Mirosław Szreder
Katedra Statystyki
Wydział Zarządzania
Uniwersytet Gdański
miroslaw.szreder@ug.edu.pl

Rola badań statystycznych w naukach ekonomicznych w świetle nowych możliwości określanych mianem *big data*

Komitet Statystyki i Ekonometrii PAN
Warszawa, 14 marca 2018

Współczesne trendy i wyzwania

1. Rosnąca popularność badań próbkowych (niewyczerpujących) w ekonomii

Zagrożenia:

A) Skłonność do zastępowania właściwszych w danej problematyce badań jakościowych badaniami statystycznymi (ankietowymi);

B) Nieświadomość zmian w strukturze całkowitego błędu badania próbkowego (w tym reprezentacyjnego) wśród części badaczy.

Wzrost w ostatnich latach roli błędów nielosowych:

- błędu pokrycia (*coverage error*) – niekompletny lub złej jakości operat losowania,
- błędu spowodowanego odmowami respondentów udziału w badaniu (*nonresponse error*),
- błędu pomiaru (*measurement error*) – zarejestrowaniem nieprawdziwych danych z winy ankietera lub respondenta,
- błędu przetwarzania zgromadzonych danych (*postsurvey processing error*).

Powszechność stosowania badań statystycznych, a także trudności z ograniczeniem oddziaływania błędów nielosowych, powodują obniżanie się jakości badań i spadek zaufania do ich wyników.

Baker, H.K., Mukherjee T.K. (2007), Survey Research in Finance: Views from Journal Editors, *International Journal of Managerial Finance*, Vol. 3(1), s.11-25.

Tablica 1. Rola badań ankietowych w czasopismach naukowych z zakresu finansów – opinie redaktorów czasopism (odpowiedzi udzieliło jedynie 23 spośród 50 redaktorów czasopism).

Które z następujących stwierdzeń najlepiej opisuje rolę, jaką powinny odgrywać badania próbkowe w literaturze z zakresu finansów	Redaktorzy kluczowych czasopism	Redaktorzy pozostałych czasopism	Ogółem (n)	Ogółem (%)
A. Badanie próbkowe powinno być traktowane na równi z innymi oryginalnymi badaniami.	-	10	10	43,5
B. Badanie próbkowe powinno pełnić rolę uzupełniającą względem innego oryginalnego badania.	4	6	10	43,5
C. Rola badania próbkowego jest ograniczona (lub nie ma ono żadnego znaczenia) w stosunku do innych oryginalnych badań.	2	1	3	13,0
D. Rola badania próbkowego powinna być następująca (podaj): …	-	-	-	-

Źrodło: [Baker i Mukherjee, 2007, s. 21].

Współczesne trendy i wyzwania

2. Integracja źródeł danych w badaniach statystycznych

Tendencja ta kształtowana jest przez czynniki popytowe i podażowe.

Popyt na dodatkową informację
→ dążenie do większej dokładności badań (techniki ważenia, kalibracji),
→ eliminowanie wpływu błędów nielosowych (i większa świadomość tych błędów)

Stwierdzenie Lesliego Kisha:

"*sampling error is «over-researched«*"

z artykułu Richarda Platka i Carla-Erika Särndala pt. *Can a statistician deliver?*, "Wiadomości Statystyczne", 2001, nr 4, dobrze charakteryzuje zmianę akcentów w analizie błędów badań próbkowych.

Podaż informacji i danych:

- rejestry urzędowe (*register-based statistics*)

  cyt. z publikacji GUS: „Statystyka publiczna – współczesne oblicze” (s. 34)

  „Tam, gdzie to tylko możliwe, staramy się ograniczać zbieranie danych bezpośrednio od obywateli i przedsiębiorców, wykorzystując nowe, alternatywne źródła danych. Zgodnie ze światową tendencją, w ostatnich latach zintensyfikowaliśmy pozyskiwanie danych z systemów administracyjnych i pozaadministracyjnych.”.

- metadata – ogół informacji o zebranych danych statystycznych (ang. data about the data), czyli w szczególności ich struktura, zakres i kontekst, np. instrumenty pomiarowe, instrukcje dla ankieterów, sposoby pomiaru sondażowego, programy do przetwarzania danych;
- paradata – zbiór szczegółowych informacji towarzyszących badaniu, na ogół trudnych do zarejestrowania, lecz użytecznych, np. czas, jaki potrzebował respondent na odpowiedzi w poszczególnych pytaniach – czas między kliknięciami w ankiecie komputerowej;

Wyzwanie – brak ugruntowanego i powszechnie akceptowanego podejścia teoretycznego.

Przez *big data* rozumie się najczęściej taki sposób zdobywania nowej wiedzy i poznawania otaczającej nas rzeczywistości, który może być zrealizowany w dużej skali, dzięki nowym możliwościom gromadzenia i przetwarzania wielkich zbiorów danych.

Wyzwania dla statystyki

Duże zbiory danych liczbowych to szansa dla statystyki, ale i zagrożenie.

Są one rzadko jednorodne i nie jest uzasadnione założenie, iż dane te stanowią realizację ciągu zmiennych losowych o jednakowych rozkładach.

Jednak ograniczanie się we wnioskowaniu statystycznym jedynie do podzbiorów danych jednorodnych pozbawia statystyka możliwości wiarygodnego opisania danego zjawiska lub procesu, a także trafnego przewidywania jego rozwoju w przyszłości.

Alan Greenspan doszukujący się przyczyn kryzysu (2007-2008) o modelach ryzyka:

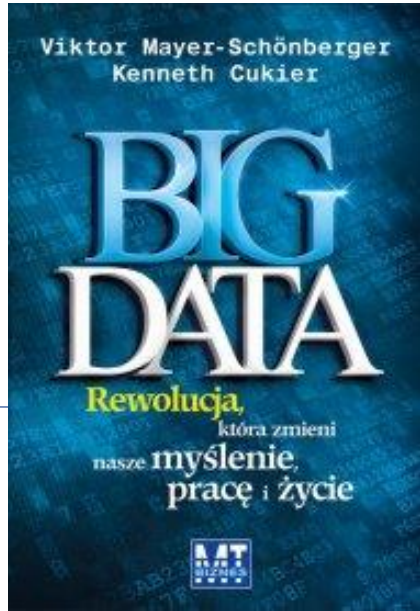
„Gdyby modele te były poprawniej dopasowane do danych historycznych, obejmujących także okresy załamania gospodarczego, określiłyby one wymogi kapitałowe na znacznie wyższym poziomie, a świat finansowy byłby teraz w znaczne lepszym stanie”.

(*„Had instead models been fitted more appropriately to historic periods of stress, capital requirements would have been much higher, and the financial world would be in far better shape, in my judgment”*)

Analityczna strona big data sprowadza się przede wszystkim do badania powiązań, współzależności i korelacji.

„Przewidywania oparte na korelacji są sercem big data”

Mayer-Schönberger V., Cukier K., „BIG DATA, Rewolucja, która zmieni nasze myślenie, pracę i życie, Wyd. MT Biznes, Warszawa 2013



Zagrożenia:

- pozorne lub sztuczne korelacje (ang. *spurious correlations*),

- próba ograniczenia celów poznania:

„W big data ważna jest odpowiedź na pytanie, co się dzieje, a nie dlaczego. Nie zawsze musimy znać przyczyny jakiegoś zjawiska, możemy po prostu pozwolić danym mówić za siebie" (s. 30).

- niepełna adekwatność technik wnioskowania statystycznego do bardzo dużych zbiorów danych z próby.

Wnioski

Oba podejścia: klasyczne badania statystyczne oraz techniki eksploracji dużych zbiorów danych (*big data*) będą jeszcze przez dłuższy okres czasu komplementarne względem siebie.

W niektórych zastosowaniach (np. w predykcji) *big data* wypiera stopniowo badania próbkowe.

W tych zagadnieniach z kolei, w których wymagana jest duża precyzja szacunku, wiodące pozostaną badania reprezentacyjne.

Dziękuję za uwagę!