejszych przypuszczeń, że kobiety są w gorszej sytuacji na rynku pracy. W ciągu pierwszych 9 miesięcy pozostawania bezrobotnym młodzi mężczyźni mieli tylko o 4,68% wyższe ryzyko przejścia ze stanu bezrobotny do pracujący niż młode kobiety. Jednak po tym okresie ryzyko to dla mężczyzn było mniejsze niż dla kobiet o 23,4%. Zatem można wnioskować, że kobiety coraz lepiej radzą sobie na rynku pracy. Również według badania GUS (2009) w grupie wieku 18-19 lat stopa bezrobocia jest wyższa dla mężczyzn (29,0%) niż dla kobiet (23,4%). Wyniki zawarte w pracach dotyczących wcześniejszego okresu oraz ogółu społeczeństwa (Socha, Sztanderska, 2000) często wskazują na gorszą sytuację kobiet na rynku pracy, niż mężczyzn, mimo że kobiety są często lepiej wykształcone i bardziej aktywnie poszukują pracy.

W przypadku cechy stan cywilny, również wcześniejsze przypuszczenia nie potwierdziły się. Osoby stanu wolnego miały mniejsze szanse na znalezienie pracy niż osoby pozostające w związku. W ciągu pierwszych 3 miesięcy ryzyko przejścia ze stanu bezrobotny do pracujący było dla osób stanu wolnego niższe o 27,52%, natomiast w okresie od 4 do 12 miesięcy niższe o 8,52%. Można przypuszczać, że osoby, które posiadały rodzinę miały większą motywację do bardziej intensywnego poszukiwania pracy.

Potwierdza się, że największe szanse na znalezienie pracy miały osoby z wykształceniem wyższym. W ciągu pierwszych 3 miesięcy było przeszło trzykrotnie bardziej prawdopodobne, że znajdą one pracę niż osoby z wykształceniem gimnazjalnym lub podstawowym. Dla osób z wykształceniem policealnym i średnim zawodowym ryzyko to było wyższe o 43,2%, dla osób z wykształceniem średnim ogólnokształcącym było wyższe o 109,68%, a dla osób z wykształceniem zasadniczym zawodowym było wyższe o 18,21%. W kolejnych miesiącach było przeszło sześciokrotnie bardziej prawdopodobne, że osoby z wykształceniem wyższym znajdą pracę niż osoby z wykształceniem gimnazjalnym lub podstawowym. Dla osób z wykształceniem policealnym i średnim zawodowym, średnim ogólnokształcącym oraz zasadniczym zawodowym było przeszło dwukrotnie bardziej prawdopodobne, że znajdą pracę niż osoby z grupy referencyjnej.

Otrzymane wyniki wskazują, że w ciągu całego badanego okresu szanse na znalezienie pracy przez osoby z wykształceniem średnim ogólnokształcącym, kształtują się podobnie, natomiast dla osób o innym poziomie wykształcenia szanse na znalezienie pracy w ciągu kolejnych 9 miesięcy wzrosły około dwukrotnie w porównaniu do osób z wykształceniem gimnazjalnym lub podstawowym. Potwierdzają to również inne badania dotyczące osób młodych (GUS, 2010) oraz całego społeczeństwa (Grzenda, 2011), wskazujące, że wykształcenie wyższe oraz średnie zawodowe warunkują lepszą sytuację na rynku pracy.

W przypadku osób młodych istotne jest dokształcanie się. Według GUS w badanym okresie około co piąta młoda osoba jednocześnie pracowała i uczyła się. W ciągu pierwszych 6 miesięcy, osoby, które nadal kształciły się miały o 39,54% wyższe ryzyko przejścia ze stanu bezrobotny do pracujący. Po upływie tego okresu, osoby, które nadal kształciły się, miały o 48,75% mniejszą wartość współczynnika ryzyka przejścia ze stanu bezrobotny do pracujący niż osoby, które nie dokształcały się.

W ciągu pierwszych 9 miesięcy osoby, które mieszkały w miastach miały nieznacznie większe szanse na znalezienie pracy niż osoby mieszkające na wsi, w przypadku osób, które mieszkały w miastach powyżej 100 tys. mieszkańców ryzyko przejścia ze stanu bezrobotny do pracujący było wyższe o 1,63%. W przypadku mniejszych miast ryzyko było wyższe o 9,41%. Natomiast po upływie tego okresu, osoby, które mieszkały w dużych miastach miały mniejsze szanse na znalezienie pracy, natomiast w przypadku mniejszych miast było przeszło dwukrotnie bardziej prawdopodobne, że ich mieszkańcy znajdą pracę w porównaniu do osób zamieszkujących na wsi. Zatem przypuszczenie, że większe szanse na znalezienie pracy mają osoby zamieszkujące duże miasta niż mieszkańcy wsi potwierdziło się tylko w początkowym okresie poszukiwania pracy. Otrzymany wynik potwierdzają rezultaty innych badań (Socha, Sztanderska, 2000). Wynika z nich, że różnice w poziomie bezrobocia między dużymi miastami, a wsią są niewielkie zarówno dla osób młodych, jak i w starszym wieku. Natomiast w małych miastach sytuacja osób młodych na rynku pracy jest znacznie lepsza niż na wsi.

Dla cechy region, otrzymano, że w ciągu pierwszych trzech miesięcy osoby zamieszkujące inne regiony niż północny miały mniejsze szanse na znalezienie pracy niż osoby zamieszkujące region wschodni. Natomiast w ciągu kolejnych miesięcy osoby zamieszkujące region południowo-zachodni i północno-zachodni miały większe szanse na znalezienie pracy niż osoby zamieszkujące region wschodni. Otrzymane wyniki nie wskazują jednoznacznie, w których regionach Polski osobom młodych łatwiej znaleźć pracę. Może to wynikać z tego, że są regiony, np. północny, w których w jednym województwie stopa bezrobocia jest niska – pomorskie, a w drugim wysoka – kujawsko-pomorskie (BAEL). Ponadto analizujemy osoby młode, dla których miejsce zamieszkania często nie jest tożsame z miejscem pobytu. Można również przypuszczać, że w regionach, gdzie wskaźnik bezrobocia jest niski, osoby młode mają wyższe oczekiwania dotyczące pierwszej pracy dotyczące np. wysokości wynagrodzenia, czy warunków pracy.

Uwzględniając cechę grupa społeczno-ekonomiczna otrzymano, że w ciągu pierwszych 3 miesięcy ryzyko przejścia ze stanu bezrobotny do pracujący było dla rolników

o 42,57% mniejsze niż dla pracowników najemnych, a dla pracujących na własny rachunek o 20,59% wyższe niż dla pracowników najemnych. W ciągu kolejnych 9 miesięcy sytuacja rolników wyglądała podobnie, natomiast dla osób pracujących na własny rachunek było przeszło trzykrotnie bardziej prawdopodobne, że znajdą pracę niż pracownicy najemni. Otrzymany wynik wskazuje, że samozatrudnienie jest jednym ze sposobów na rozwiązanie problemu bezrobocia, szczególnie wśród osób młodych. Ponadto według badania GUS mężczyźni częściej niż kobiety zakładali własną firmę czy działalność gospodarczą (6,0% wobec 3,9%).

Osoby w wieku do 19 lat w ciągu pierwszych 3 miesięcy miały o 110,98% wyższe ryzyko przejścia ze stanu bezrobotny do pracujący niż osoby w wieku powyżej 25 lat, natomiast po upływie tego okresu ryzyko to było o 5,93% niższe niż dla osób w wieku powyżej 25 lat. Osoby w wieku 20-24 lat w ciągu pierwszych 3 miesięcy miały o 41,76% wyższe ryzyko przejścia ze stanu bezrobotny do pracujący niż osoby w wieku powyżej 25 lat, natomiast po upływie tego okresu ryzyko to było o 9,47% wyższe w stosunku do osób w wieku powyżej 25 lat. Otrzymany wynik może wskazywać na to, że osoby najmłodsze, które ukończyły szkoły średnie lub zasadnicze zawodowe, jeśli znajdą pracę, kończą swoją edukację. Najtrudniej znaleźć pracę osobom w wieku powyżej 25 lat, otrzymany wynik może niepokoić, ponieważ w tej grupie są osoby, które ukończyły studia wyższe. Jednak rezultaty dla cechy wykształcenie pokazały, że osoby z wykształceniem wyższym mają największe szanse na znalezienie pracy. Otrzymane rezultaty są zbieżne z badaniami dla województwa śląskiego (Kryńska, 2004), zgodnie z którymi w grupie osób w wieku 18-24 w latach 2000-2003 zaobserwowano sukcesywny spadek wskaźnika bezrobocia, natomiast wśród osób w wieku 25-34 wzrost.

Wykorzystany w pracy model umożliwił zbadanie występowania zależności pomiędzy wybranymi czynnikami demograficznymi oraz społeczno-ekonomicznymi a długością czasu pozostawania bez pracy osób młodych. Zaletą modeli przeżycia w badaniu aktywności zawodowej jest uwzględnienie całej historii badanej jednostki. Ponadto zaproponowany model umożliwił analizę zmiennych, których stopień wpływu na czas trwania badanego zjawiska jest różny w badanym okresie. Podejście bayesowskie dało możliwość wykorzystania w badaniu informacji pochodzącej z wcześniejszych badań.

Szczegółowa charakterystyka sytuacji osób młodych na rynku pracy może pozwolić na odpowiednie kształtowanie polityki społeczno-gospodarczej zmierzającej do ograniczenia bezrobocia wśród tych osób.

Szkoła Główna Handlowa w Warszawie

LITERATURA

[1] Balcerowicz-Szkutnik M., Dyduch M., Szkutnik W., (2010), *Wybrane modele i analizy rynku pracy: uwarunkowania rynku pracy i wzrostu gospodarczego*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.
[2] Biggeri L., Bini M., Grilli L., (2001), T*he Transition from University to Work: a Multilevel Approach to the Analysis of the Time to Obtain the First Job*, Journal of the Royal Statistical Society, 164, 293-305.
[3] Blossfeld H.P., Rohwer G., (2002), *Techniques of Event History Modeling*, New Approaches to Causal Analysis, Lawrence Erlbaum Associates, New Jersey.
[4] Bolstad W.M., (2007), *Introduction to Bayesian Statistics*, Wiley & Sons, New Jersey.
[5] Collier W., (2003), *The Impact of Demographic and Individual Heterogeneity on Unemployment Duration: A Regional Study*, Studies in Economics, 0302.
[6] Congdon P., (2007), *Bayesian Statistical Modelling*, Wiley & Sons, New York.
[7] Cox D.R., (1972), *Regression Models and Life Tables (with discussion)*, Journal of the Royal Statistical Society, 34, 187-220.
[8] Cox D.R., (1975), *Partial Likelihood*, Biometrika, 62, 269-276.
[9] Drobnič S., Frątczak E., (2001), *Employment Patterns of Married Women in Poland*, w: Blossfeld HP., Drobnič S. (red.), Careers of Couples in Contemporary Societies, Oxford University Press, 281-306.
[10] Geweke J., (1992), *Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments*, w: Bernardo J., Berger J., Dawiv A., Smith, A., Bayesian Statistics, 4, 169-193.
[11] Gilks W., Best N., Tan K., (1995), *Adaptive Rejection Metropolis Sampling with Gibbs Sampling*, Applied Statistics, 44, 455-472.
[12] Grzenda W., (2011), *Wykorzystanie modeli drzew decyzyjnych oraz regresji logistycznej do analizy czynników demograficznych oraz społeczno-ekonomicznych wpływających na szanse znalezienia pracy*, Studia Ekonomiczne, 95, 271-277.
[13] GUS, (2010), *Monitoring Rynku Pracy*, Wejście ludzi młodych na rynek pracy.
[14] Ibrahim J.G., Chen M.-H., Sinha D., (2001), *Bayesian Survival Analysis*, Springer-Verlag, New York.
[15] Kryńska E. (red.), (2004), *Polski rynek pracy – niedopasowania strukturalne*, IPiSS, Warszawa.
[16] Kwiatkowski E., (2007), *Bezrobocie*, Podstawy teoretyczne, WN, Warszawa.
[17] Lancaster T., (1979), *Econometric Methods for the Duration of Unemployment*, Econometrica, 47, 939-956.
[18] Merrick J.R., Soyer R., Mazzuchi A., (2002), *A Bayesian Semi-parametric Analysis of the Reliability and Maintenance of Machine Tools*, Technometrics 48, 58-69.
[19] Osiewalski J., (2001), *Ekonometria bayesowska w zastosowaniach*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
[20] Pipień M., (2006), *Wnioskowanie bayesowskie w ekonometrii finansowej*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
[21] Sinha D., Dey D.K., (1997), *Semiparametric Bayesian Analysis of Survival Data*, Journal of the American Statistical Association, 92, 1195-1212.
[22] Sinha D., Ibrahim J.G., Chen, M.-H., (2003), *A Bayesian Justification of Cox's Partial Likelihood*, Biometrika, 90(3), 629-641.
[23] Socha M., Sztanderska U., (2000), *Strukturalne podstawy bezrobocia w Polsce*, PWN, Warszawa.
[24] Szreder M., (1994), *Informacje a priori w klasycznej i bayesowskiej estymacji modeli regresji*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

BADANIE DETERMINANT POZOSTAWANIA BEZ PRACY OSÓB MŁODYCH Z WYKORZYSTANIEM SEMIPARAMETRYCZNEGO MODELU COXA

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

Obecnie wśród osób rozpoczynających karierę zawodową obserwuje się szczególnie dużą wartość wskaźnika bezrobocia. Celem niniejszego opracowania jest identyfikacja czynników demograficznych oraz społeczno-ekonomicznych wpływających na długość czasu pozostawania bez pracy tych osób. W badaniu wykorzystano m.in. bayesowski semiparametryczny model Coxa dla danych indywidualnych. Wykorzystanie modelu przeżycia daje możliwość analizy jednoczesnego wpływu wybranych zmiennych objaśniających na czas pozostawania bez pracy. Natomiast podejście bayesowskie umożliwia uwzględnienie w badaniu, za pomocą rozkładów a priori, dodatkowej informacji spoza próby. Estymację modeli przeprowadzono z wykorzystaniem metod Monte Carlo opartych na łańcuchach Markowa, a dokładniej algorytmu ARMS.

Słowa kluczowe: bezrobocie, semiparametryczny model Coxa, wnioskowanie bayesowskie, metody MCMC

AN ANALYSIS OF UNEMPLOYMENT DURATION DETERMINANTS AMONG YOUNG PEOPLE USING SEMIPARAMETRIC COX MODEL

A b s t r a c t

High unemployment rates are observed among people beginning job careers nowadays. The aim of the work is to identify demographic and socio-economic factors influencing the unemployment duration in this age group. In this research, Bayesian semiparametric Cox model for individual data has been used. The advantage of survival model is the possibility of the analysis of the impact of selected independent variables on unemployment duration. The Bayesian approach with a priori distribution makes the use of out of the sample knowledge possible. The model has been estimated using Markov chain Monte Carlo method with ARMS algorithm.

Key words: unemployment, semiparametric Cox model, Bayesian inference, Markov chain Monte Carlo method

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

MARIA SZMUKSTA-ZAWADZKA, JAN ZAWADZKI

Z BADAŃ NAD METODAMI PROGNOZOWANIA NA PODSTAWIE NIEKOMPLENTYCH SZERGÓW CZASOWYCH Z WAHANIAMI OKRESOWYMI (SEZONOWYMI)

1. SYNTEZA WYNIKÓW BADAŃ NAD METODAMI PROGNOZOWANIA BRAKUJĄCYCH DANYCH SEZONOWYCH

Przystępując do modelowania zmiennych ekonomicznych w wielu przypadkach spotykamy się z sytuacją, gdy w szeregach czasowych, zwłaszcza dla danych o krótszych niż rok okresach jednostkowych, występują luki. W świetle jednego z klasycznych założeń dotyczących modelowania ekonometrycznego byłoby to jednoznaczne z rezygnacją z jego przeprowadzenia. Jednak jak wykazują wyniki wieloletnich badań autorów zawarte m.in. w pracach Szmuksta-Zawadzka, Zawadzki (2002a, 2002b, 2002c, 2003) oraz jako rozdziały w monografiach pod redakcją Zawadzkiego (1999, 2003), przy spełnieniu określonych założeń dotyczących m.in. skali wahań sezonowych, odsetka brakujących danych czy sekwencyjności luk, możliwe jest modelowanie i prognozowanie zmiennych w warunkach braku pełnej informacji pozwalające na otrzymanie prognoz dopuszczalnych. Pojęcie prognozy dopuszczalnej może być rozpatrywane w różnorodny sposób[1]. W pracy za dopuszczalne uważać będziemy takie prognozy, których spodziewana dokładność wyrażona w liczbach absolutnych lub względnych dla założonego horyzontu prognozy jest akceptowalna przez użytkownika.

Do budowy prognoz zmiennych ekonomicznych wykazujących wahania sezonowe mogą być wykorzystywane zarówno metody dla danych oryginalnych (z sezonowością) jak i danych oczyszczonych z sezonowości. Poniżej przedstawione zostaną metody wykorzystujące te dwa rodzaje danych. Z uwagi na to, że większość niżej wymienionych metod prognozowania na podstawie kompletnych szeregów czasowych posiada bogatą literaturę, ograniczymy się jedynie do wskazania w kolejności alfabetycznej autorów ważniejszych prac z tego zakresu. Więcej uwagi poświęcimy natomiast tym metodom,

[1] Zdaniem Zeliasia (1997) sposoby definiowania kryteriów dopuszczalności zależą m.in. od charakteru zmiennej prognozowanej i warunków predykcji oraz stopnia pewności prognozy. W pracy Dittmann i inni (2009) stopień niepewności (dopuszczalności) należy rozpatrywać z punktu widzenia: błędów *ex ante*, wiarygodności prognozy, błędów *ex post* oraz oceny słownej. Najobszerniej na ten temat wypowiada się Guzik (2003) poświęcając cały rozdział swojej pracy. Formułuje dwie zasadnicze klasy kryteriów: μ-kryterium oraz ω – kryterium. Pierwsze z nich odnosi się do miernika niedokładności prognozy a drugie do wiarygodności. Następnie rozpatruje to pojęcie w aspekcie: miernika błędu prognozy, przedziału ufności, funkcji sygnalnych oraz horyzontu prognozy. Autorzy przywołanych wyżej prac na pierwszy plan wysuwają spełnianie przez nie kryterium dokładności ustalonego przez użytkownika lub prognostyka.

których stosowanie łączy się ze znacznym stopniem komplikacji w przebiegu procesu modelowania. Odnosić się to będzie zwłaszcza do przypadku występowania systematycznych luk w danych. Będziemy odwoływać się także do wybranych prac, w których prezentowane są wyniki modelowania i prognozowania zmiennych wykazujących wahania sezonowe warunkach braku pełnej informacji.

Do pierwszej grupy metod, wykorzystujących dane oryginalne, można zaliczyć:

a) klasyczne modele szeregu czasowego z wahaniami sezonowymi (Czerwiński, Guzik, 1980; Dittmann, 2008; Zawadzki (red.), 1999),
b) modele hierarchiczne szeregu czasowego (Zawadzki (red.), 2003; Szmuksta- Zawadzka, Zawadzki, 2000, 2002a, 2002b,2002c, 2003),
c) modele adaptacyjne uwzględniające wahania sezonowe (np. Holta-Wintersa) (Dittmann, 2008; Pawłowski, 1973, 1982; Zeliaś, 1997; Zeliaś, Pawełek, Wanat, 2003)
d) modele sieci neuronowych (podstawy metodologiczne metody sztucznych sieci neuronowych oraz przykłady ich zastosowań można znaleźć m.in. w pracach: Dittmanna, 2008; Gajdy, 2001; Luli, 1999; Markiewicz, Zawadzkiego, 2005; Witkowskiej, 2002)

Przebieg procesu modelowania związany z możliwościach wykorzystania metod wymienionych w punkcie a) i b) zależy przede wszystkim od rodzaju luk. W szeregach z wahaniami sezonowymi występować mogą luki niesystematyczne lub systematyczne. Z lukami niesystematycznymi mamy do czynienia wtedy, gdy dysponujemy przynajmniej jedną informacją dla każdego podokresu. Do prognozowania brakujących danych w warunkach występowania luk niesystematycznych można wykorzystać w zasadzie wszystkie wymienione wyżej modele. Jedynie w przypadku modeli wyrównywania wykładniczego konieczne jest dysponowanie określona liczbą danych początkowych.

Natomiast luki systematyczne występują wtedy, gdy nie mamy informacji przynajmniej dla jednego podokresu (miesiąca, dekady, dnia itp.) we wszystkich latach okresu estymacyjnego. Następstwem występowania tego rodzaju luk jest niemożność użycia części modeli, a w przypadku innych łączy się to ze znaczną komplikacją przebiegu procesu modelowania i prognozowania, polegającą na szacowaniu dużej liczby równań o tych samych własnościach predyktywnych.

W skład drugiej grupy metod wchodzą:

a) klasyczne modele trendu (Cieślak, 1997; Czerwiński, Guzik, 1980; Zeliaś, 1997),
b) modele adaptacyjne bez wahań sezonowych (np. Browna i Holta) (Dittmann, 2008; Pawłowski, 1973,1982; Zeliaś, 1997; Zeliaś, Pawełek, Wanat, 2003),
c) metody numeryczne (np. odcinkowa, łuków, wielomianowa Lagrange'a) (Fortuna, 2001; Grabiński, Wydymus, Zeliaś, 1979; Zboś, 1995).

Modelowanie i prognozowanie dla danych oczyszczonych z sezonowości przebiega w trzech etapach. W etapie pierwszym dokonuje się eliminacji wahań sezonowych ze zmiennej prognozowanej. Na podstawie tak otrzymanych danych przeprowadzane jest szacownie parametrów odpowiednich modeli na podstawie, których następnie wyznaczane są prognozy inter- i ekstrapolacyjne. Końcowym etapem jest ich przemnożenie przez wskaźniki sezonowości.

Za pomocą niektórych metod numerycznych (np. odcinkowej i łuków) można wyznaczać tylko prognozy interpolacyjne. Prognozy ekstrapolacyjne wyznacza się dla „pełnego" szeregu (uzupełnionego o prognozy interpolacyjne) za pomocą np. modeli trendu lub modeli adaptacyjnych.

Z wieloletnich badań empirycznych zawartych m.in. w przywołanych wyżej pracach autorów, związanych ze stosowaniem różnych metod wynika, że dokładność prognoz otrzymywanych na podstawie szeregów z wahaniami sezonowymi zawierających luki w danych zależy od:

– rodzaju i skali wahań sezonowych,
– struktury harmonicznej zmiennych,
– rodzaju luk w danych,
– regularności w przebiegu zmiennych,
– liczby i rozmieszczenia luk względem minimów i maksimów sezonowych oraz od sekwencyjności luk,
– długości okresu jednostkowego,
– horyzontu prognozy.

Z doświadczeń, o których mowa, wynika także, że prognozy otrzymane na podstawie modeli o najlepszych własnościach predyktywnych dość często nie charakteryzują się minimalnymi ocenami błędów. Ponadto minimalne oceny błędów prognoz interpolacyjnych otrzymuje się na podstawie innych modeli niż prognoz ekstrapolacyjnych.

W tabeli poniżej przedstawiona została syntetyczna charakterystyka wymienionych wyżej metod w aspekcie prognozowania brakujących danych. Dokonując tabelarycznego zestawienia metod wykorzystywanych w prognozowaniu w warunkach braku pełnej informacji, brano pod uwagę: rodzaj luk, możliwość wyznaczania albo obu rodzajów prognoz albo tylko prognoz interpolacyjnych, liczbę obserwacji początkowych oraz liczby potencjalnych modeli, spośród których dokonuje się wyboru dla celów prognozowania.

2. PRZYKŁAD EMPIRYCZNY MODELOWANIE I PROGNOZOWANIA ZMIENNEJ EKONOMICZNEJ

Ilustracją przeprowadzonych powyżej rozważań wynikających z wieloletnich badań nad wykorzystaniem metod prognozowania brakujących informacji będzie przykład empiryczny dotyczący prognozowania miesięcznych wielkości skupu mleka (w tys. l) dla luk systematycznych występujących w II i IV kwartale każdego roku.

Przedział czasowy „próby" obejmował lata 1999-2004, przy czym rok 2004 będzie rokiem empirycznej weryfikacji prognoz. Luki w danych otrzymano przez „wymazanie" odpowiednich danych z pełnego szeregu.

Na rysunku 1 przedstawiono kształtowanie się wielkości skupu mleka według miesięcy.

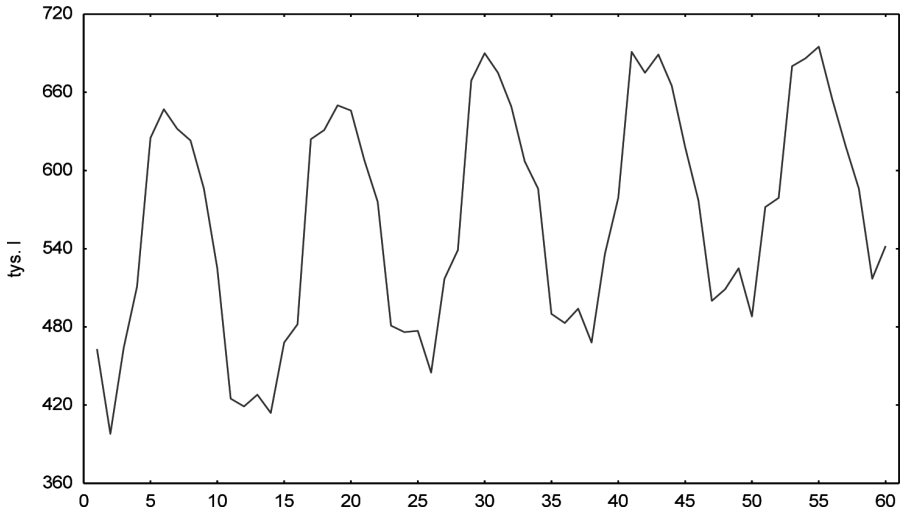
Tabela 1.

Syntetyczne zestawienie własności metod prognozowania brakujących danych sezonowych

Lp	Metody(modele)	Rodaj luk	Rodzaj prog.	L.obs.początkowych	Liczba modeli
I	Dane oryginalne(z sezonowością)				
1	Klasyczne ze zmiennymi zerojedynkowymi	NS	I+E		1
2	Klasyczne z wielomianem trygonometrycznym.	NS+S	I+E		NS -1 S-zależy od układu luk i str. harmonicznej
3	Hierarchiczne	NS+S	I+E		NS -7 S-zależy od układu luk i str. harmonicznej.
4	Holta - Wintersa	NS	I+E	13	Kilkanaście wybranych z 729 (dla param. wygładzania zmieniających się co 0,1)
5	Sieci neuronowych	NS+S	I+E		Kilka (kilkanaście) wybranych z kilkuset
II	Dane oczyszczone z sezonowości				
	A. wyrównywania wykładniczego				
6	Browna -prosty	NS	I+E	1	Kilka wybranych z 9 (dla parametrów wygładzania zmieniających się co 0,1)
7	Browna -liniowy	NS	I+E	2	Kilka wybranych z 9 (dla parametrów wygładzania zmieniających się co 0,1)
8	Browna -kwadratowy	NS	I+E	3	Kilka wybranych z 9 (dla parametrów wygładzania zmieniających się co 0,1)
9	Holta - liniowy	NS	I+E	2	Kilkanaście wybr. z 81 (dla parametrów wygładzania zmieniających się co 0,1)
	B. Metody numeryczne				
10	Odcinkowa/MNK(WW)	NS+S	I	1	1
11	Łuków I/MNK(WW)	NS+S	I	2	1
12	Łuków II/MNK(WW)	NS+S	I	1	1
13	Wielomianowa Lagrange'a	NS+S	I+E	1	1
14	Wielomianowa Lagrange'a z opt. Czebyszewa 3	NS+S	I+E	1	1
15	MNK - metoda trendu	NS+S	I+E		Zależy od liczby postaci anal. trendu

Objaśnienia skrótów: NS - luki niesystematyczne; S-luki systematyczne; I- prognozy interpolacyjne; E - prognozy ekstrapolacyjne.

Źródło: opracowanie własne

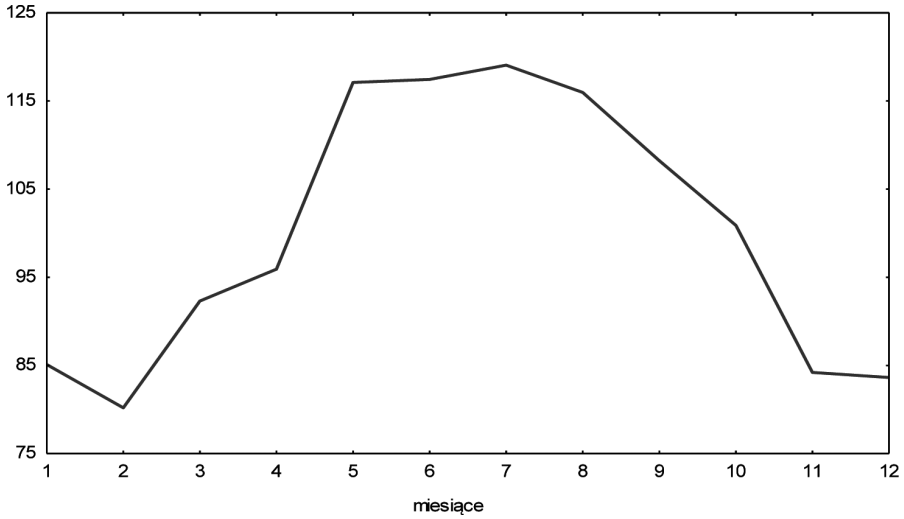


Rysunek 1. Skup mleka w latach 1999-2004

Źródło: opracowanie własne

Z powyższego rysunku wynika, że badana zmienna charakteryzuje się występowaniem dość silnych wahań sezonowych.

Natomiast na rysunku 2 przedstawiono kształtowanie się wskaźników sezonowości. Informacje zawarte na rysunku wskazują, że zmienna prognozowana charakteryzuje się dość silnym natężeniem wahań sezonowych, ponieważ amplituda ocen wskaźników sezonowości wynosi ok. 40 punktów procentowych.



Rysunek 2. Kształtowanie się wskaźników sezonowości

Źródło: opracowanie własne

W celu obliczenia udziałów poszczególnych składowych harmonicznych w wyjaśnieniu wariancji sezonowej najpierw został oszacowany model szeregu czasowego z liniowym trendem i periodycznym składnikiem sezonowym a następnie dokonano eliminacji trendu.

W tabeli 1 zestawione zostały udziały procentowych składowych harmonicznych oraz harmonik w wyjaśnianiu wariancji sezonowej zmiennej prognozowanej (składowe sinusoidalne oznaczono przez S_i a kosinusoidalne przez C_i).

Tabela 2.

Procentowe udziały składowych harmonicznych sinuso- i kosinusoidalnych oraz harmonik w wyjaśnianiu wariancji sezonowej (w %)

Składowe sinusoidalne	**Odsetki**	**Składowe kosinusoidalne**	**Odsetki**	**Suma**
S_1	18,17	C_1	75,89	94,05
S_2	0	C_2	0,8	0,8
S_3	2,06	C_3	0,10	2,17
S_4	0,09	C_4	0,03	0,12
S_5	2,20	C_5	0,28	2,48
–	–	C_6	0,38	0,38
Suma	22,53		77,47	

Źródło: obliczenia własne

Z informacji zawartych w tabeli wynika, że najwyższym udziałem w wyjaśnianiu ogólnej wariancji sezonowej charakteryzuje się składowa kosinusoidalna o cyklu 12 miesięcznym (75,89%) a następnie składowa sinusoidalna o takiej samej długości cyklu (18,17%). Pierwsza harmonika wyjaśnia zatem w ponad 94 procentach całkowitą zmienność sezonową. Ponadto wskaźniki udziałów jeszcze tylko dwóch składowych sinusoidalnych (S_3 i S_5) o cyklach wynoszących odpowiednio: 4 oraz 2,4 miesiąca przekraczają 2 procent. Łączna wielkość udziałów pozostałych 7 składowych wynosi 1,69%, przy czym dla wszystkich składowych wskaźniki udziałów są niższe od 1 procenta.

Rozpatrywany będzie jeden wariant luk systematycznych, będą one występować w II i IV kwartale każdego roku. Luki te otrzymano przez „wymazanie" odpowiednich danych z pełnego szeregu.

W prognozowaniu brakujących danych zostaną wykorzystane następujące metody:

I. Dla danych oryginalnych (z sezonowością):
   1) klasyczny model szeregu czasowego z liniowym trendem i periodycznym składnikiem sezonowym opisanym za pomocą wielomianu trygonometrycznego,
   2) modele hierarchiczne,
   3) model Holta –Wintersa,
   4) sieci neuronowe,

II. Dla danych oczyszczonych:
5) metoda trendu(MNK),
6) model prosty Browna,
7) model liniowy Browna,
8) model kwadratowy Browna,
9) model liniowy Holta (dwuparametryczny),
10) metoda odcinkowa/MNK
11) metoda łuków I/MNK,
12) metoda łuków II/MNK,
13) metoda Lagrange'a dla 3 i 4 węzłów interpolacyjnych rozmieszczonych w równych odległościach ,
14) metoda Lagrange'a z optymalizacją Czebyszewa z 3 i 4 węzłami.

Zanim przejdziemy do zestawienia tabelarycznego wyników prognozowania inter- i ekstrapolacyjnego dla najlepszych predyktorów, przedstawimy kolejno wyniki modelowania oraz prognozowania dla metod wymienionych w punktach: 1,3,6-9 oraz 11-14.

W przypadku metody klasycznej z wielomianem trygonometrycznym, występowanie luk systematycznych oznacza m.in., że macierz współczynników korelacji składowych harmonicznych przestaje być macierzą jednostkową. W naszym przypadku przyjęła ona postać:

Tabela 3.

Oceny współczynników korelacji liniowej składowych harmonicznych dla rozpatrywanego wariantu luk w danych

	S_1	S_2	S_3	S_4	S_5	C_1	C_2	C_3	C_4	C_5
S_1	1,00	0,00	0,50	0,00	-0,50	-0,61	0,00	0,61	0,00	0,00
S_2	0,00	1,00	0,00	0,00	0,00	0,00	0,76	0,00	1,00	0,00
S_3	0,50	0,00	1,00	0,00	0,50	-0,61	0,00	0,00	0,00	0,61
S_4	0,00	0,00	0,00	1,00	0,00	0,00	-0,65	0,00	0,00	0,00
S_5	-0,50	0,00	0,50	0,00	1,00	0,00	0,00	-0,61	0,00	0,61
C_1	-0,61	0,00	-0,61	0,00	0,00	1,00	0,00	0,25	0,00	0,25
C_2	0,00	0,76	0,00	-0,65	0,00	0,00	1,00	0,00	0,76	0,00
C_3	0,61	0,00	0,00	0,00	-0,61	0,25	0,00	1,00	0,00	0,25
C_4	0,00	1,00	0,00	0,00	0,00	0,00	0,76	0,00	1,00	0,00
C_5	0,00	0,00	0,61	0,00	0,61	0,25	0,00	0,25	0,00	1,00

Źródło: obliczenia własne

Z powyższej tabeli wynika, że 14 elementów leżących powyżej i poniżej głównej przekątnej jest różnych od zera, przy czym cztery z nich są ujemne. Otrzymanie elementów różnych od zera może oznaczać wystąpienie liniowych kombinacji składowych.

Wyboru statystycznie istotnych składowych harmonicznych dokonano za pomocą metody *selekcji a priori*, będącej jedną z procedur regresji krokowej. Polega ona, najogólniej rzecz biorąc, na wprowadzaniu do modelu kolejnej składowej (oprócz pierwszej) najsilniej skorelowanej z resztami modelu otrzymanego w poprzednim etapie. Opis tej metody można znaleźć m.in. w pracy Drapera, Smitha (1973).

W trakcie wyboru składowych okazało się, że kombinacje liniowe tworzą: para składowych: S_3 i C_1 oraz trójka: S_5, C_3 oraz C_5. Oznacza to, że do modelu można wprowadzić tylko po jednej zmiennej wchodzącej w skład pary lub trójki. Tym samym więc szacowanych będzie 6 wersji modelu – zostały one zestawione w poniższej tabeli. Zawiera ona także oceny błędów prognoz interpolacyjnych oraz prognoz ekstrapolacyjnych dla horyzontu h=3, 6, 9 i 12 miesięcy.

Tabela 4.

Oceny parametrów struktury stochastycznej modeli z wielomianem trygonometrycznym oraz błędy prognoz inter- i ekstrapolacyjnych

Lp	**Składowe**	**R^2**	**S_e**	**V_{Se}(%)**	**MAPE [%]**				
			(tys.l)		**inter.**	**h=3**	**h=6**	**h=9**	**h=12**
1	S145C1	0,98	13,05	2,37	5,42	3,65	3,53	3,66	4,22
2	S14C13	0,98	13,05	2,37	6,97	3,65	5,17	4,76	4,53
3	S14C15	0,98	13,05	2,37	6,38	3,65	5,17	4,76	5,32
4	S1345	0,98	13,05	2,37	22,48	3,65	10,85	8,54	12,98
5	S134C3	0,98	13,05	2,37	21,47	3,65	10,18	8,1	12,29
6	S134C5	0,98	13,05	2,37	22,17	3,65	10,64	8,4	12,81

Źródło: obliczenia własne

W kolumnie drugiej podane zostały zestawy składowych wchodzących w skład poszczególnych wersji modelu. Z informacji zawartych w kolejnych trzech kolumnach wynika, że wersje te charakteryzują się identycznymi ocenami: współczynników determinacji (R^2), odchyleń standardowych składników losowych (S_e) oraz współczynników zmienności losowej (V_{Se}). Oznacza to, że są one nierozróżnialne z punktu widzenia własności predyktywnych. Mogą się natomiast różnić ocenami błędów prognoz inter- lub/i ekstrapolacyjnych. Oceny względnych błędów prognoz (MAPE) podane zostały w ostatnich pięciu kolumnach.

W kształtowaniu się ocen błędów prognoz interpolacyjnych w sposób zdecydowany wyodrębnić można dwie grupy predyktorów. Do pierwszej grupy należą predyktory wymienione w wierszach od drugiego do czwartego. Charakteryzują się one tym, że zawierają jednocześnie składowe C_1 oraz S_1 tj. składowe o największych udziałach w wyjaśnianiu wariancji sezonowej. Oceny błędów prognoz kształtują się w granicach od 5,42 do 6,97 procent. Oceny błędów dla pozostałych trzech predyktorów, zaliczonych do grupy drugiej, przyjmują wartości znacznie wyższe i zawierają się w przedziale

21,47-22,48 procent. Spośród pary składowych o najwyższych udziałach występuje tylko składowa S_1.

W przypadku prognoz ekstrapolacyjnych wyodrębnienie tych grup jest możliwe dla horyzontu prognozy wynoszącego co najmniej 6 miesięcy, ponieważ dla h=3 otrzymano identyczne oceny błędów wynoszące 3,65 procent. Dokładność predyktorów należących do grupy pierwszej jest od około 4 do około 8 punktów procentowych wyższa od dokładności predyktorów należących do grupy drugiej.

Większa dokładność predyktorów zaliczonych do grupy pierwszej wynika, podobnie jak to miało miejsce w przypadku prognoz interpolacyjnych, z faktu, że zawierają one składową kosinusoidalną o cyklu dwunastomiesięcznym (C_1). Udział jej w wyjaśnieniu wariancji sezonowej dla pełnych danych, o czym była mowa wyżej, przekracza 75 procent.

W tabeli 4 zestawiono oceny błędów prognoz otrzymanych za pomocą metod numerycznych. Z uwagi na fakt, że za pomocą metody odcinkowej oraz metody łuków można wyznaczyć tylko prognozy interpolacyjne, prognozy ekstrapolacyjne wyznaczono MNK szacując trendy liniowe na podstawie szeregów z lukami uzupełnionych o prognozy interpolacyjne.

Tabela 5.

Oceny błędów prognoz inter- i ekstrapolacyjnych otrzymanych za pomocą metod numerycznych

Metoda	**MAPE [%]**				
	interpol.	**h=3**	**h=6**	**h=9**	**h=12**
Odcinkowa/MNK	1,88	5,12	3,45	2,68	4,13
Łuków I/MNK	1,41	5,99	4,73	4,18	6,18
Łuków II/MNK	2,04	5,34	3,75	2,94	4,51
Lagrange'a WP 3	3,29	2,26	5,57	8,48	8,66
Lagrange'a WP 4	4,75	33,87	40,09	48,2	58,34
Lagrange'a z opt. Czebyszewa 3	3,89	7,8	5,71	4,24	4,99
Lagrange'a z opt. Czebyszewa 4	3,04	15,46	16,97	19,62	24,63

Źródło: obliczenia własne

W przypadku metody Lagrange'a z 4 węzłami interpolacyjnymi rozmieszczonymi w równych odległościach wyraźnie widoczny jest wynik oddziaływania tzw. efektu Rungego, polegającego na pogorszeniu się jakości interpolacji wielomianowej wraz ze zwiększeniem liczby węzłów. W naszym przypadku odnosi się to do prognoz ekstrapolacyjnych.

Dla metody Lagrange'a z optymalizacją Czebyszewa dla 4 węzłów efekt ten występuje w skali znacznie słabszej, ale i tak oceny błędów są zdecydowanie wyższe niż ma to miejsce dla 3 węzłów.

W tabeli 5 zestawione zostały modele (wersje modeli) charakteryzujące się najniższymi wartościami ocen błędów prognoz inter- i ekstrapolacyjnych. Dlatego też niektóre metody (modele) zwłaszcza adaptacyjne są reprezentowane przez dwa a nawet trzy predyktory, różniące się stałymi wygładzenia lub w przypadku modelu ze składowymi harmonicznymi zestawem składowych harmonicznych. W kolumnie drugiej podane zostały liczby modeli, spośród których zostały one wybrane. „Tłustym" drukiem zaznaczone zostały predyktory o najniższych ocenach: mierników wartości predyktywnych, błędów prognoz interpolacyjnych lub ekstrapolacyjnych. Ponadto w dwóch ostatnich kolumnach zestawione zostały miejsca (rangi) poszczególnych predyktorów uszeregowane według rosnących ocen błędów prognoz inter- i ekstrapolacyjnych.

W pierwszych sześciu wierszach zestawione zostały predyktory na podstawie, których wyznaczano prognozy dla danych oryginalnych (z sezonowością). Z analizy mierników charakteryzujących własności predyktywne wynika, że najmniej korzystne ich oceny otrzymano dla sieci neuronowych.

Oceny błędów prognoz interpolacyjnych dla pięciu spośród sześciu predyktorów przyjęły wartości z przedziału 5,42 do 7,08 procent. Znacznie *in minus* od nich odbiega ocena tego błędu dla predyktora hierarchicznego – jest ona ponad dwukrotnie wyższa. Zróżnicowanie ocen błędów prognoz ekstrapolacyjnych jest znacznie niższe. Dla pięciu predyktorów zawierają się one w przedziale od 3,27 do 6,12 procent. Jedynie w przypadku modelu hierarchicznego błąd ten przekracza 8 procent. Oznacza to tym samym, że sekwencyjny układ systematycznych luk w danych, obejmujący miesiące wchodzące w skład 2 i 4 kwartału, najsilniej oddziaływał w przypadku modeli hierarchicznych. Najbardziej stabilny z punktu widzenia względnych ocen obu rodzajów błędów okazał się model klasyczny z wielomianem trygonometrycznym zajmując odpowiednio 1 i 2 miejsce wśród modeli szacowanych dla danych oryginalnych (sezonowością). Następnym w kolejności okazał się model Holta-Wintersa.

Z analizy rang metod dla błędów prognoz interpolacyjnych wynika, że modele dla danych oryginalnych zajmują 6 spośród 7 miejsc w dole tabeli – zostały one jedynie przedzielone liniowym modelem Holta sklasyfikowanym na 20 miejscu.

Natomiast w przypadku prognoz ekstrapolacyjnych mamy do czynienia z bardzo dużym zróżnicowaniem miejsc (rang). Najlepszą metodą spośród 22 analizowanych predyktorów okazała się sieć neuronowa z 13 warstwami ukrytymi. W przypadku modeli sieci neuronowych (S3MLP) zaobserwowano największe zróżnicowanie rang ocen błędów. Otrzymano dla niego najniższą oceną błędu prognoz ekstrapolacyjnych i najwyższą ocenę błędu prognoz interpolacyjnych. W przypadku modelu MLP1 różnica rang była niewielka, ale został on sklasyfikowany na odpowiednio: na 18 i 16 miejscu. Prezentowane w tabeli modele sieci neuronowych, wybrane spośród 400 wcześniej oszacowanych, charakteryzowały się względnie stabilną relacją błędów prognoz inter- i ekstrapolacyjnych. W wielu przypadkach modele o niższych ocenach błędów prognoz interpolacyjnych dawały bardzo duże błędy prognoz ekstrapolacyjnych i odwrotnie. Predyktor klasyczny z wielomianem trygonometrycznym został sklasyfikowany na 8 miejscu. Miejsce 15 zajął model Holta – Wintersa charakteryzujący się najniższą oceną

Tabela 6.
Zestawienie wyników modelowania i prognozowania inter- i ekstrapolacyjnego oraz rangowania metod

Metody(modele)	L. mod.	Stałe wygładzania			MAPE wart.wyr. lub V_{Se}**	Błędy prognoz		Ranga metody prognozowania	
		α	β	γ		interpol.	Ekstrapol. h=12	inter.	ekstr.
Klasyczne z wiel. tryg.(s145c1)	6				2,37**	5,42	4,22	16	8
Hierarchiczny(H43)*	14				2,34**	15,02	8,17	22	18
Holta Wintersa	13	0,1	0,1	0,5	**2,26**	**5,56**	5,51	17	15
Holta Wintersa	13	0,3	0,3	0,8	4,38	6,58	**4,96**	19	12
Sieci neuronowe (S3-MLP3)*	6				**5,61****	7,08	**3,27**	21	1
Sieci neuronowe (S2-MLP1)*	6				6,58**	**5,8**	6,12	18	16
Browna -prosty	9	0,9			**2,21**	3,63	5,33	11	14
Browna -prosty	9	0,4			2,66	**3,32**	4,51	9	10
Browna -prosty	9	0,3			2,78	3,4	**4,42**	10	9
Browna -liniowy	9	0,2			**2,82**	**2,95**	3,75	7	6
Browna -liniowy	9	0,1			2,97	3,66	**3,56**	12	2
Browna -kwadratowy	8	0,1			**3,09**	4,14	**3,7**	15	5
Browna -kwadratowy	8	0,2			3,59	**2,88**	10,9	4	21
Holta-liniowy	18	0,9	0,9		**3,64**	**4,11**	13,03	14	22
Holta-liniowy	18	0,3	0,4		8,27	6,83	**8,79**	20	20
Odcinkowa/MNK	1					1,88	4,13	2	7
Łuków I/MNK	1					1,41	6,18	1	17
Łuków II/MNK	1					2,04	4,51	3	11
Lagrange'a WP 3	2					3,29	8,66	8	19
Lagrange'a z opt. Czebyszewa 3	2					3,89	4,99	13	13
MNK-trend liniowy	1				2,78**	2,95	3,66	6	3
MNK-trend wykładn. o stałej stopie wzrostu	1				2,82**	2,95	3,69	5	4

Źródło: opracowanie własne oraz (*) Markiewicz, Zawadzki, (2005)
(**) współczynnik zmienności losowej

błędu interpolacyjnego. Natomiast predyktor hierarchiczny sklasyfikowany został na 18 miejscu.

Obecnie przechodzimy do analizy własności predyktywnych mierników dokładności prognoz dla danych oczyszczonych z sezonowości. Najpierw analizie poddane zostaną predyktory oparte na modelach Browna i liniowym modelu Holta, a następnie

predyktory wykorzystujące metody numeryczne. Dla trzech modeli wyrównania wykładniczego otrzymano po 2 predyktory charakteryzujące się minimalnymi ocenami własności predyktywnych i jednego rodzaju prognoz. Dla prostego modelu Holta były to nawet 3 wersje modelu różniące się stałymi wygładzenia.

Zróżnicowanie błędów wartości wyrównanych dla 8 z 9 modeli było niewielkie – oceny te mieściły się w przedziale od 2,21 do 3,64 procent. Znacznie wyższą ocenę otrzymano dla liniowego modelu Holta o stałych wygładzenia α i β wynoszących odpowiednio: 0,3 i 0,4. Dla modelu tego otrzymano minimalną ocenę błędu prognoz ekstrapolacyjnych. W przypadku prognoz interpolacyjnych dla 8 wersji modeli oceny te mieściły się w przedziale 2,95 do 4,14 procent. Natomiast dla wymienionego wyżej modelu błąd ten był wyższy od najwyższego o blisko 2,7 punktu procentowego.

Oceny błędów prognoz ekstrapolacyjnych otrzymanych na podstawie prostych i liniowych modeli Browna kształtują się w przedziale od 3,56 do 5,33 procent. Dla modelu Holta z minimalną oceną błędu tego rodzaju prognoz, ocena ta jest o około 3,5 punktu procentowego wyższa. Oceny błędów prognoz przekraczające 10 procent otrzymano dla modelu kwadratowego i modelu Holta, charakteryzujących się minimalnymi ocenami błędów prognoz interpolacyjnych.

Z kształtowania się rang błędów prognoz interpolacyjnych wynika, że mieszczą się one na ogółw grupie środkowej od 7 – 15 miejsca. Wysoką czwartą pozycję zajmuje kwadratowy model Browna dla stałej wygładzenia równej 0,2 a dwudziestą model Holta o minimalnej ocenie błędu prognoz ekstrapolacyjnych.

W przypadku rang ocen błędów prognoz ekstrapolacyjnych występuje większe zróżnicowanie – od 2, 5 i 6 miejsca dla liniowego i kwadratowego modelu Browna do 21 i 22 miejsca dla modeli Holta. Zwraca uwagę znaczne zróżnicowanie rang metod dla błędów prognoz inter- oraz ekstrapolacyjnych. Z największym zróżnicowaniem mamy do czynienia w przypadku modelu kwadratowego Browna sklasyfikowanym odpowiednio: na 4 i 21 miejscu. W przypadku liniowego i kwadratowego modelu Browna o minimalnych ocenach błędów prognoz różnica rang wynosi 10. Oznacza to, że inne predyktory powinny być wykorzystane w budowie prognoz interpolacyjnych oraz inne w przypadku prognozowania ekstrapolacyjnego.

Obecnie przechodzimy do analizy błędów prognoz oraz do rankingu metod numerycznych. W przypadku metody odcinkowej oraz łuków bezpośrednio za ich pomocą otrzymano prognozy interpolacyjne. Prognozy ekstrapolacyjne wyznaczono metodą MNK na podstawie liniowych równań trendów. W przypadku metody Lagrange'a wybrano metodę z trzema węzłami rozmieszczonymi proporcjonalnie lub zgodnie z metodą optymalizacji Czebyszewa. Prognozy ekstrapolacyjne wyznaczone na ich podstawie charakteryzują się brakiem omówionego wcześniej efektu Rungego.

Zróżnicowanie ocen błędów prognoz interpolacyjnych jest niewielkie i waha się od 1,42 do 3,89 procent, przy czym metody: odcinkowa i łuków zostały sklasyfikowane na trzech pierwszych miejscach. Natomiast oba warianty metody Lagrange'a odpowiednio: na 8 i 13 miejscu.

Oceny błędów prognoz ekstrapolacyjnych mieszczą się w przedziale od 4,13 dla metody odcinkowej/MNK do 8,66 dla metody Lagrange'a z 3 węzłami rozmieszczonymi proporcjonalnie. W klasyfikacji metod zajmują one miejsca: 7, 11, 13, 17 i 19. Z najmniejszą różnicą miejsc dla predyktorów o najniższym błędzie mamy do czynienia w przypadku metody odcinkowej/MNK. Natomiast najbardziej stabilny okazał się predyktor Lagrange'a z optymalizacją Czebyszewa, zajmując w obu klasyfikacjach 13 miejsce z ocenami błędów niższymi odpowiednio: od 4 i 5 procent.

Z informacji zawartych w dwóch ostatnich wierszach tabeli wynika, że oceny błędów prognoz otrzymanych na podstawie modeli trendu: liniowego i wykładniczego są identyczne dla prognoz interpolacyjnych oraz prawie takie same dla prognoz interpolacyjnych. W przypadku pierwszego rodzaju prognoz predyktory te zajmują miejsca: 6 i 5 a dla ekstrapolacyjnych: 3 i 4.

3. PODSUMOWANIE

Z przedstawionych wyżej analiz mierników charakteryzujących proces modelowania i prognozowania wynika, że różne metody (lub różne warianty (wersje)) charakteryzują się minimalnymi ocenami parametrów (własności predyktywnych) oraz prognoz inter- i ekstrapolacyjnych.

Oznacza to, że proces prognozowania *ex ante* zmiennych charakteryzujących się występowaniem co najmniej dość silnym natężeniem skali wahań sezonowych, powinien być poprzedzony analizą błędów zarówno prognoz interpolacyjnych jak i przede wszystkim błędów prognoz *ex post*. Predyktory charakteryzujące się minimalnymi ocenami błędów tych ostatnich prognoz powinny być wykorzystane do budowy prognoz *ex ante*. Nie należy wykluczyć wykorzystania metod „łamanych" tzn. szacowania modeli wykorzystanych w budowie prognoz *ex post,* które zostaną następnie będą wykorzystane do budowy prognoz *ex ante*.

Dokonując próby podsumowania rozważań przeprowadzonych w pracy możemy stwierdzić, że metody statystyczno-ekonometryczne mogą być z powodzeniem wykorzystywane w prognozowaniu brakujących danych zmiennych ekonomicznych o dość silnym natężeniu wahań sezonowych tzn. charakteryzujących się rozstępem skrajnych ocen wskaźników sezonowości nie przekraczających na ogół 40-60 punktów procentowych.

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

LITERATURA

[1] Cieślak M.(red.), (1997), *Prognozowanie gospodarcze,* PWN, Warszawa.
[2] Czerwiński Z., Guzik B., (1980), *Prognozowanie ekonometryczne*, PWE, Warszawa.
[3] Dittmann P., (2008), *Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie,* Walters Kluwer Polska.

[4] Dittmann P., Szabela-Pasierbińska E., Dittmann I., Szpulak A., (2011), *Prognozowanie w zarządzaniu sprzedażą i finansami przedsiębiorstwa,* Walters Kluwer Polska-Oficyna.
[5] Draper N. R., Smith H., (1973), Analiza regresji stosowana, PWN, Warszawa.
[6] Fortuna Z., (2001), *Metody numeryczne*, WNT, Warszawa.
[7] Gajda J., (2001), *Prognozowanie i symulacje a decyzje gospodarcze*, C.H.Beck, Warszawa.
[8] Grabiński T., Wydymus S., Zeliaś A., (1979), *Z badań nad metodami predykcji brakujących informacji*, Zeszyty Naukowe AE w Krakowie Nr 114, Kraków.
[9] Grabiński T., Wydymus S., Zeliaś A., (1981), *Modele ekonometryczne w procesie prognozowania*, Akademia Ekonomiczna w Krakowie, Kraków, str. 74-75.
[10] Guzik B.,(2003), *Wstęp do teorii prognozowania i symulacji*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.
[11] Lula P., (1999), *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*, Wyd. AE w Krakowie, Kraków.
[12] Markiewicz. A, Zawadzki J., (2005), *Dokładność prognoz skupu mleka dla systematycznych luk w danych*, Folia Universitatis Agriculturae Stetinensis, seria: Oeconomica 245, s. 411-416.
[13] Pawłowski Z., (1973), *Prognozy ekonometryczne*, PWN, Warszawa.
[14] Pawłowski Z., (1982), *Zasady predykcji ekonometrycznej*, PWN, Warszawa.
[15] Szmuksta-Zawadzka M., Zawadzki J., (1995), *O metodzie prognozowania brakujących informacji dla danych sezonowych*, Przegląd Statystyczny, nr 3-4, str. 377-385.
[16] Szmuksta-Zawadzka M., Zawadzki J., (2000), *On hierarchic models of time series with seasonal fluctuations*, Wyd. UMK, Toruń, str. 25-30.
[17] Szmuksta-Zawadzka M., Zawadzki J., (2002a), *Forecasting on basis of time series hierarchic models with variable seasonality*, [w:] Dynamics Econometrics Models No. 5, Toruń.
[18] Szmuksta-Zawadzka M., Zawadzki J., (2002b), *Prognozowanie na podstawie hierarchicznych modeli szeregu czasowego z lukami w danych*, w: Analiza szeregów czasowych na początku XXI wieku, Toruń, str. 93-102.
[19] Szmuksta-Zawadzka M., Zawadzki J., (2002c), *Hierarchiczne modele szeregów czasowych z wahaniami sezonowymi. Budowa. Estymacja. Prognozowanie.*, [w:] Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych, Wyd. AE, Kraków, str. 193-204.
[20] Szmuksta-Zawadzka M., Zawadzki J., (2003), *O modelach hierarchicznych dla danych dekadowych z wahaniami sezonowymi*, [w:] Dynamiczne modele ekonometryczne, Toruń.
[21] Witkowska D., (2002), *Sztuczne sieci neuronowe i metody statystyczne*, Wyd. C.H. Beck, Warszawa.
[22] Zawadzki J.(red.), (1999), *Ekonometryczne metody predykcji dla danych sezonowych w warunkach braku pełnej informacji,* Uniwersytet Szczeciński, Szczecin.
[23] Zawadzki J.(red.), (2003), *Zastosowanie hierarchicznych modeli szeregów czasowych w prognozowaniu zmiennych ekonomicznych z wahaniami sezonowymi*, Akademia Rolnicza, Szczecin.
[24] Zboś D., (1995), *Metody numeryczne*, Politechnika Krakowska, Kraków.
[25] Zeliaś A., (1987), *Estymacja parametrów równania regresji liniowej w warunkach niepełnej informacji*, Zeszyty Naukowe AE w Krakowie, Nr 243, Kraków.
[26] Zeliaś A., (1997), *Teoria prognozy*, PWE, Warszawa.
[27] Zeliaś A., Pawełek B., Wanat S., (2003), *Prognozowanie ekonomiczne. Teoria, przykłady, zadania.*, PWN, Warszawa.

Z BADAŃ NAD METODAMI PROGNOZOWANIA NA PODSTAWIE NIEKOMPLENTYCH SZERGÓW CZASOWYCH Z WAHANIAMI OKRESOWYMI (SEZONOWYMI)

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

Praca została poświęcona syntetycznemu omówieniu wyników wieloletnich badań autorów nad zastosowaniami metod prognozowania w warunkach braku pełnej informacji w szeregach czasowych z wahaniami sezonowymi. Rozważania odnosić się będą do dwóch rodzajów luk w danych: systematycznych i niesystematycznych. Z lukami systematycznymi mamy do czynienia wtedy, gdy nie są dostępne informacje liczbowe przynajmniej o jednym podokresie w całym przedziale czasowym „próby". Rozpatrywane będą metody prognozowania zarówno dla danych oryginalnych (z sezonowością) jak i danych, z których wyeliminowano wahania sezonowe. Egzemplifikacją rozważań o charakterze teoretycznym będzie przykład empiryczny.

Słowa kluczowe: szeregi czasowe, wahania sezonowe, brakujące dane, prognozowanie

STUDIES OF METHODS APPLIED TO FORECASTING INCOMPLETE DATA IN SEASONAL TIME SERIES

A b s t r a c t

This work presents discussion about results of long-term of authors research on applications of different forecasting methods in condition of lack of full information. There will be considered two types of gaps in data: systematic and unsystematic. The systematic gaps in data are only when we have not any information about at least one sub-period in the whole of analyzed data. There will be presented two types of methods applied to time series with and without seasonal component. Exemplification of theoretical considerations will be an empirical example.

Key words: time series, seasonal fluctuations, missing data, forecasting

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

TOMASZ KLIMANEK

USING INDIRECT ESTIMATION WITH SPATIAL AUTOCORRELATION IN SOCIAL SURVEYS IN POLAND[1]

1. BACKGROUND

First attempts at applying various approaches to parameter estimation for small areas in Poland were undertaken after the International Conference on Small Area Statistics held in Warsaw in 1992 (eds. Kalton, Kordos, Platek, 1993). There were only a few applications of small area estimation (SAE) methods to measure the scope of unemployment, poverty, household structure and in agriculture related surveys (Kordos, Paradysz, 2000). Further applications and examination of "standard" indirect estimators properties were undertaken within the EURAREA project[2] after the IASS Satellite Conference held in Riga in 1999. The first important study to apply SAE methodology in LFS was conducted in 2003. The authors (Bracha, Lednicki and Wieczorkowski, 2003) estimated totals and rates of several labour market characteristics by region, subregion and poviat (NUTS2, NUTS3 and NUTS4). They used direct, synthetic and composite estimators.

The second important study was conducted in 2004 by E. Gołata. The study was intended to rely on EURAREA project experiences. The Polish database - the so-called super-population labelled POLDATA - was created on the basis of 3 data sources: the 1995 Micro-census, the 1995 Household Budget Survey and the Local Data Bank. POLDATA represented as closely as possible the characteristic of Poland in 1995 with respect to the new administration division of the country which was introduced in January 1999. For the purposes of applying the standard estimators, the proportion of ILO unemployed (in the whole population over 15) in an area was estimated (Gołata, 2004).

One should also mention the study conducted in 2004 by Kubacki. The parameter of interest in the study was unemployment size for NUTS 2 and NUTS 4 level. Registered unemployment constituted the additional data source used in the study (with covariates: the number of unemployed persons, the number of employed persons, the

[1] The results presented in the paper are the outputs of Work Package 5 (Case studies) of EUROSTAT project ESSnet on SAE 61001.2009.003-2009.859 (2009-2012).

[2] The EURAREA project no. IST-2000-26290 entitled *Enhancing Small Area Estimation Techniques to meet European needs* is part of 5^{th} framework programme for research, technological development and demonstration of EU. Its main co-ordinator is ONS – Office for National Statistics, UK.

number of economically inactive persons, the number of dwellings and the number of persons aged 15 and above, Kubacki, 2004).

Both design and model based types of estimators were applied:

– design based estimators including post stratification methods (both ratio and regression estimator), synthetic estimator (both ratio and regression estimator),
– model based estimators include empirical Bayes (EB) estimator and hierarchical Bayes (HB) estimator.

Recent years have seen a growing interest in new possibilities and tools developed to meet the growing demand for estimates at local level. Special projects were carried in the Central Statistical Office (CSO) in cooperation with university researchers. The projects referred to different subjects, for example: labour market, household structure, business statistics including small business. From the perspective of the Population Census 2011, which was in progress at the time, special research was undertaken within the Central Statistical Office and newly established Centre for Small Area Statistics in Poznan. It was aimed at examining administration registers, their quality and usefulness as sources of auxiliary variables in small domain estimation. But practical application of SAE methods of official statistics in Poland is still not part of normal practice.

2. GENERAL SETUP

The aim of the Author was to continue explorations of estimating labour market characteristics for small domains. First, data infrastructure referring to economic activity and unemployment is presented and discussed. The intention was to include all variable categories, which experience has shown to be effective. In the case of ILO[3] unemployment the 'standard variables' are: age, sex, education, employment status and housing. In the study data from the following sources were used:

– Register of Unemployed,
– Vital Statistics Register,
– Tax Register data will be used in an indirect form via Commuting Survey which was based on data from Tax Register – the first edition of the survey is available for 2006.

Taking into account special features of the labour market in Poland, especially its high territorial differentiation, various estimation techniques will be analysed. Theoretical approaches to estimation with spatial effects proposed by R. Chambers & A. Saei (2003) together with the SAS software provided by EURAREA project are considered

[3] ILO unemployment – unemployment defined according to International Labour Organization, which is comparable among European and non-European countries. According to this definition the unemployed in general comprise all persons above a specified age who during the reference period were: without work, that is, were not in paid employment or self employment during the reference period; currently available for work, that is, were available for paid employment or self-employment during the reference period; and seeking work, that is, had taken specific steps in a specified recent period to seek paid employment or self-employment.

and adjusted to fit specific arrangements. The standard EURAREA estimators used in the study are: direct (for comparative reasons), generalised regression estimator – GREG, synthetic and EBLUP. Also EBLUPGREG, which takes into account spatial correlation structure, was applied.

The structure of economically active population and, especially, unemployment and its structure are of exceptional social interest in Poland. Unemployment, since the very beginning of 90's has assumed alarming dimensions and is characterised by great territorial differentiations at the national as well as regional level. This characteristic is due to structural differences in economy and regional inequalities shaped by distinct historical experiences as well as the transformation process. Regularities observed at the national level, in most cases, cannot be generalised and differ from region to region. For example, in June 2011 the highest registered unemployment rate in Poland was observed in the Warmia-Masuria Province – 19,5%, and the lowest in the Wielkopolska Province – 9,1% (province (voivodship) refers to the NUTS2 level according to Eurostat territorial division).This situation requires advanced studies reflecting regional specificities.

Data available from the Labour Force Survey enable estimation of employed and unemployed for the whole country by age, sex and place of residence: urban and rural areas. But at the regional level (NUTS2) only aggregated data can be obtained from LFS differentiated by sex and place of residence (into urban and rural areas), but not by age.

The goal is then to estimate the percentage of unemployed people in the population of 15 and older[4] at the NUTS3 level based on data from the Labour Force Survey 1st quarter of 2008. Although there had been previous attempts to estimate unemployment at the NUTS4 but they were not very satisfactory.

3. DESCRIPTION OF THE SURVEY AND EMPLOYMENT-RELATED POPULATION FLOW STUDY

Labour Force Survey

The LFS methodology is based on the definitions of the economically active population, the employed and the unemployed adopted by the Thirteenth International Conference of Labour Statisticians (October 1982) and recommended by the International Labour Office. The survey concentrates on the situation from the point of view of economic activity of population, i.e. the fact of being employed, unemployed or economically inactive in the reference week. The labour force survey is a probability sample survey.

Sampling for the LFS follows the two-stage household sampling. The primary sampling units subject to the first stage selection, are census units called census clasters

[4] It is different from the unemployment rate but such an assumption will significantly simplify the computations, especially as far as the MSE of the estimators is concerned.

– CCs in towns, while in rural areas they are enumeration districts – EDs[5]. Second stage sampling units are dwellings[6]. The primary sampling units (PSU) are sampled with the application of the so-called stratification. Strata correspond to provinces (voivodships). Strata within provinces were created depending on the size of a place; rural areas were included into the smallest ones.

The estimation process consists of defining the appropriate generalizing factors, referred to as weights. This is achieved in three steps. The first step provides primary weights, which basically are the reciprocals of selection probabilities for ultimate sampling units (i.e. dwellings), which compensates for the disproportionate construction of the sample. The secondary weights are calculated in the next step by dividing primary weights by R, where R rate depends on the category of a place of residence of a given dwelling (the rural area or one of the five town classes mentioned above). The secondary weights are also final for the results concerning households and families. Final weights for the results concerning the population are calculated in the third step. The purpose of this step is to adjust the LFS results to the current demographic estimates. It is given by calculating the so-called modifiers for each of 48 categories defined by the place-of-residence (urban/rural), sex and 12 age groups. Final weights result from multiplying secondary weights by adequate modifiers.

The variances of complex estimators obtained in LFS cannot be estimated with ordinary methods and special, approximate procedures must be employed. Since 2003 one of the most popular approximate methods has been chosen for this purpose, based on the resampling and bootstrap rule. The detailed description of the bootstrap procedure applied in the variance estimation for LFS estimates in Poland is presented in details in Bracha et al. (2003).

Tax Register – general characteristics of an employment-related population flow study in 2006

The use of administrative registers in Polish public statistics is at the initial stage. The only larger study in which they played a supporting role was the 2011 National Census, which relies on information from various sources in order to collect certain data (reducing the burden on respondents), update the sampling frame for sample surveys and to update the database of buildings and dwellings. Wider access to administrative databases provides an opportunity for Polish public statistics to develop methods of

[5] In rural areas application of smaller first stage sampling units is useful for organizational reasons, but negatively influences precision of results for these areas. In order to improve this, the principle of the so-called overrepresentation of rural areas was applied, i.e. the number of dwellings sampled from rural areas is about 10% higher than the number resulting from the so-called proportional allocation (related to the number of dwellings in the whole population).

[6] Sampling of primary sampling units and dwellings is conducted on the basis of the Domestic Territorial Division Register, including among others a list of territorial statistical units and a list of dwelling addresses within CC's and EDs.

how they can be used in statistical reporting, and consequently, to upgrade the statistical infrastructure.

A study of employment-related population flow was conducted on the basis of data from the tax system collected in the POLTAX database in the Statistical Office in Poznan in 2009. The study was intended to provide estimates about the volume and directions of commuter traffic involving people in paid employment, using data from 31 December 2006. The results and the methodological details of the study have been made partially available in the Regional Data Bank and in the book entitled "Commuting in Poland", edited by K. Kruszka, Poznan, in October 2010.

Registers, after verification and cleaning, turned out to be a good source of information about the structure of economic activity from the territorial perspective. Additionally, the set included characteristics of sex and age (variable derived from the variable "birth date"), which enabled another breakdown. One disadvantage is their incompleteness. In addition, they do not cover all the characteristics reported in other studies (such as education, class of places of residence, etc., which are included in LFS); this means that these datasets cannot fully replace the previously used measurement tools. Nevertheless, registers can play a supporting role and provide a good source for auxiliary variables in indirect estimation of labour market characteristics.

4. APPROACH TO THE PROBLEM

The problem was to estimate the percentage of unemployed people at the lower level of aggregation than presented in the CSO's publications. Having in mind that data available from the Labour Force Survey enable direct estimates of employed and unemployed for the whole country by age, sex and place of residence and for aggregated data at the regional level (NUTS2), it was decided to try to get small area estimates at the NUTS3 level.

Another problem was to evaluate the results. The applied software of course enables us to compute MSEs for the studied estimators, but there was a serious difficulty to validate the estimates against population values. It was decided that despite of small differences in the definition of LFS and registered unemployment the latter will be used as a kind of benchmark.

The natural choice was to use the EURAREA code.

The variable *status on the labour market* was recoded into binary variable getting 1 if the person was unemployed and 0 otherwise. This way target variable was created.

As potential covariates in the model we chose:

– commuting to work,
– place of residence,
– sex,
– 6 age groups.

All covariates were recoded into binary variables producing nine variables:
X1 – commuting to work (1 if a person commutes to work, 0 otherwise),

X2 – place of residence (1 if a person lives in rural area, 0 otherwise),
X3 – sex (1 if a person is a male, 0 otherwise),
X4 – group of age (1 if a person is up to 20 years of age, 0 otherwise),
X5 – group of age (1 if a person is 20-24 years of age, 0 otherwise),
X6 – group of age (1 if a person is 25-34 years of age, 0 otherwise),
X7 – group of age (1 if a person is 35-44 years of age, 0 otherwise),
X8 – group of age (1 if a person is 45-54 years of age, 0 otherwise),
X9 – group of age (1 if a person is over 55 years of age, 0 otherwise).

Then we used stepwise selection to get the following model (two variables were excluded from the model: X6 was not significant and X9 because of collinearity):

Table 1.

Model parameter estimates (domain level) after excluding X6 and X9

Variable	Parameter estimate	Standard error	t Value	Pr > |t|
Intercept	0.22376	0.18512	1.21	0.2317
X1	-0.26409	0.09227	-2.86	0.0058
X2	0.00630	0.02145	0.29	0.7700
X3	-0.50223	0.45085	-1.11	0.2699
X4	2.40624	0.63456	3.79	0.0004
X5	-1.11075	0.53435	-2.08	0.0421
X7	-1.25566	0.30082	-4.17	0.0001
X8	1.00178	0.24339	4.12	0.0001

And its goodness of fit is as follows

Table 2.

Model goodness of fit (domain level) after excluding X6 and X9

Root MSE	0.01236	R-Square	0.7191
Dependent Mean	0.05522	Adj R-Sq	0.6852
Coeff Var	22.37504		

Although another two variables (X2 and X3) are not significant we decided to include them in the model as in our opinion they are of special importance (place of residence and sex).

5. METHODS

The seven "standard" estimators were applied:
– synthetic population level estimator NSMEAN,
– direct estimator,

– GREG with a standard linear regression model,

– synthetic estimator considered under two different models:

a) a linear two-level model with individual data,

b) a linear model with area-level covariates and a pooled sample estimate of within-area variance,

– EBLUP estimator using models:

a) a linear two-level model with individual data,

b) a linear model with area-level covariates and a pooled sample estimate of within-area variance.

The eighth method was also EBLUP estimator however based on the assumption of the existence of spatial autocorrelation – EBLUPGREG SPATIAL – a linear two-level model with individual data taking into account the spatial correlation structure.

Direct estimator

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id} y_{id}, \tag{1}$$

where: $\hat{N}_d = \sum_{i \in u_d} w_{id}, \quad w_{id} = 1/\pi_{id}$.

MSE Estimator:

$$M\hat{S}E(\hat{\bar{Y}}_d^{DIRECT}) = \left(\frac{1}{\hat{N}_d}\right)^2 \sum_{i \in u_d} w_{id}(w_{id} - 1)(y_{id} - \hat{\bar{Y}}_d^{DIRECT})^2 \tag{2}$$

(assuming, that $\pi_{id,jd'} = 0$, for all $d \neq d'$ or $i \neq j$.

GREG Estimator

$$y_{id} = \mathbf{x}_{id}^T \boldsymbol{\beta} + \varepsilon_{id}, \tag{3}$$

$$\mathrm{E}(\varepsilon_{id}) = 0, \qquad \mathrm{Var}\ (\varepsilon_{id}) = \sigma_\varepsilon^2,$$

$$\hat{Y}_d^{GREG} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i} + \left(\bar{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{\mathbf{x}_i}{\pi_i}\right)^T \hat{\boldsymbol{\beta}}, \tag{4}$$

where $\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i}$ and $\hat{\boldsymbol{\beta}}$ are estimated by using LSM weighted by weights resulting from the sampling process:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in u_d} w_{id} x_{id} x_{id}^T\right)^{-1} \sum_{i \in u_d} w_{id} x_{id} y_{id}, \tag{5}$$

$$\mathrm{E}(\varepsilon_{id}) = 0, \qquad \mathrm{Var}\ (\varepsilon_{id}) = \sigma_\varepsilon^2.$$

Assuming: $r_{id} = y_{id} - \mathbf{x}_{id}^T \hat{\boldsymbol{\beta}}$ and $\hat{Y}_d^{GREG} = \sum_{i \in u_d} w_{id} g_{id} y_{id}$,

with weights g : $g_{id} = 1 + (\bar{X}_{.d} - \bar{x}_{.d})^T (\sum_{i \in u_d} w_{id} x_{id} x_{id}^T)^{-1} x_{id}$,

$$M\hat{S}E(\hat{\bar{Y}}_d^{GREG}) = \sum_{i \in u_d} \sum_{j \in u_d} \frac{\pi_{ijd} - \pi_{id}\pi_{jd}}{\pi_{ijd}\pi_{id}\pi_{jd}} g_{id} r_{id} g_{jd} r_{jd}. \tag{6}$$

SYNTHETIC ESTIMATORS

MODEL A

Standard two level linear model:

$$y_{id} = x_{id}^T \boldsymbol{\beta} + u_d + e_{id}, \tag{7}$$

$$u_d \sim iid \;\; N(0, \sigma_u^2), \qquad e_{id} \sim iid \;\; N(0, \sigma_e^2),$$

$$\hat{\bar{Y}}_d^{SYNTH} = \bar{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}, \tag{8}$$

z $\bar{X}_{.d} = (\bar{X}_{.d,1}, ..., \bar{X}_{.d,p})^T$

Estimator does not respect sampling weights

$$M\hat{S}E(\hat{\bar{Y}}_d^{SYNTH}) = \hat{\sigma}_u^2 + \bar{\mathbf{X}}_{.d} \hat{\mathbf{V}} \bar{\mathbf{X}}_{.d}^T, \tag{9}$$

where $\hat{\mathbf{V}}$ is the covariance matrix of covariates.

MODEL B

Model for domain is as follows:

$$\bar{y}_{.d} = \bar{x}_{.d}^T \boldsymbol{\beta} + \xi_d, \tag{10}$$

where $\xi_d \sim iid \;\; N(0, \sigma_u^2 \;\; + \frac{\sigma_e^2}{n_d})$ and n_d denotes sample size for area d.

Variance σ_e^2 is estimated according to the formula:

$$\hat{\sigma}_e^2 = \frac{1}{n - na} \sum_i \sum_d (y_{id} - \bar{y}_{.d})^2, \tag{11}$$

where: $n-$ sample size;

na – number of domains in the sample.

One level regression model with β i σ_u^2 estimated itreratively from:

$$\hat{\boldsymbol{\beta}} = (x^{\mathbf{T}} \mathbf{D}^{-\mathbf{1}} x)^{-\mathbf{1}} x^{\mathbf{T}} \mathbf{D}^{-\mathbf{1}} y, \tag{12}$$

where $y-$ vector of the sample elements y,

x – matrix with rows consisting of x_d^T,

D – diagonal matrix with iteratively updated values $(\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d})$ on the diagonal.

$$\hat{\bar{Y}}_d^{SYNTH} = \bar{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}, \tag{13}$$

$$\hat{MSE}(\hat{\bar{Y}}_d^{SYNTH}) = \hat{\sigma}_u^2 + \bar{\mathbf{X}}_{.d}^T \hat{\mathbf{V}} \bar{\mathbf{X}}_{.d}, \tag{14}$$

where $\hat{\mathbf{V}}$ is the estimate of the covariance matrix $(x^{\mathbf{T}}\mathbf{D}^{-1}x)^{-1}$.

EBLUP ESTIMATORS

MODEL A

$$\hat{\bar{Y}}_d^{EBLUP} = w_d^{EBLUP} \hat{\bar{Y}}_d^{GREG} + (1 - w_d^{EBLUP}) \hat{\bar{Y}}_d^{SYNTH}, \tag{15}$$

$$w_d^{EBLUP} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d}. \tag{16}$$

In more details the models may be presented as follows:

EBLUP using MODEL A

$$\hat{\bar{\bar{Y}}}_d = \gamma_d (\bar{y}_{.d} - \bar{\mathbf{x}}_{\mathbf{.d}}^{\mathbf{T}} \hat{\beta}) + \bar{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}}, \tag{17}$$

where:

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d}, \tag{18}$$

$\bar{y}_{.d}$, $\bar{\mathbf{x}}_{\mathbf{.d}}^{\mathbf{T}}$, are corresponding sample means of y and the covariates for domain d. $\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2, \hat{\sigma}_u^2$are parameters estimated using standard two-level linear model.

MSE Estimators.

ESTIMATOR 1

$$\hat{MSE}(\hat{\bar{\bar{Y}}}_d) = \frac{\gamma_d \hat{\sigma}_e^2}{n_d} + (1 - \gamma_d)^2 \bar{\mathbf{X}}_{.d}^T \hat{\mathbf{V}} \bar{\mathbf{X}}_{.d} \tag{19}$$

ESTIMATOR 2

$$\begin{aligned} \hat{MSE}(\hat{\bar{\bar{Y}}}_d) = {} & \frac{\gamma_d \hat{\sigma}_e^2}{n_d} + (1 - \gamma_d)^2 \left(\bar{\mathbf{X}}_{.d}^T \hat{\mathbf{V}} \bar{\mathbf{X}}_{.d} \right) + \\ & 2 \times \left(\frac{\hat{\sigma}_e^2}{n_d} \right)^2 \left(\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_d \right)^{-3} \left(\hat{Var}(\hat{\sigma}_u^2) + \frac{\hat{\sigma}_u^4}{\hat{\sigma}_e^4} \hat{Var}(\hat{\sigma}_e^2) - 2 \frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2} \hat{Cov}(\hat{\sigma}_u^2, \sigma_e^2) \right), \end{aligned} \tag{20}$$

where:

$$\begin{aligned} \hat{Var}(\hat{\sigma}_e^2) &= \frac{2\hat{\sigma}_e^4}{m_1 - m_2 - p}, \\ \hat{Cov}(\hat{\sigma}_e^2, \hat{\sigma}_u^2) &= \frac{2\hat{\sigma}_e^2 \hat{\sigma}_u^2}{m_1 - m_2 - p}, \\ \hat{Var}(\hat{\sigma}_u^2) &= \hat{Var}(\hat{\rho}) \hat{Var}(\hat{\sigma}_e^2) + \hat{\sigma}_e^4 \hat{Var}(\hat{\rho}) + \hat{\rho}^2 \hat{Var}(\hat{\sigma}_e^2), \end{aligned}$$

m_1- number of selected units,
m_2- number of domains,
$\rho = \sigma_u^2/\sigma_e^2$ is the variance ratio,

$$\hat{Var}(\hat{\rho}) = \frac{2}{\sum\limits_d n_d^2/(1 + n_d\hat{\rho})^2},$$

where $\hat{\mathbf{V}}$ is the covariance matrix of covariates.

$\hat{Var}(\hat{\sigma}_e^2)$ is estimated variance $\hat{\sigma}_e^2$ and $\hat{Var}(\hat{\sigma}_u^2)$ is estimated variance $\hat{\sigma}_u^2$. Estimator 1 is a corresponding approximation if the number of domains is large. Estimator 2 may be applied in any case.

MODEL B

$$\hat{\bar{Y}}_d^{EBLUP} = w_d^{EBLUP}\hat{\bar{Y}}_d^{DIRECT} + (1 - w_d^{EBLUP})\hat{\bar{Y}}_d^{SYNTH}, \tag{21}$$

$$w_d^{EBLUP} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\psi}_d}. \tag{22}$$

EBLUP using MODEL B.

$$\hat{\bar{Y}}_d^{EBLUP} = \gamma_d\hat{\bar{Y}}_d^{direct} + (1 - \gamma_d)\bar{\mathbf{X}}_{.d}^T\hat{\boldsymbol{\beta}}, \tag{23}$$

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}, \tag{24}$$

where

$$\hat{\beta} = (x^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}x)^{-\mathbf{1}}x^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}y, \tag{25}$$

where:

$y-$ vector of the sample elements y,
x – matrix with rows consisting of $\bar{x}_{.d}^T$,
D – diagonal matrix with iterative updated values $(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$ on the diagonal.

MSE Estimators

ESTIMATOR 1

$$\hat{MSE}(\hat{\bar{Y}}_d) = \gamma_d\hat{\psi}_d + (1 - \gamma_d)^2\left(\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right), \tag{26}$$

$\hat{\psi}_d$ is an estimator of residual variance inside domains $\hat{\psi}_d = \dfrac{\sigma_e^2}{n_d}$.

ESTIMATOR 2

$$\hat{MSE}(\hat{\bar{Y}}_d) = \gamma_d\hat{\psi}_d + (1 - \gamma_d)^2\left(\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right) + \hat{\psi}_d^2\left(\hat{\sigma}_u^2 + \hat{\psi}_d\right)^{-3}\hat{Var}(\hat{\sigma}_u^2) \tag{27}$$

ESTIMATOR 3

$$M\hat{S}E(\hat{\bar{Y}}_d) = \gamma_d\hat{\psi}_d + (1-\gamma_d)^2\left(\bar{\mathbf{X}}_{.d}^T\hat{\mathbf{V}}\bar{\mathbf{X}}_{.d}\right) + 2\times\hat{\psi}_d^2\left(\hat{\sigma}_u^2+\hat{\psi}_d\right)^{-3}V\hat{a}r(\hat{\sigma}_u^2), \tag{28}$$

where $\hat{\mathbf{V}}$ is the covariance matrix of covariates and $V\hat{a}r(\hat{\sigma}_u^2)$ is an estimate of $\hat{\sigma}_u^2$.

EBLUP using spatial correlation (EBLUPGREG_SPATIAL)

Software prepared by ISTAT was based on re-formulation of the expressions contained in Saei and Chambers (2003) in order to obtain a more efficient SAS code.

The considered model is the general linear mixed model

$$y = X\boldsymbol{\beta} + Zu + e, \tag{29}$$

where:

X and **Z** are known matrices of order N×P and N×DOM respectively;

X is the matrix of the population values of the covariates and **Z** is the incidence matrix for the spatial random area effect;

e and **u** are vectors of random variables with mean and variance and covariance matrices expressed respectively by the couples:

$$N \sim [0,\ \sigma^2 I_N],$$

$$N \sim [0,\ \sigma_U^2 A],$$

I_N being the identity matrix of order N and A square matrix of order DOM allowing a spatial correlation structure to be included in the model. The generic element of A $a_{dd'}$ is given by

$$\mathbf{a}_{(dd')} = \left[1+\delta_{(dd')}\exp\left(\frac{\boldsymbol{dist(d,d')}}{\alpha}\right)\right]^{-1}, \tag{30}$$

$$\delta_{(dd')} = \begin{cases} 0\ \boldsymbol{for} & \boldsymbol{d=d'}, \\ 1\ \boldsymbol{for} & \boldsymbol{d\neq d'}, \end{cases}$$

$\boldsymbol{dist(d,d')}$ is the Euclidean distance between centroids of area d and d'.

6. SOFTWARE

The software used in the study was SAS. We started from using PROC SURVEYMEANS to compute the *direct* estimator. Then the EURAREA code was implemented. Not only were Seven Standard Estimators computed but the EBLUPGREG software with its options was tested to find out whether to allow for spatial autocorrelation or not. The SAS software was also used for computing coordinates of the centroids which are of special importance in estimation conducted via EBLUPGREG taking into account the spatial dependence.

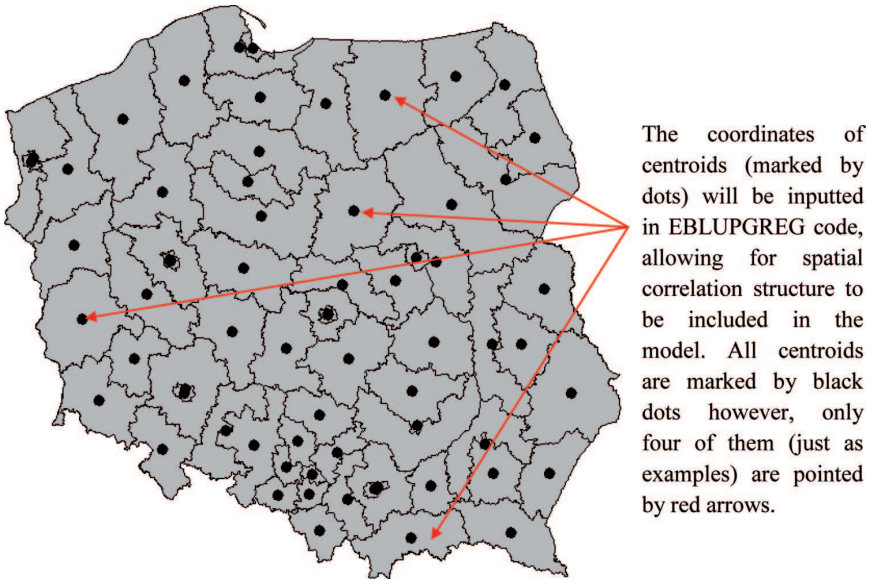


Figure 1. Centroids of small domains – NUTS4

7. RESULTS

Figures 2 and 3 show that the spatial pattern of the direct estimates for NUTS3 level fits quite well to the pattern of the values coming from the administrative registers.

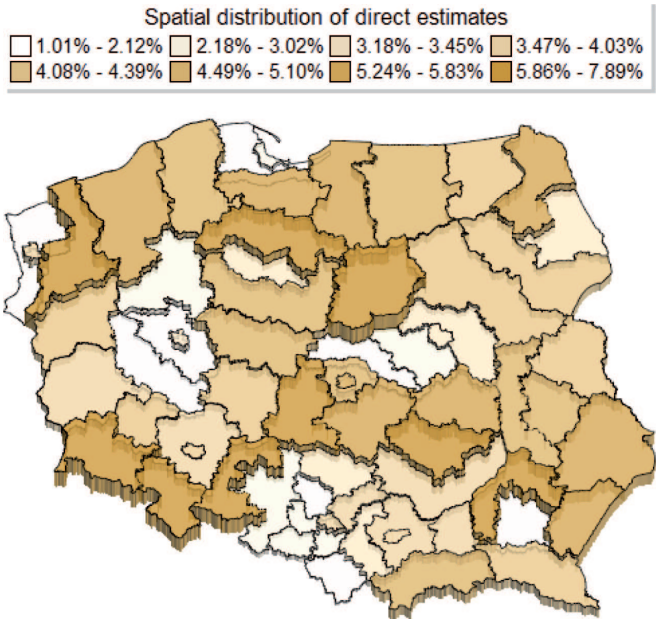


Figure 2. Direct estimates

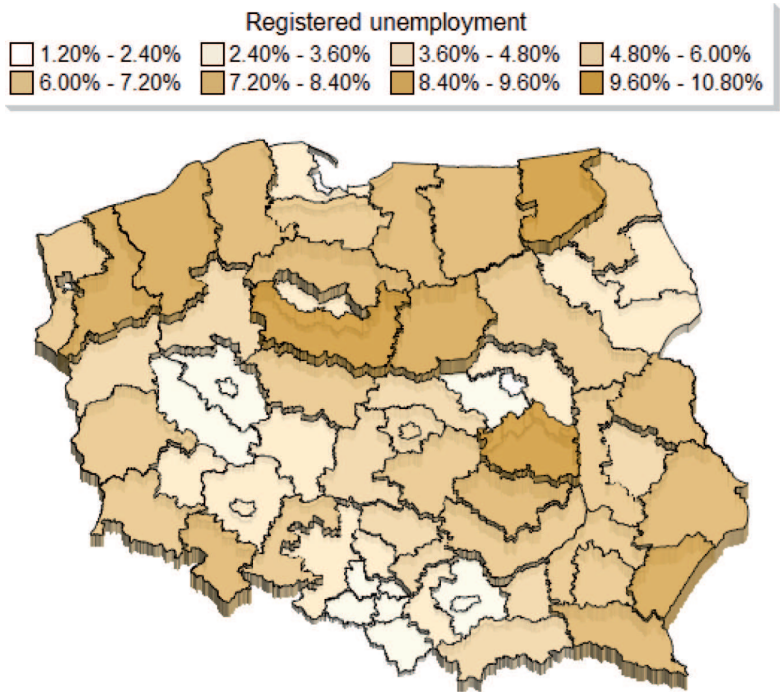


Figure 3. Registered unemployment

The following figures (Fig. 4 – Fig.10) give some overview on how the applied estimators reproduce the registered unemployment. The Figure 4 showed that direct estimates of ILO (International Labor Bureau) unemployment are in most of the cases lower for the NUTS3 areas than registered unemployment. Quite different situation is presented on the Figure 5. In case of model assisted estimator (GREG) the estimated unemployment from LFS is higher than registered. However if the ranges of unemployment are compared one could see that direct estimates have slightly more narrow range when compared with GREG. The registered unemployment ranges from 1.41% to 10.77%, direct estimates range from 1.01% to 7.89% and finally GREG from 2.92% to 10.89%.

Synth_A estimator has the smallest variance but the bias in this case is unacceptable. Synth_B estimates are quite similar to direct ones. However the range for the estimates obtained while applying Synth_B is very short – from 2.35% to 5.94%. So one could conclude that smoothing of the estimates is too strong.

The comparison of the EBLUP estimators based on different assumptions revealed that they should be of a special interest in applications of small area methodology in Labor Force Survey. The results obtained suggest their bias is relatively small with relatively small variation. For instance the range of estimates produced by EBLUP-GREG_SPATIAL (which takes into consideration the spatial structure of data) is from 4.50% to 10.34%. One should notice that even some basic assumptions relating to the

Figure 4. Comparison of registered unemployment rate and direct estimates

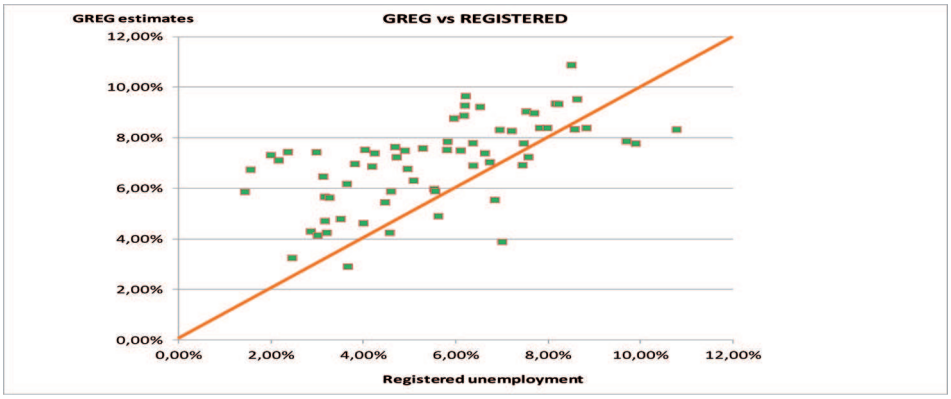


Figure 5. Comparison of registered unemployment rate and greg estimates

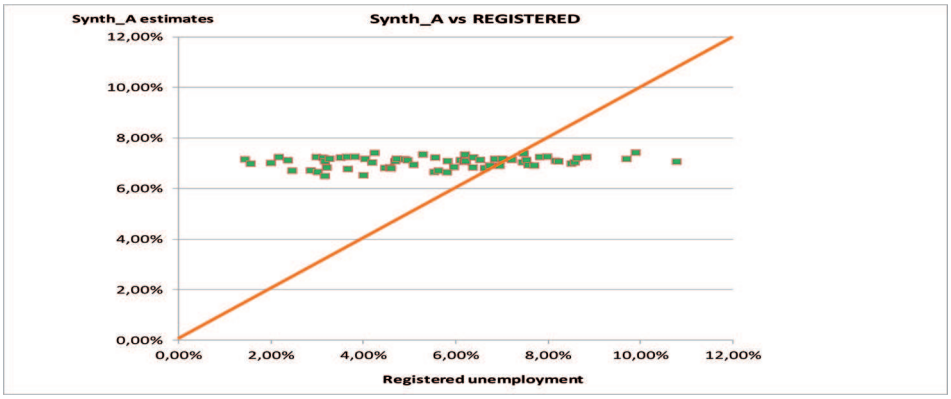


Figure 6. Comparison of registered unemployment rate and Synth_A estimates

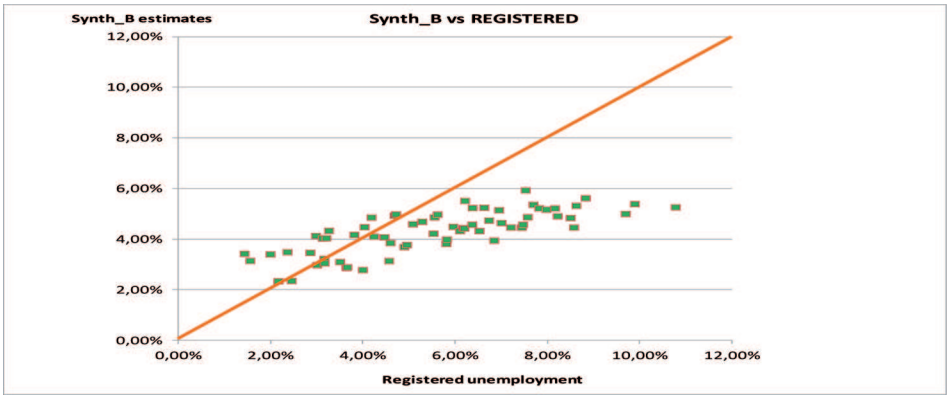


Figure 7. Comparison of registered unemployment rate and Synth_B estimates

nature of variable are violated (number of unemployed people aging 15 and more is not a continuous variable and the distribution could not be normal), the behavior of the EBLUPs is quite well.

The estimates obtained by application of EBLUP_A and EBLUP_B models show the same patters as their direct components. EBLUP_A is a linear combination of GREG and SYNTH_A while EBLUP_B is a linear combination of DIRECT and SYNTH_B.

We also compared the MSE of the studied estimators. The software produced in EURAREA project computes the MSE of seven standard estimators and also for spatial version of EBLUPGREG estimator. However the approach presented by us has two simplifications. In fact for DIRECT and GREG estimators we have mainly variance as these two estimators are design unbiased. The second problem while applying EURAREA code the reader should realize is the fact, that the quality assessment measures are computed assuming simple random sampling. In fact the sampling design applied in LFS surveys is rather stratified and two-stage in most cases.

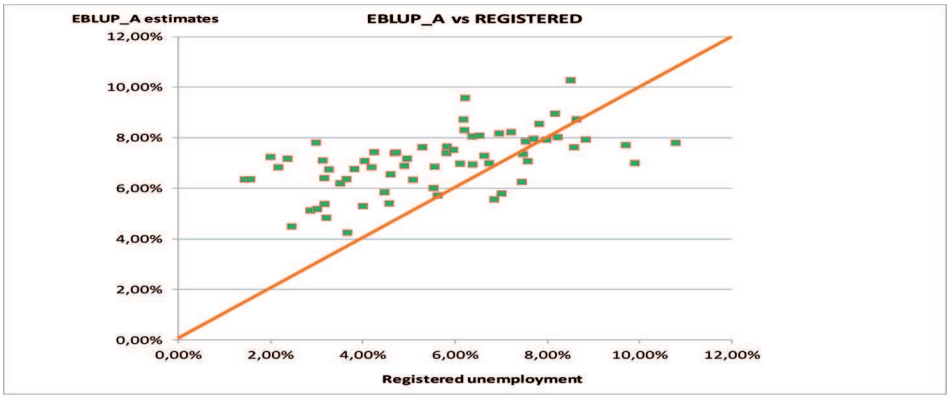


Figure 8. Comparison of registered unemployment rate and EBLUP_A estimates

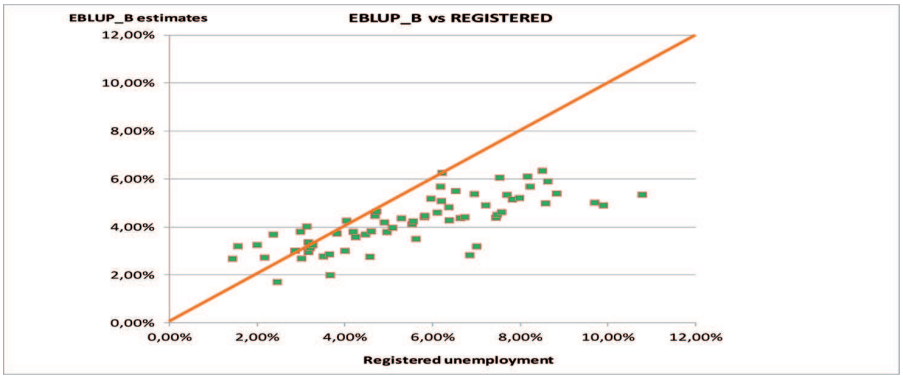


Figure 9. Comparison of registered unemployment rate and EBLUP_B estimates

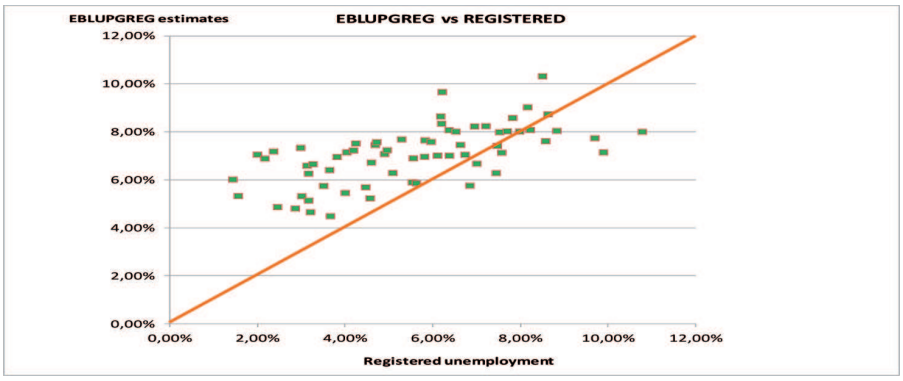


Figure 10. Comparison of registered unemployment rate and EBLUPGREG_SPATIAL estimates

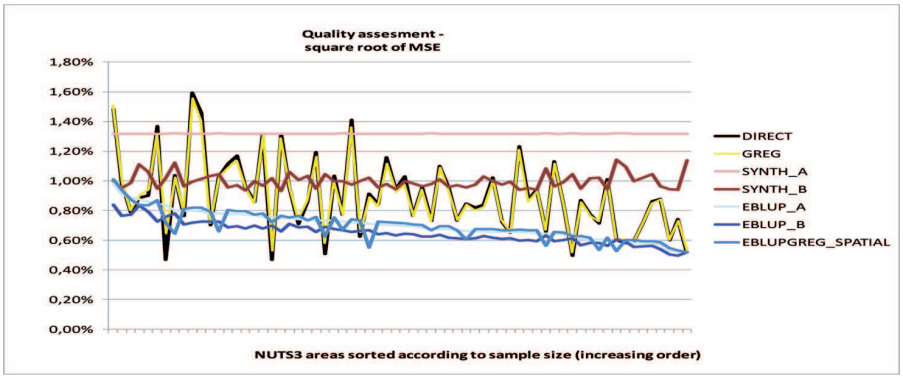


Figure 11. Distribution of MSEs (NUTS3 ordered according to increasing sample size)

The last Figure (11) presents the distribution of MSEs where the NUTS3 were ordered according to the increasing sample size. The highest values of MSEs are connected with SYTH_A estimator. In this case the important input in its value is the input of bias. The design unbiased estimators – DIRECT and GREG show a quite high amount of variance which is slightly smaller in the case of GREG. However when sample size increases the variance of the DIRECT estimator is decreasing. In the case of SYNTH the variance is rather constant. The best performance as far as the behavior of the estimation is concerned is connected with EBLUPs. They are quite similar and have the smallest MSE.

Poznan University of Economy

LITERATURA

[1] Bracha, C., Lednicki, B. and Wieczorkowski, R., (2003), *Estimation of Data from the Polish Labour Force Surveys by poviats (counties) in 1995—2002* (in Polish), Central Statistical Office of Poland, Warsaw.
http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_estymacja_danych_z_bad_na_poziomie_pow_dla_lat_1995_2002.pdf

[2] Chandra H., Salvati N., Chambers R., (2009), *Small Area estimation for Spatially Correlated Populations – A Comparison of Direct and Indirect Model-Based Methods,* Southampton Statistical Sciences Research Institute, Methodology Working Paper M07/09, University of Southampton.

[3] D'Alò M., Falorsi S., Solari F., (2004), *EURAREA Documentation on SAS/IML program on Linear Mixed Model with Spatial Correlated Area Effects in Small Area Estimation*, EURAREA Deliverable 3.3.2.

[4] *EURAREA_Project_Reference_Volume,* http://www.statistics.gov.uk/eurarea.

[5] *EURAREA EBLUPGREG Software Documentation*, Statistics Finland EURAREA Consortium, Deliverables D2.3.2, D3.3.2, 2004.

[6] Gołata E., (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo AE w Poznaniu, Poznań.

[7] Kruszka K., (ed.), (2010), *Commuting in Poland*, Statistical Office in Poznań, Poznań.

[8] Kubacki J., (2004), *Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey*, Statistics in Transition, 6 (5), 785-796, Warsaw.

[9] Saei A., Chambers R., (2003), *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*, University of Southampton.

[10] Saei A., Chambers R., (2004), *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, University of Southampton.

WYKORZYSTANIE ESTYMACJI POŚREDNIEJ UWZGLĘDNIAJĄCEJ KORELACJĘ PRZESTRZENNĄ W BADANIACH SPOŁECZNYCH W POLSCE

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

Artykuł przedstawia propozycję wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania pewnych charakterystyk rynku pracy w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w 2008 roku. Jest to bardziej szczegółowy poziom agregacji przestrzennej niż ten prezentowany w publikacjach Głównego Urzędu Statystycznego

opartych na wynikach Badania Aktywności Ekonomicznej Ludności. Drugim celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (empirical best linear unbiased predictor) oraz estymatora typu EBLUPGREG_SPATIAL (uwzględniającego korelację przestrzenną).

Słowa kluczowe: statystyka małych obszarów, autokorelacja przestrzenna, bezrobocie, Badanie Aktywności Ekonomicznej Ludności (BAEL)

USING INDIRECT ESTIMATION WITH SPATIAL AUTOCORRELATION IN SOCIAL SURVEYS IN POLAND

Abstract

The article presents possible application of indirect estimation methods (including the method accounting for spatial correlation) to estimate some characteristics of labor market in the population of people aged 15 and over at the level of NUTS3 in Poland in 2008. This is a more detailed spatial aggregation of data compared with that found in publications of the Central Statistical Office based on Labour Force Survey results. The second aim of the article is to compare the precision measures of the direct estimator with those of the EBLUP estimator (*empirical best linear unbiased predictor*) and the EBLUPGREG_SPATIAL estimator (which takes into account spatial correlation).

Key words: small area statistics, spatial autocorrelation, unemployment, Labour Force Survey (LFS)

PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

ANDRZEJ CZYŻEWSKI, ALEKSANDER GRZELAK

MOŻLIWOŚCI WYKORZYSTANIA STATYSTYKI BILANSÓW PRZEPŁYWÓW MIĘDZYGAŁĘZIOWYCH DLA MAKROEKONOMICZNYCH OCEN W GOSPODARCE

1. WSTĘP

Ocena procesów makroekonomicznych stwarza zapotrzebowanie na wykorzystanie różnych teorii i modeli. Jedną z takich możliwości jest model przepływów międzygałęziowych. Poprzez analizę typu dostawca – odbiorca (input – output) konkretyzuje on idee funkcjonowania mechanizmu gospodarczego (rynkowego i budżetowego), jego wewnętrzne powiązania, zależności. Jest jednocześnie użytecznym instrumentem oceny funkcjonowania gospodarki (Tomaszewicz, 1994). Opierając się na założeniach teorii równowagi ogólnej pozwala na analizę wytworzonych efektów makroekonomicznych, procesów redystrybucji budżetowej, związków danych sektorów z otoczeniem, oddziaływania procesów globalnych na gospodarkę poprzez eksport i import. Zaletą tego modelu jest także możliwość wnioskowania dedukcyjnego, co umożliwia jego szerokie wykorzystanie, także w procesie kształcenia akademickiego. Głównym celem artykułu jest ocena możliwości wykorzystania modelu przepływów międzygałęziowych do analizy procesów makroekonomicznych w gospodarce. Przedstawione rozważania mają charakter teoretyczny. Zawierają także określone propozycje modyfikacji modelu input-output wynikające z dotychczasowych doświadczeń autorów w zakresie praktycznego jego zastosowania jak i doświadczeń badawczych.

2. PROPEDEUTYKA MODELU INPUT-OUTPUT

Procesy gospodarcze charakteryzują się złożonością. Związane jest to ze wzajemnymi powiązaniami między krajami, sektorami, podmiotami. Gospodarka krajowa, czy globalna jest systemem naczyń połączonych. Stąd zmiany sytuacji w jednych obszarach gospodarki oddziałują na pozostałe. Istnienie przepływów produktów między sektorami tworzy zapotrzebowanie na analizę nakładów i wyników w skali poszczególnych grup przedsiębiorstw oraz całej gospodarki. Zazwyczaj traktuje się ją jako model ustalania ilościowych związków między różnymi sektorami, prowadzących do ogólnej równowagi gospodarczej. Pierwszym, który zauważył i wykorzystał ten sens analizy przepływów, był F. Quesnay 1758 (Quesnay, 1928), który w formie tablicy ekonomicznej przedstawił przepływy dóbr między trzema działami gospodarki: rolnictwem (klasa produkcyjna), sferą pozarolniczą (klasa jałowa) i właścicielami (władza świecka i duchowna). Na tej

podstawie przedstawił współzależności sfery wytwórczej gospodarki, podziału wytworzonego produktu społecznego oraz sfery dochodowej (Czyżewski, 2011).

Współczesne analizy opierające się na bilansie przepływów międzygałęziowych nawiązują co do idei, do modelu przepływów input-output W. Leontiefa (Leontief, 1936). Jego istota sprowadza się do założenia, iż gospodarka narodowa stanowi agregat zasobów i strumieni składających się z kilku sprzężonych ze sobą układów: produkcyjnego i usług oraz zagranicy, gospodarstw domowych, budżetu i banków, które opisano metodą nakładów i wyników (input-output) w formie tabelarycznej (szachownicowej). Model przepływów międzygałęziowych składa się z czterech części. W I części zawarto poszczególne fazy produkcji, określające zaspokojenie popytu pośredniego sektorów. W II części odzwierciedlono popyt finalny. Odbiorcami jego są: konsument indywidualny, budżet, a także sfera inwestycyjna zaopatrująca się w środki trwałe i obrotowe. Część III przedstawia makroekonomiczne efekty tworzone w poszczególnych sektorach, z perspektywy dochodów (składników wartości dodanej). Część IV tablicy z kolei dotyczy podziału wytworzonych dochodów. W praktyce gospodarki centralnie sterowanej, jak i rynkowej zazwyczaj nie jest wypełniana przez GUS[1] (Czyżewski, 2011). Konstrukcja taka (cztero-częściowa) zapewnia z jednej strony kompleksowość analiz, z drugiej szczegółowość ocen w zakresie wzajemnych powiązań sektorów w gospodarce, co jest niemożliwe do realizacji przy standardowych zagregowanych danych GUS.

3. OCZEKIWANIA EKONOMISTÓW A PRAKTYKA GUS

Idea przepływów międzygałęziowych ma swój wymiar zarówno teoretyczny, głęboko osadzony w historii myśli ekonomicznej, jak i aplikacyjny związany z publikowaniem przez GUS stosownych bilansów. Pierwszy z wymienionych wymiarów doczekał się wielu interpretacji i modyfikacji. Pozwolimy sobie w tej części tego opracowania przedstawić przykładową interpretację teoretyczną (tab. 1), wykorzystywaną do objaśniania współzależności makroekonomicznych.

Oznaczenia schematu (tab.1) (Czyżewski, 2011):

n – liczba sektorów na które zdezagregowano gospodarkę,

w_{ij} – przepływ dóbr i usług „i" wytworzonych w danej gospodarce bądź pochodzących z importu uzupełniającego produkcję tej gospodarki (wraz z marżami handlowymi), a zużytych przez sektor „j";

$w_{ij} = x_{ij}+y_{ij}+m_{ij;}$

x_{ij} – wartość przepływu dóbr i usług „i" wytworzonych w danej gospodarce, a zużytych przez sektor „j";

y_{ij} – wartość przepływu dóbr i usług pochodzących z importu uzupełniającego produkcję „i" danych sektorów, a zużytych przez sektor "j";

[1] Wykorzystywano ją jedynie sporadycznie dla koordynacji strony finansowej planów gospodarczych. Ostatnie publikowane dane z tego zakresu dotyczą bilansu gospodarki polskiej za 1957 r. (Opracowanie..., 1958).

Tabela 1

Bilans pieniężnych przepływów międzygałęziowych (międzysektorowych)

i \ j		Część I Popyt pośredni				Część II Popyt finalny podmiotów gospodarczych (wydatki)					Razem popyt końcowy	Razem produkcja globalna i eksport
		Sektory			Razem popyt pośredni	Konsumenci indywidualni	Budżet	Banki	Inwestorzy	Popyt restytucyjny		
		1	.....	n	Σ	I	II	III	IV	V	Σ	**ΣΣ**
Koszty materiałowe i usługi	1	w_{11}	.....	w_{1n}	*Σ*	C_1'	C_1''	k_1	I_1	p_1	Σ	**Σ**
	.	.	.....	.	.	.	.	.	.	.	.	.
	.	.	.....	.	.	.	.	.	.	.	.	.
	.	.	.....	.	.	.	.	.	.	.	.	.
	n	w_{n1}	....	w_{nn}	*Σ*	C_n'	C_n''	k_n	I_n	p_n	Σ	**Σ**
Razem koszty materiałowe i usługi		Σ	.....	Σ	ΣΣ	Σ	Σ	Σ	Σ	Σ	ΣΣ	**ΣΣΣ**
Część III Dochody	1	V_1	.....	V_n	Σ	V_c'	V_c''	V_k	V_I	-	Σ	
	2	B_1	.....	B_n	.	B_c'	B_c''	B_k	B_I	-	.	
	3	A_1	.....	A_n	.	-	-	-	-	A_p	.	
	4	Z_1	.....	Z_n	.	Z_c'	Z_c''	Z_k	Z_I	-	.	
	5	m_1	.....	m_n	.	m_c'	m_c''	m_k	m_I	-	.	
	6	TA	.....	TA	Σ	TA_c	TA_c	TA_k	TA_i	-	Σ	
Razem dochody brutto		Σ	.....	Σ	ΣΣ	Σ	Σ	Σ	Σ	Σ	ΣΣ	
Razem produkcja globalna i import (koszty i dochody brutto		**Σ**	**.....**	**Σ**	**ΣΣΣ**	Część IV						

Źródło: (Czyżewski, 2011)

m_{ij} – wartość marż handlowych zrealizowanych na dostawach dóbr i usług „i" wyprodukowanych w danej gospodarce bądź pochodzących z importu „uzupełniającego" produkcję tej gospodarki, a zużytych przez sektor „j";

V_i (1) – płace pracowników (konsumentów);

B_i (2) – dochody poszczególnych sektorów z tytułu dotacji budżetowych;

A_i (3) – przychody poszczególnych sektorów z tytułu amortyzacji (w tym nadwyżka amortyzacyjna);

Z_i(4) – dochody poszczególnych sektorów z tytułu nadwyżki ekonomicznej;

m_i(5) – dochody poszczególnych sektorów z tytułu zmiany ilości pieniądza pozostającego do ich dyspozycji (głównie z tytułu kredytów, a także emisji);

TA (6) – podatki bezpośrednie i pośrednie

C_i' (I) – struktura popytu finalnego (wydatków) konsumentów indywidualnych na dobra „i";

C_i" (II) – struktura popytu finalnego budżetu (wydatków) na dobra „i" (z wyłączeniem funduszów celowych);

F_i', F_i" (IIa i Iib) – struktura popytu finalnego (wydatków) na dobra „i" w ramach przykładowych funduszów celowych (IIa i IIb);

k_i(III) – zakupy bankowe; struktura popytu finalnego sektora bankowego na dobra „i";

I_i (IV) – struktura popytu finalnego sfery inwestycji na dobra „i" (zakupy inwestycyjne);

p_i(V) – struktura popytu restytucyjnego podmiotów gospodarczych (wydatki na rzecz amortyzacji)[2].

Zaproponowany model pieniężnych przepływów międzygałęziowych (tab.1) składa się z 4 części[3]. Część pierwsza określa popyt pośredni zgłaszany przez sektory gospodarki. Zostały tu zaprezentowane wzajemne transakcje między sektorami. W wierszach przedstawiony został strumień popytu pośredniego. W tym przypadku wielkości odnoszące się do danego sektora oznaczają wartość strumienia jego produktów (usług) na które został zrealizowany popyt przez inne sektory w celu dalszego ich przetworzenia. W kolumnach natomiast pokazano strukturę kosztów (zakupów) produktów i usług poszczególnych sektorów. Stąd ćwiartka ta służy głównie do oceny wzajemnych zależności gospodarczych pomiędzy sektorami.

W II części przedstawiono popyt finalny: konsumentów indywidualnych, budżetu, sektora bankowego, a także sfery inwestycji (akumulacja). Na podstawie zawartych

[2] W oznaczeniach poszczególnych składników drugiej, trzeciej oraz czwartej ćwiartki tabeli przepływów użyto tylko jednego indeksu – „i". Drugi indeks – „j" – zastąpiony został odpowiednim zróżnicowaniem liter (np. C_i', C_i", F_i' itd.).

[3] W materiałach GUS w odniesieniu do przepływów publikowane są trzy części. Wynika to stąd, że na poziomie zagregowanym nie przedstawia się części czwartej, odnoszącą się do podziału dochodów. Niemniej jednak procesy, które związane są z tymi zjawiskami są istotne dla funkcjonowania gospodarki i dlatego w rozważaniach w niniejszym artykule zostały uwzględnione na poziomie teoretycznym (Czyżewski, 2011).

tam danych można dokonać oceny strumieni odnoszących się do rozdysponowania produktów danego sektora służących zaspokojeniu finalnego popytu przez podmioty gospodarcze. Jednocześnie ćwiartka ta z punktu widzenia teoretycznego może być rozpatrywana przez pryzmat popytu potencjalnego (aspiracje). Związany jest on z oczekiwaniami podmiotów występujących na rynku. Popyt ten jest wyższy od efektywnego (zrealizowanego). Różnica między tymi kategoriami na gruncie ekonomii keynesowskiej sprowadza się do istnienia luki popytowej (równowaga w warunkach niepełnego wykorzystania czynników wytwórczych). W gospodarce rynkowej uwidocznione to jest w nierównowadze podażowej. Po prostu produkcja, której wartość w uproszczeniu (ze względu na brak uwzględnienia kosztów niepełnego wykorzystania zasobów produkcyjnych) może być odzwierciedleniem łącznego popytu aspiracyjnego, nie znajduje popytu. Rośnie presja na konsumenta poprzez działania promocyjne, a z drugiej strony zwiększają się względne oszczędności (Czyżewski, Grzelak, 2009).

Część III tablicy ilustruje z kolei proces tworzenia dochodów brutto. W wierszach wyróżnione są poszczególne elementy wartości dodanej, m.in. płace, nadwyżka operacyjna. W oddzielnych wierszach zapisano wpływ budżetu państwa (poprzez m.in. dotacje), podatków i banków na dochody przedsiębiorstw. Z informacji zawartych w tej części możemy oceniać makroekonomiczne efekty działalności poszczególnych sektorów, w tym zwłaszcza wielkość nadwyżki operacyjnej, jak również wartości dodanej (Czyżewski, 2011), co pozwala to na identyfikację źródeł wzrostu gospodarczego.

Z kolei część IV poświęcona jest głównie podziałowi dochodu narodowego, (brutto) który z kolei służy realizacji popytu finalnego. W ćwiartce tej dochód zostaje rozdzielony pomiędzy podmioty występujące w gospodarce. Uzyskują one dochody z różnych tytułów na pokrycie wydatków składających się na ich popyt finalny, zaprezentowany w II części tablicy przepływów. W ćwiartce tej następuje dalsza modyfikacja pierwotnego podziału wytworzonych dochodów. Można sądzić, że procesy globalizacji oddziałują na zmniejszenie się roli tej części bilansu przepływów międzygałęziowych ze względu na presję wywieraną na zmniejszenie roli państwa w procesach gospodarczych (Czyżewski, Grzelak 2009). Z drugiej strony ostatnie doświadczenia związane z globalnym kryzysem, a zwłaszcza z instrumentami jego przeciwdziałania, wskazują na istotność ocen procesów gospodarczych dotyczących tej części modelu.

Z kolei model przepływów międzygałęziowych wykorzystywany przez GUS nie wykorzystuje IV ćwiartki (tab. 2). W okresie gospodarki centralnie planowanej wynikało to z istnienia nierównowagi popytowej, która odzwierciedlała się w tym, że dochody podmiotów gospodarczych nie były odzwierciedlone w produkcji i usługach (realnie ćwiartka III>ćwiartki IV i II). Natomiast w gospodarce rynkowej sytuacja jest odwrotna. Jednocześnie ćwiartki II i III w opracowaniach GUS służą bardziej do interpretacji procesów równowagi popytowo-podażowej w procesach gospodarczych niż oceny poziomu nierównowagi. Brakuje także sfery polityki pieniężnej i związanej z tym kreacji pieniądza w gospodarce poprzez kredyt. Jednocześnie wyeksponowane jest znaczenie rezydentów i nierezydentów w zakresie kreowania popytu.

Tabela 2

Bilans pieniężnych przepływów międzygałęziowych (międzysektorowych) według GUS

j / i		Część I Zużycie pośrednie (PKD)			Część II Popyt końcowy							Ogółem zużycie pośrednie i popyt końcowy
					spożycie			akumulacja brutto		export	razem	
		Sektory		Razem popyt pośredni	przez gosp. dom.	przez inst. niekomercyjne	przez instytucje rządowe i samorządowe	nakłady brutto na środki trw.	przyrost rzecz. środ. obrot.			
Produkty (PKWiU)	1	w_{11} ..... w_{1n}		Σ	C_1	C_1'	C_1''	N_1	O_1	Ex_1	Σ	$\Sigma\Sigma$
	.	.....		.	.	.	.	.	.	.	.	.
	.	.....		.	.	.	.	.	.	.	.	.
	.	.....		.	.	.	.	.	.	.	.	.
	n	w_{n1} ..... w_{nn}		Σ	C_n	C_n'	C_n''	N_n	O_n	Ex_n	Σ	$\Sigma\Sigma$
Razem zużycie pośrednie*		Σ	....	$\Sigma\Sigma$	Σ	Σ	Σ	Σ	Σ	Σ	$\Sigma\Sigma$	$\Sigma\Sigma\Sigma$
Część III	n+1	.....		Σ								
	n+2	.....		.								
	n+3	.....		.								
	n+4	.....		.								
	n+5	.....		Σ								
Wartość dodana brutto		Σ	.....	Σ	$\Sigma\Sigma$							
Produkcja globalna		Σ	.....	Σ	$\Sigma\Sigma\Sigma\Sigma$							

* razem produkty oraz wykorzystanie produktów importowanych (cif), jak również podatki od prod. pomniejszone o dot. do prod.

Źródło: Rachunek podaży i wykorzystania wyrobów i usług w 2005 roku, Warszawa 2009

Objaśnienia do tab.2:
(n+1) – koszty związane z zatrudnieniem;
(n+2) – podatki od producentów – dotacje dla producentów;
(n+3) – amortyzacja środków trwałych;
(n+4) – nadwyżka operacyjna netto;
(n+5) – nadwyżka operacyjna brutto;

Jakiego rodzaju analiz, wobec powyższych informacji, można dokonać opierając się na materiałach GUS dotyczących bilansów przepływów międzygałęziowych? Mogą służyć one do oceny struktury strumieni produktów „zasilających" dany sektor (odpowiednie kolumny, np. kolumna w I cz. tablicy przepływów międzygałęziowych). Pozwala to na określenie zakresu samozaopatrzenia, czy wzajemnych powiązań pomiędzy sektorami w układzie przedmiotowym i dynamicznym (dla ujęć wieloletnich). Oceniając z kolei struktury rozdysponowania produktów danych sektorów, a w szczególności elementy popytu końcowego (spożycie, akumulację) można dokonać analizy ich pozycji w gospodarce w kontekście surowcowego, bądź finalnego charakteru. Z perspektywy okresu transformacji gospodarki w Polsce można odnieść wrażenie, że niejednokrotnie decydenci gospodarczy przy tworzeniu programów restrukturyzacyjnych (prywatyzacyjnych) poszczególnych sektorów nie uwzględniali właśnie tych strumieni określających wzajemne powiązania w gospodarce (Czyżewski, Grzelak, 2007).

Na podstawie tablicy przepływów międzygałęziowych można zbadać także strukturę bezpośrednich i pośrednich nakładów bieżących oraz nakładów majątkowych, a przez odwrócenie tzw. współczynników „chłonności" określić efektywność poszczególnych rodzajów nakładów. Celowi temu służą m.in. współczynniki produktochłonności (materiałochłonności). Najczęściej stosowany jest współczynnik bezpośredniej materiałochłonności, zwany technicznym współczynnikiem produkcji. Określa on relację wartości dóbr zużytych bezpośrednio przez badany sektor (grupę przedsiębiorstw) do wartości wytworzonej produkcji. Współczynniki te służą określeniu efektywności poszczególnych sektorów, ich znaczenia w kształtowaniu procesów rozwojowych w gospodarce (Czyżewski, 2011).

Powiązanie poszczególnych sektorów gospodarki z zagranicą można analizować przez pryzmat zmian udziału eksportu produktów danego sektora w łącznym, bądź finalnym popycie na produkty tego sektora, jak również z perspektywy jego importochłonności danego sektora. Pierwsza z wymienionych możliwości pozwala na ocenę zmian konkurencyjności zewnętrznej danego sektora. W przypadku natomiast współczynnika importochłonności (bezpośredniej) definiowanej jako wartość produktów zużytych bezpośrednio przez sektor, a pochodzących z importu, odniesioną do produkcji globalnej tego sektora, ocenić możemy znaczenie zasileń przez strumienie produktów z importu. Zależności te pozwalają na ustalenie znaczenia importu w rozwoju badanych sektorów w gospodarce.

Analiza tablicy przepływów międzygałęziowych pozwala także na ocenę efektywności makroekonomicznej poszczególnych sektorów. Rozumiana może być ona jako udział wartości dodanej brutto w łącznej wartości dodanej brutto wytworzonej w go-

spodarce, lub w produkcji globalnej danego sektora. Innym sposobem może być ocena relacji popytu finalnego do wartości strumieni zasilających badany sektor (precyzyjniej to efektywność powiązań międzygałęziowych) (Czyżewski, Grzelak, 2007). Wskaźniki te służą do badania pozycji konkurencyjnej danego sektora względem pozostałych sektorów. Pośrednio więc mogą wskazywać na transfer wypracowanych efektów i potencjalnych rent w danym sektorze do otoczenia (głównie poprzez system cen). Jednocześnie pozwalają na identyfikację źródeł procesów rozwojowych w gospodarce. Określają także poprawę, bądź pogorszenie rozwoju danego sektora przez pryzmat kreacji efektów dochodowych. Należy w tym miejscu podkreślić, że omawiane współczynniki efektywności powiązań międzygałęziowych obarczone są uproszczeniem wynikającym z techniczno-bilansowego ujęcia danych w bilansach przepływów międzygałęziowych. Chodzi tu o założenie jednoczesnej transformacji nakładów na efekty (Czyżewski, Helak 1991). Mimo to nie przekreśla to wnioskowania o zaistniałych tendencjach w tym zakresie. Tablica przepływów międzygałęziowych może być także wykorzystana do oceny dynamiki poszczególnych składników przy porównaniach wieloletnich (Grzelak, 2008). Chodzi tu w szczególności o te elementy, które związane są ze zmianami popytu finalnego na produkty danego sektora, wartości dodanej, strumieni dóbr i usług zasilających dany sektor. Istotne znaczenie w tym przypadku ma wycena poszczególnych składników, a zwłaszcza zastosowanie właściwych deflatorów przy porównaniach w cenach stałych. Oceny uwzględniające porównania wieloletnie umożliwiają określenie kierunków zmian procesów gospodarczych, w tym zwłaszcza znaczenia poszczególnych sektorów (Czyżewski, Grzelak 2007).

Istnieje także możliwość wykorzystania modelu input-output do ocen w układzie regionalnym. Pozwalają one na badania sytuacji badanych sektorów w zakresie omawianych powyżej wskaźników, ale na poziomie regionalnym (wojewódzkim). W przypadku takiego podejścia należy jednak liczyć się z istotnymi niedostatkami źródłowymi. Brak jest bowiem publikacji tablic przepływów międzygałęziowych w układzie regionalnym przez urzędy statystyczne. W związku z tym zachodzi potrzeba samodzielnej dekompozycji zawartych danych w tablicy przepływów międzygałęziowych na podstawie określonego podejścia: programowania matematycznego, czy współczynników lokalizacji w celu stworzenia stosownych macierzy (Malaga, 1992; Zawalińska, 2009).

W polskiej literaturze tematu istnieją przykłady wykorzystania przepływów input-output do ocen makroekonomicznych (Tomaszewicz, 1994) i badań regionalnych (Malaga, 1992; Tomaszewicz, Trębska, 2005; Zawalińska, 2009), jak i w relatywnie szerszym zakresie, np. sektora rolnego (Lonc, 1985; Czyżewski, Helak 1991; Kujaczyński, 2008; Czyżewski, Grzelak, 2009; Mrówczyńska-Kamińska, Czyżewski, 2011; Woś, 1973). Można jednak sądzić, że bilanse przepływów międzygałęziowych są wciąż w niewielkim zakresie wykorzystywane przez ekonomistów. Wynika to z jednej strony ze złożoności takich analiz, a z drugiej z ograniczeń o których mowa w dalszej części opracowania.

4. POTRZEBA MODYFIKACJI STATYSTYK BILANSÓW PRZEPŁYWÓW MIĘDZYGAŁĘZIOWYCH

Z perspektywy dotychczasowych doświadczeń związanych z wykorzystaniem bilansów przepływów międzygałęziowych do ocen makroekonomicznych możemy sformułować kilka zasadniczych postulatów. Przede wszystkim brakuje wypełnienia i odrębnej interpretacji ćwiartki IV tablicy przepływów. Istotnie ogranicza to możliwości analizy w obszarze podziału dochodu narodowego i związanych z tym procesami redystrybucji budżetowej. Na poziomie zagregowanych ocen umożliwiłoby to pełniejsze określenie roli budżetu państwa w procesach rozwojowych w gospodarce. Istotną kwestią dla rozumienia zjawisk gospodarczych w modelu przepływów międzygałęziowych są zagadnienia dotyczące bilansów. Mamy do czynienia z bilansami wewnątrz i między ćwiartkowymi. Wiążą się one z tożsamościami ekonomicznymi. I tak przykładowo wydatki jednych podmiotów są przychodami innych, a jednocześnie łączna suma przepływów międzygałęziowych równa jest sumie nakładów, z kolei dochody w układzie podmiotowym równe są dochodom z punktu widzenia ich źródeł. Zdecydowanie jednak priorytetową, ze względu na poruszaną problematykę badawczą, jest kwestia bilansów między ćwiartkami, czyli równowagi pomiędzy ćwiartkami II, III i IV. Zarówno popyt efektywny (zrealizowany), wytworzone dochody, jak i podzielone efekty powinny być sobie równe (por. np. teorię równowagi ogólnej Walrasa). Wynika to z tożsamości w rachunku makroekonomicznym. Tak więc strumienie wydatków podmiotów gospodarczych pokrywają się z łącznymi ich dochodami, a dokonuje się to poprzez mechanizm rynkowy. Jak wobec tego wyjaśnić nierównowagę podażową z perspektywy modelu przepływów międzygałęziowych? Gospodarka z reguły nie wykorzystuje swoich możliwości w zakresie pełnego wykorzystania zdolności produkcyjnych. W tym wypadku można mówić o równowadze, ale zazwyczaj w warunkach niepełnego wykorzystania mocy wytwórczych w gospodarce (luka popytowa). Sytuacja ta w gospodarce rynkowej objawia się nadwyżką produkcji nad popytem, czy wzrostem poziomu zapasów. Jednocześnie w modelu przepływów międzygałęziowych mogłoby to zostać uwidocznione w niezrealizowanych aspiracjach konsumentów i producentów w II ćwiartce. W gospodarce rynkowej II i IV ćwiartka są większe (co do aspiracji popytowych) od III. Skutkiem tego są nadwyżkowe zapasy, niewykorzystane zasoby produkcyjne (zwłaszcza pracy), a w gospodarce nierynkowej przymusowe oszczędności. Jest to również konsekwencją nienadążania opłaty pracy w odniesieniu do wzrostu wydajności tego czynnika oraz niższej wydajności podatkowej podmiotów gospodarczych (problem ukrywania dochodów, cen transferowych) (Czyżewski, Grzelak, 2009). W związku z powyższym interesujące byłoby podjęcie próby pokazania przybliżonej[4] nierównowagi poprzez przedstawienie (alternatywnie) w ćwiartce II popytu potencjalnego, tj. efektywnego i niezrealizowanych aspiracji. Możnaby tego dokonać z poprzez oszacowanie wartości wytworzonych dóbr finalnych, wraz z zapasami (oszczędnościa-

[4] Przybliżonej w tym sensie, że nie sposób jest zaprezentować wszystkie aspekty nierównowagi podażowej, takie chociażby jak niewykorzystane zasoby produkcyjne.

mi) i zestawienie tego z popytem zrealizowanym, zapisanym w tym przypadku w IV części tablicy pieniężnych przepływów międzygałęziowych. Niedosyt pozostawia także przedstawienie sfery odnoszącej się do relacji pieniężnych w gospodarce. Chodzi tu o szacunkowe zaprezentowanie (w ćwiartce III) wielkości depozytów i kredytów (lub nowo otwieranych/udzielanych) dla poszczególnych sektorów. Umożliwiłoby to bardziej kompleksowe określenie roli polityki pieniężnej i sektora banków w gospodarce, w tym w szczególności znaczenie kreacji pieniądza dla procesów gospodarczych. Z kolei dla ocen regionalnych wskazane byłoby opracowanie bilansów przepływów międzygałęziowych w układach regionalnych (wojewódzkich). Pozwoliłoby to na analizę procesów gospodarczych w ujęciu przestrzennym z uwzględnieniem zakresu powiązań pomiędzy regionami. Takie oceny z punktu widzenia kreowania polityki regionalnej byłyby szczególnie użyteczne. Publikowanie bilansów międzygałęziowych, co 5 lat i to ze znacznym przesunięciem czasowym (ok. 4-letnim) znacznie ogranicza wykorzystanie tych analiz jeśli chodzi o ich aktualność oraz bieżące monitorowanie zjawisk gospodarczych. Zdając sobie sprawę ze znacznego wysiłku organizacyjnego, nakładów związanych przy opracowaniu wyników modelu input-output ze strony GUS, a także ograniczonej ich bieżącej dostępności, stoimy na stanowisku, że możnaby realizować te publikacje szybciej, tak aby ukazywały się maksymalnie z 3-letnim opóźnieniem. Ważne też by dawały one podstawę dla pełnej oceny podziału wytworzonych dochodów.

5. ZAKOŃCZENIE

Powyższe rozważania skłaniają do kilku konkluzji:

1. model przepływów międzygałęziowych pozwala na pogłębienie analiz makroekonomicznych i stanowi interesujący instrument oceny i interpretacji zjawisk gospodarczych. Zasadniczym przesłaniem tablicy przepływów międzygałęziowych jest ukazanie wzajemnych zależności w gospodarce, które decydują o jej rozwoju jako całości, jak i jej części.
2. tablica przepływów międzygałęziowych pozwala nie tylko ująć powiązania między poszczególnymi sektorami gospodarki narodowej, ale umożliwia dokonanie kompleksowych ocen podstawowych relacji ekonomicznych, charakteryzujących strukturę badanych zjawisk i zachodzące między nimi współzależności, także w porównaniach w układzie dynamicznym i międzynarodowym. Ocena danych sektorów z perspektywy modelu input-output umożliwia w szczególności poszerzenie perspektywy badawczej na zagadnienia dotyczące ich pozycji w gospodarce, efektywności makroekonomicznej, procesów rozwojowych, zależności międzygałęziowych, poziomu rynkowego zrównoważenia.
3. istnieje potrzeba poszerzenia zapisu i interpretacji tablicy przepływów o ćwiartkę czwartą (podział dochodów), szersze uwzględnienie powiązań badanych podmiotów z sektorem bankowym i tym samym pokazanie skali i struktur kreacji pieniądza emisyjnego i kredytowego w gospodarce. Do rozważenia pozostaje także wykorzystanie tablicy input-output do określenia zakresu nierównowagi podażo-

wej w gospodarce poprzez zestawienie popytu potencjalnego i efektywnego oraz określenie w ten sposób przybliżonego poziomu nierównowagi. Rodzi to potrzebę częstszych i pełniejszych publikacji GUS bilansów przepływów międzygałęziowych w gospodarce.

Uniwersytet Ekonomiczny w Poznaniu

LITERATURA

[1] Czyżewski A., (2011), *Przepływy międzygałęziowe jako makroekonomiczny model gospodarki*, Wyd. UE w Poznaniu, Poznań.
[2] Czyżewski A., Grzelak A., (2009), *Możliwości oceny rozwoju rolnictwa w warunkach globalnych z zastosowaniem tabeli przepływów międzygałęziowych*, Roczniki Naukowe SERiA, Tom XI, z. 2.
[3] Czyżewski A., Grzelak A., (2007), *The use of Input-Output to Evaluate the Agriculture Situation in Poland after 1990*, Management, Vol.11, No 2.
[4] Czyżewski A., Helak K., (1991), *Przekształcenia w kompleksie gospodarki żywnościowej w Polsce*, Wieś i Rolnictwo, nr 3.
[5] Grzelak A., (2008), *Związki gospodarstw rolnych z rynkiem w Polsce po roku 1990. Próba określenia intensywności i efektywności*, Wyd. AE w Poznaniu, Poznań.
[6] Kujaczyński T., (2008), *Zmiany struktur wytwórczych w gospodarce żywnościowej w świetle przepływów międzygałęziowych*, Roczniki Ekonomiczne Kujawsko-Pomorskiej Szkoły Wyższej, Bydgoszcz.
[7] Leontief W., (1936), *Quantitative input and output relations in the economic system of the United States*, The Review of Economics and Statistics, vol. XVIII, August.
[8] Lonc T., (1985), *Związki rolnictwa z gospodarką narodową na początku lat 80-tych*, Zagadnienia Ekonomiki Rolnej, nr 6.
[9] Malaga K., (1992), *Struktura produkcyjna gospodarki żywnościowej Wielkopolski w świetle przepływów międzygałęziowych*, [w:] Gospodarka żywnościowa w Polsce i regionie, Czyżewski A. (red.), Wyd. PWE, Warszawa.
[10] Mrówczyńska-Kamińska A., Czyżewski B., (2011), *Zaopatrzenie materiałowe rolnictwa w Polsce i Niemczech w świetle bilansów przepływów międzygałęziowych*, Roczniki SERiA, Tom XIII, z. 3.
[11] *Opracowanie Biura Ekonomicznego NBP,* (1958), NBP, Warszawa.
[12] Quesnay F., (1928), *Pisma wybrane* (tłum. B.J. Pietkiewiczówna), Gebetnber i Wolf, Warszawa.
[13] Tomaszewicz Ł., (1994), *Metody analizy input-output*, PWE, Warszawa.
[14] Tomaszewicz Ł., J. Trębska, (2005), *Regional and Interregional Input-Output Tables for Poland*, [w:] Modeling Economies in Transition, Welfe W., Wdowiński P. (red.), Wyd. UŁ, Łódź.
[15] Woś A., (1973), *Rolnictwo w bilansie przepływów międzygałęziowych*, Zagadnienia Ekonomiki Rolnictwa, nr 1.
[16] Zawalińska K., (2009), *Regionalne efekty wsparcia Unii Europejskiej dla rozwoju obszarów wiejskich*, IRWiR, Warszawa.

MOŻLIWOŚCI WYKORZYSTANIA STATYSTYKI BILANSÓW PRZEPŁYWÓW MIĘDZYGAŁĘZIOWYCH DLA MAKROEKONOMICZNYCH OCEN W GOSPODARCE

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

W artykule zaprezentowano możliwości wykorzystania modelu przepływów międzygałęziowych w zakresie ocen makroekonomicznych gospodarki. Zasadniczym przesłaniem bilansu przepływów międzyga-

łęziowych jest ukazanie wzajemnych zależności w gospodarce, które decydują o jej rozwoju jako całości, jak i jej części. Stwierdzono, że oceny dokonywane za pośrednictwem bilansów przepływów międzygałęziowych umożliwiają i poszerzają perspektywę badawczą uwzględniając pozycję badanych sektorów (grup produktów) w gospodarce, a także ich efektywność makroekonomiczną i współzależności w procesie rozwoju. Dostrzeżono, że model przepływów międzygałęziowych wykazuje znaczące walory dla pełniejszego zrozumienia funkcjonowania mechanizmu rynkowego i budżetowego, schematu przepływu strumieni dochodów i wydatków w gospodarce oraz kształtowania się podstawowych parametrów makroekonomicznych. Istnieje także potrzeba poszerzenia zapisu i interpretacji tablicy przepływów międzygałęziowych o ćwiartkę czwartą (podział dochodów), szersze uwzględnienie powiązań badanych podmiotów z sektorem bankowym i tym samym pokazanie skali oraz struktur kreacji pieniądza emisyjnego i kredytowego w gospodarce. Do rozważenia pozostaje także wykorzystanie tablicy input-output do określenia zakresu nierównowagi podażowej w gospodarce poprzez zestawienie popytu potencjalnego i efektywnego oraz określenie w ten sposób przybliżonego poziomu nierównowagi.

Słowa kluczowe: model przepływów międzygałęziowych, gospodarka, makroekonomia

POSSIBILITIES OF USING THE STATISTIC OF INPUT-OUTPUT IN MACROECONOMICS EVALUATION OF THE ECONOMY

A b s t r a c t

The possibilities of using of the input-output table for macroeconomics evaluation of economy was presented in the article. The principal message of input-output model is appearance mutual interdependence in economy, which decide about her development as the whole, and as her part. It has stated that the evaluation conducted by use of this model makes possible extension of investigative perspective regarding such questions as: position of a examined sectors in the economy, macroeconomics efficiency, and interdependences in developmental processes. One had noticed that the input-output model points out considerable values in for fuller understanding of functioning of the market and budget mechanism, the roundabout pattern of streams of incomes and the expenses in economy, and the shaping of the basic macroeconomics parameters. There is a need extension of record and the interpretation of input-output model about fourth quarter (the division of incomes), the wider regard of connections of studied subjects with bank sector and the same show the scale and the structures of creation of issue and credit money in the economy. It stays for considering also the use of this model for evaluation the range of the supply non-equilibrium in the economy by composition of potential demand and effective and qualification in this way an approximate level the non-equilibrium.

Key words: the input-output model, economy, macroeconomics

SPIS TREŚCI

Numer Specjalny 1

Numer Specjalny 2

CONTENTS

Special issue 1

Special issue 2