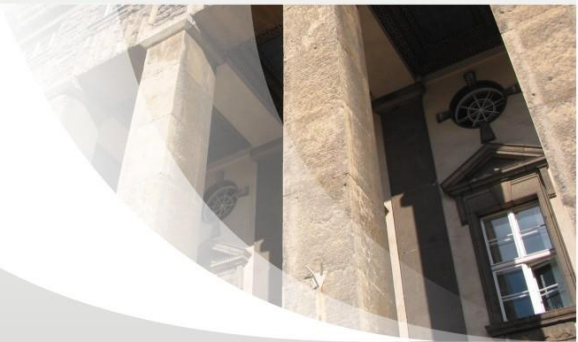
Katedra Statystyki
Wydział Informatyki i Gospodarki Elektronicznej

UNIWERSYTET EKONOMICZNY
W POZNANIU

Statystyczna integracja danych w badaniach społeczno-ekonomicznych

Wojciech Roszka
Uniwersytet Ekonomiczny w Poznaniu

UNIWERSYTET EKONOMICZNY
W POZNANIU



Plan wykładu

1. Problemy badawcze
2. Integracja danych w badaniach statystycznych
3. Przykłady zastosowań integracji danych w statystyce publicznej
4. Statystyczna integracja danych
5. Podsumowanie

Problemy badawcze

- Zwiększający się popyt na rzetelną i wielowymiarową informację na szczeblu lokalnym.
- Towarzyszące temu wzrostowi ograniczenia budżetowe.
- Jakość i bezpieczeństwo danych.
- Zwiększający się szum informacyjny powodujący zakłócenia w pomiarze zjawisk społeczno-ekonomicznych („zła” informacja wypiera „dobrą”).

Problemy badawcze

- Wymogi obniżenia kosztów zbierania i przetwarzania danych;
- Efektywność wykorzystania istniejących źródeł informacji;
- Zwiększenie zasobów informacyjnych istniejących źródeł danych;
- Obniżenie obciążenia respondentów;
- Zmiany zachodzące w podejściu do statystyki publicznej;

Integracja danych w badaniach statystycznych

Idea integracji danych

- Celem integracji jest stworzenie nowego, bogatszego, zbioru danych, opisującego pewną populację docelową.
- Ze względu na różne architektury systemów administracyjnych oraz ich niestatystyczne przeznaczenie, rejestry wymagają dostosowania do potrzeb organów statystyki publicznej.
- Łączenie następuje na podstawie unikalnego klucza połączeniowego, np. numer PESEL, dane adresowe.

Źródła danych

1. Spis powszechny
2. Badania reprezentacyjne (statystyki publicznej i spoza systemu statystyki publicznej)
3. Rejestry administracyjne

Spis powszechny

1. **Zalety**
   - badanie pełne (z częścią reprezentacyjną)
   - umożliwia tworzenie szczegółowych charakterystyk dla małych jednostek terytorialnych
   - wykorzystanie rejestrów administracyjnych

2. **Wady**
   - przeprowadzany raz na dekadę
   - długi czas przetwarzania danych i publikacji wyników
   - problemy z częścią reprezentacyjną

Badania reprezentacyjne statystyki publicznej

1. **Zalety**

- przeprowadzane cyklicznie
- zharmonizowana metodologia, porównywalne
- możliwość wnioskowania na całą populację generalną

2. **Wady**

- szczegółowość ograniczona liczebnością próby
- brak łącznej obserwacji wszystkich zjawisk społeczno-ekonomicznych (jedno badanie – jedna grupa zjawisk)
- wysokie koszty
- błędy nielosowe

Badania reprezentacyjne statystyki publicznej

Nazwa	Rekord/ jednostka	Liczba rekordów	Wybrane zmienne/zagadnienia
BAEL	osoba w wieku 15 lat i więcej	111 807	płeć
			wiek
			wykształcenie
			stan cywilny
			status na rynku pracy
			wymiar czasu pracy
			zawód wykonywany
	gospodarstwo domowe	59 994	źródło utrzymania gosp. dom.
			typ gospodarstwa dom.
BBGD	osoba	107 124	płeć
			wiek
			stan cywilny
			status na rynku pracy
			wykształcenie
	gospodarstwo domowe	34 767	wydatki gosp. dom. w przekroju szczegółowych kategorii
			dochody gosp. dom. w przekroju szczegółowych kategorii
			charakterystyka lokalu zajmowanego przez gosp. dom.
			wyposażenie gosp. dom.
			wielkość spożycia towarów i usług w gosp. dom.

Nazwa	Rekord/ jednostka	Liczba rekordów	Wybrane zmienne/zagadnienia
EU-SILC	osoba w wieku 16 lat i więcej	36 590	płeć
			wiek
			stan cywilny
			wykształcenie
			dostęp do różnych usług
			dochody osobiste w przekroju szczegółowych kategorii
			różne aspekty jakości życia
	gospodarstwo domowe	14 914	dochody gosp. dom. w przekroju szczegółowych kategorii
			charakterystyka lokalu zajmowanego przez gosp. dom
			wyposażenie gosp. dom.
			subiektywna sytuacja materialna
			warunki mieszkaniowe
			wskaźnik ubóstwa materialnego
DS	osoba w wieku 16 lat i więcej	26 420	różne charakterystyki społeczno-demograficzne
			stan zdrowia
			systemy wartości i postawy społeczne
			opinie dotyczące otoczenia społeczno-gospodarczego
	gospodarstwo domowe	12 387	sytuacja dochodowa
			warunki mieszkaniowe
			uczestnictwo w kulturze i wypoczynek
			różne aspekty wykluczenia społecznego

Rejestry administracyjne

1. **Zalety**
   - wysokie pokrycie
   - brak konieczności przeprowadzania pomiaru
   - bogata zawartość merytoryczna

2. **Wady**
   - cel powstania to podejmowanie decyzji administracyjnych
   - definicje zmiennych i wariantów wynikają z aktów prawnych
   - brak kontroli statystycznej błędów
   - brak standaryzacji między gestorami
   - bardzo ograniczona dostępność

Rejestry administracyjne

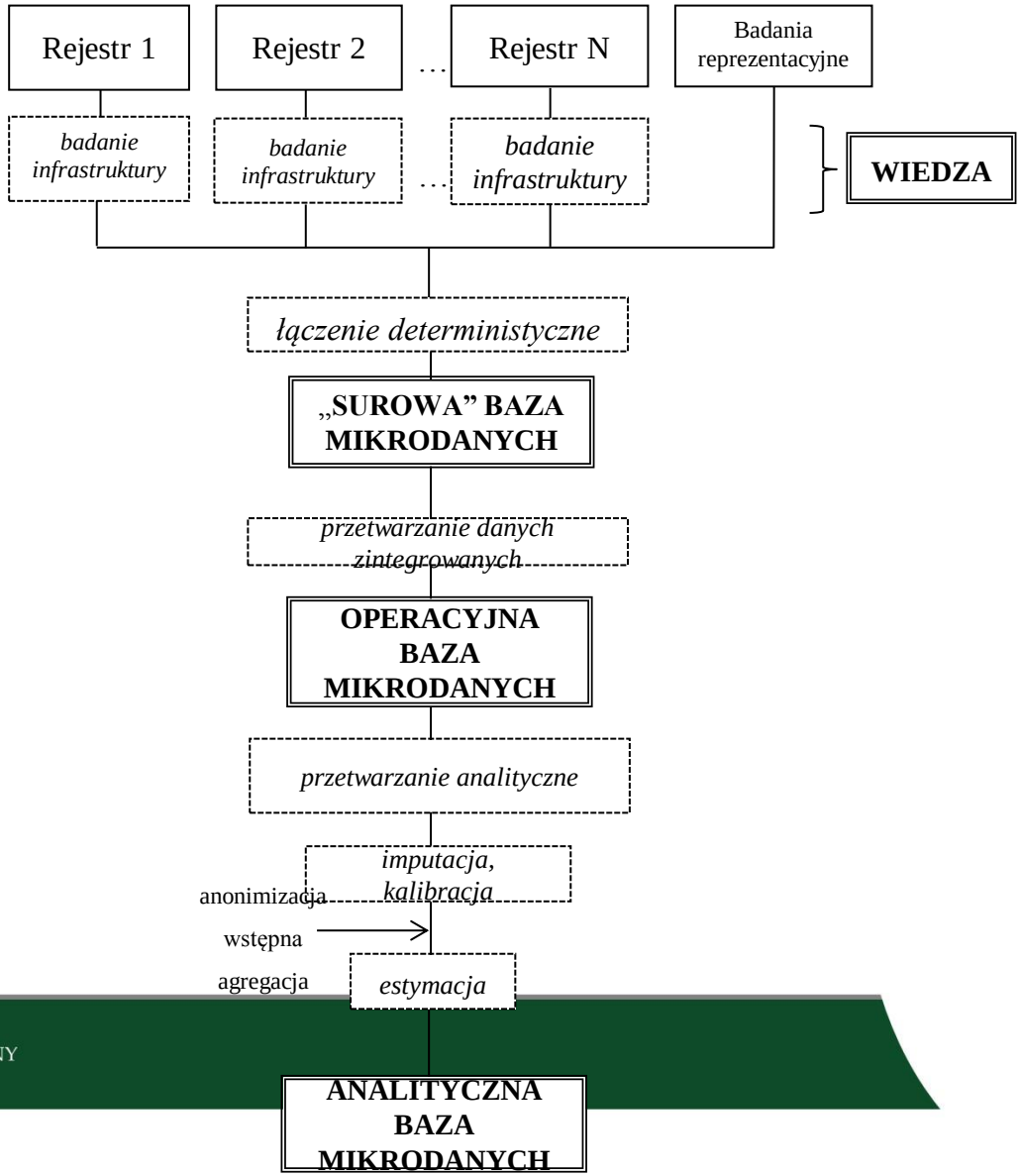
Charakterystyka		System administracyjny	System statystyczny
1. Cel powstania		Podejmowanie decyzji administracyjnych	Dokonywanie szacunków i analiz
2. Zbiorowość		Wszystkie podmioty (jednostki) prawnie podległe danemu gestorowi	Wszystkie podmioty (jednostki) objęte badaniem statystycznym
3. Jednostka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

Źródło: opracowanie własne na podstawie [Penneck 2007]

Rejestry administracyjne

Nazwa	Rekord/ jednostka	Liczba rekordów	Wybrane zmienne/zagadnienia
PESEL	osoba	ok. 38,7 mln	płeć
			data urodzenia
			stan cywilny
			adres
ZUS	płatnik ubezpieczenia społecznego	16 214 456	płeć
			data urodzenia
			adres
			status na rynku pracy
			wymiar czasu pracy
			status emerytalno-rentowy
			stopień niepełnosprawności
NFZ	osoba	38 647 138	płeć
			data urodzenia
			adres
			status na rynku pracy[1]
POLTAX	pracujący, emeryci, renciści	ok. 19 mln	adres miejsca zamieszkania
			adres miejsca pracy
			przychód
			dochód
			wysokość podatku
			wysokość składek ubezpieczeniowych

Integracja repozytoriów danych pochodzących z różnych źródeł



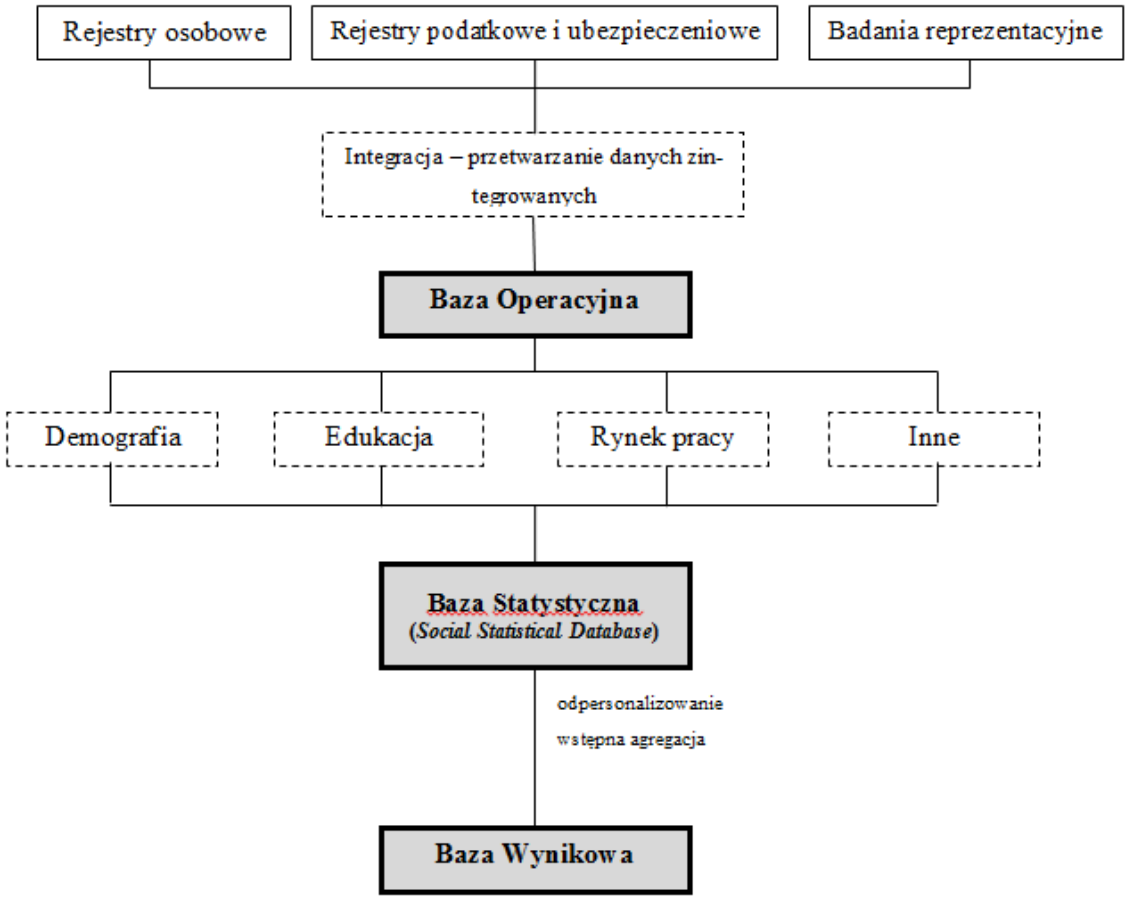
Integracja danych - problemy

- W „**klasycznym**” podejściu do integracji potrzebny jest unikatowy **klucz** połączeniowy – dostępny w każdym źródle.

- Przy integracji źródeł pełnych (np. rejestrów administracyjnych) z częściowymi (np. badaniami reprezentacyjnymi) **łączna obserwacja** różnych charakterystyk zapewniona jest **jedynie dla jednostek poddanych pomiarowi**.

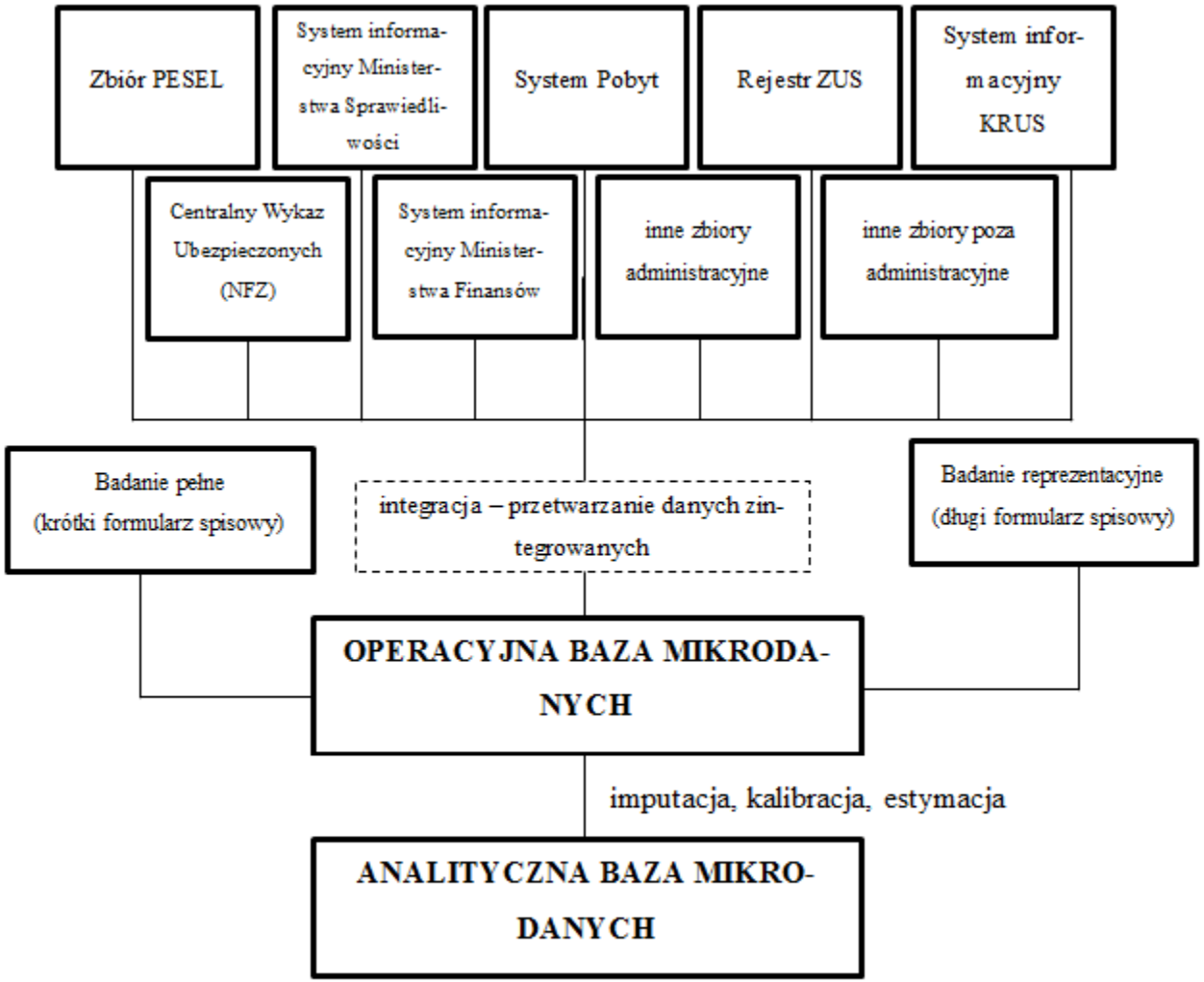
Przykłady zastosowań integracji danych w statystyce publicznej

Spis wirtualny w Holandii



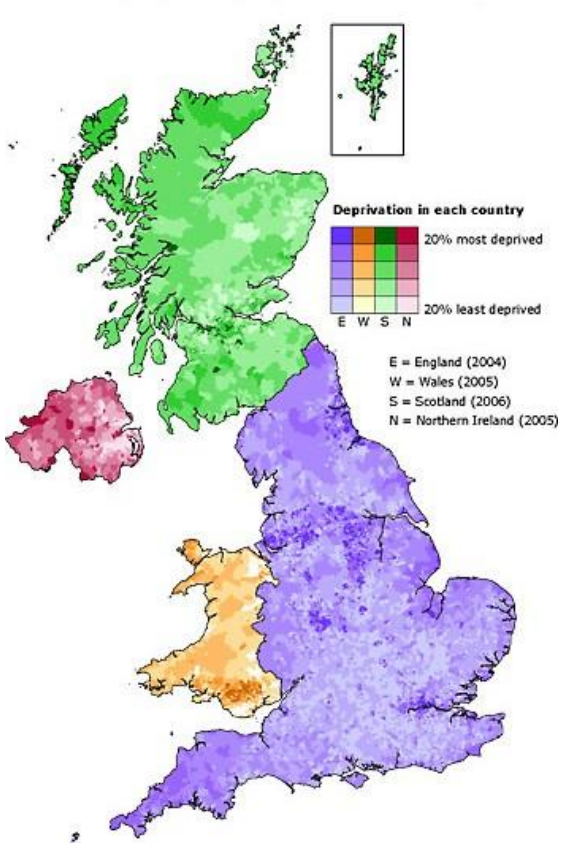
Źródło: opracowanie własne na podstawie [Everaers, van der Laan 2003]

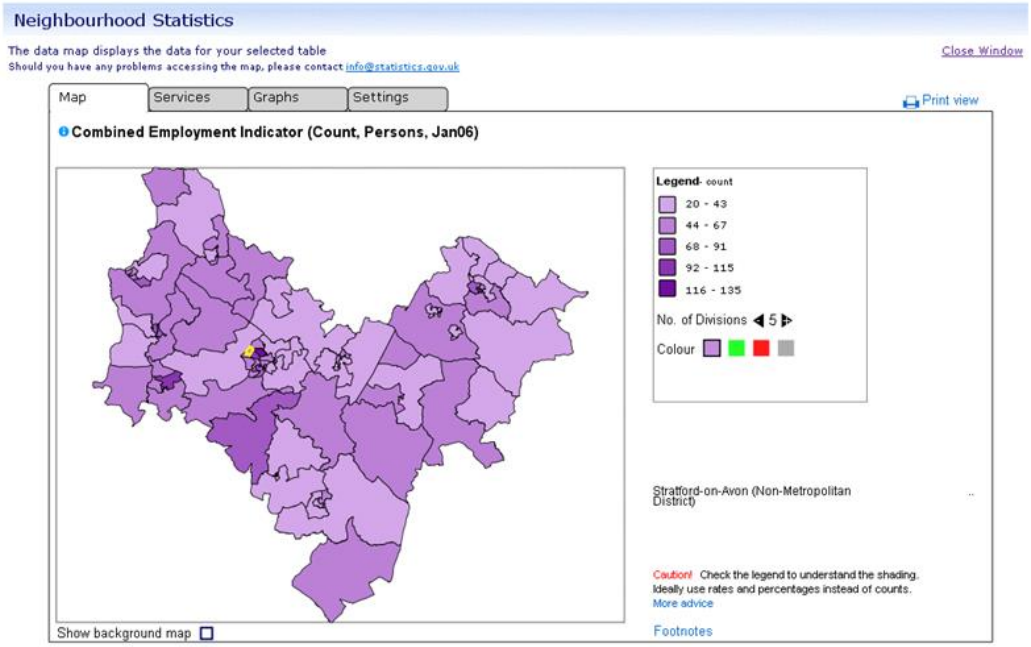
NSP 2011



Źródło: opracowanie własne na podstawie [Dygaszewicz 2011]

System statystyki sąsiedztwa w Wielkiej Brytanii





Statystyczna integracja danych

Klasyfikacja metod

Wyróżnia się dwie podstawowe grupy metod statystycznej integracji danych:

1. **Łączenie rekordów (*record linkage*)**

- Integrowane bazy danych zawierają informacje o tych samych jednostkach (np. rejestry administracyjne, dane spisowe).
  - **deterministyczne łączenie rekordów (*exact record linkage*)**

    Możliwe do zastosowania w przypadku, gdy integrowane zbiory danych zawierają unikalny klucz połączeniowy, który nie posiada braków danych ani błędów (np. PESEL)
  - **probabilistyczne łączenie rekordów (*probabilistic record linkage*)**

    Stosuje się, gdy unikalny klucz połączeniowy nie jest dostępny (np. usunięty w celu ochrony danych osobowych) lub gdy jest dostępny, ale zawiera braki danych lub różnego rodzaju błędy. Łączenie odbywa się na podstawie oszacowania prawdopodobieństwa, że porównywana para rekordów należy do tej samej jednostki.

2. **Parowanie statystyczne (*statistical matching, data fusion*)**

- Grupa metod służących do integracji dwóch (lub więcej) źródeł danych (zwykle pochodzących z **badań próbkowych**) odnoszących się do tej samej populacji. Celem jest analiza związków pomiędzy zmiennymi nieobserwowanymi łącznie w pojedynczym źródle.

Probabilistyczne łączenie rekordów

ID	Nazwa	Adres	Nr telefonu
1432	Świnka sp. j.	ul. Mickiewicza 1a	+22 7456969
1433	Rowerek	Trzeciego Maja 15	591987321
1434	Bucik sp. z o.o.	al. Niepodległości 10	581596325

ID_rec	Nazwa	Adres	Nr telefonu
D1215	Rowerk	3 Maja 15	12 591-987-321
D1354	Świnka	Mickiewicza 1	745-69-69
D1236	Bucik	Niepodległości 10/1	14 581596325

Źródło: opracowanie własne na podstawie [Fortini *et al.* 2006]

Probabilistyczne łączenie rekordów

Głównym zadaniem metody probabilistycznego łączenia rekordów jest ustalenie, **czy para rekordów należy do tego samej jednostki czy nie**.

Decyzję tę podejmuje się najczęściej na podstawie **prawdopodobieństwa** (lub jego przekształceń), że dana para rekordów należy (lub nie) do tej samej jednostki.

Najczęściej tworzy się pewne miary, zwane **wagami zgodności i niezgodności**.

Probabilistyczne łączenie rekordów

Niech ***m*** oznacza **empiryczne prawdopodobieństwo zgodności** wartości zmiennych parujących przy założeniu, że porównywane rekordy należą do tej samej jednostki.

Niech ***u*** oznacza **empiryczne prawdopodobieństwo niezgodności** wartości zmiennych parujących przy założeniu, że porównywane rekordy nie należą do tej samej jednostki.

$$w_z = \frac{ln\left(\frac{m}{u}\right)}{ln2}$$

$$w_n = \frac{ln\left(\frac{1-m}{1-u}\right)}{ln2}$$

Wagi oblicza się dla wszystkich cech, na podstawie których dokonuje się integracji. Następnie wagi się sumuje otrzymując **wagę łączną - R**.

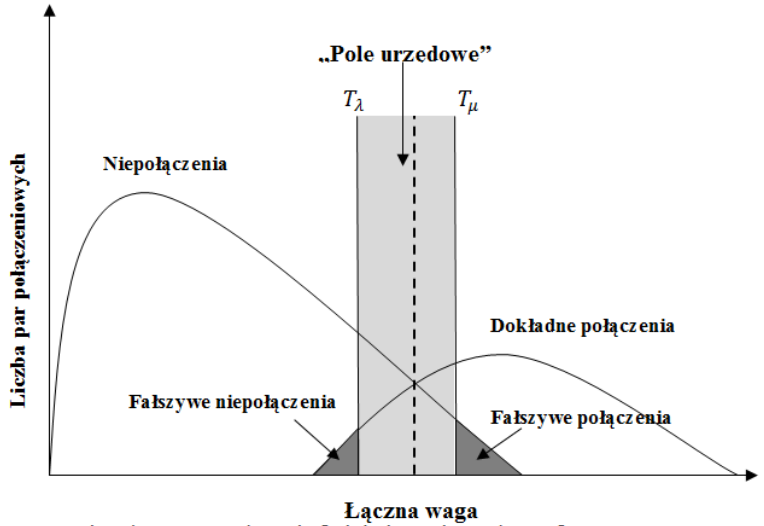
Probabilistyczne łączenie rekordów

Wartość R jest następnie porównywalna z wartościami progowymi T_μ i T_λ, które są określone granicznymi błędami a *priori* na, odpowiednio, fałszywe połączenie i fałszywe niepołączenie. Jeżeli spełniony będzie warunek:

$R > T_\mu$ – to para jest uważana za dokładne połączenie,

$T_\lambda \leq R \leq T_\mu$ – to połączenie jest możliwe; przedział ten nazywany jest „polem niedecyzyjnym” lub „polem urzędowym”,

$R < T_\lambda$ – to para uznawana jest za niepołączenie.



Źródło: opracowanie własne na podstawie [Blakely, Salmond 2002]

Probabilistyczne łączenie rekordów

- Procedura probabilistycznego łączenia rekordów narażona jest na błędy analogiczne jak w przypadku wnioskowania statystycznego: zaklasyfikowanie jako niepołączenie pary rekordów w rzeczywistości odnoszące się do tej samej jednostki (błąd I rodzaju) oraz, przeciwnie, zaklasyfikowanie jako połączenie pary rekordów nie odnoszącej się do tej samej jednostki (błąd II rodzaju).

	Dokładne połączenie	**Niepołączenie**	**Wartość predyk-cyjna**
Prawdopodobne połączenie	Prawdziwie pozytywne - n_m	Fałszywie pozytywne (błąd I rodzaju) - n_{fp}	**Dodatnia** $\frac{n_m}{(n_m+n_{fp})}$
Prawdopodobne niepołączenie	Fałszywie negatywne (błąd II rodzaju) - n_{fn}	Prawdziwie negatyw-ne - n_u	**Ujemna** $\frac{n_u}{(n_u+n_{fn})}$
Suma	N_m	N_u	
	Czułość $\frac{n_m}{N_m}$	**Swoistość** $\frac{n_u}{N_u}$	

Źródło: opracowanie własne na podstawie na podstawie [Blakely, Salmond 2002]

Parowanie statystyczne

Zbiór A

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

Zbiór B

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

Źródło: opracowanie własne

Y_1	...	Y_Q	X_1	...	X_P	Z_1	...	Z_R
y^A_{11}	...	y^A_{1Q}	x^A_{11}	...	x^A_{1P}	$\tilde{z}^B_{11}$	...	$\tilde{z}^B_{1R}$
...	...	...	...	...	...	...	...	...
y^A_{a1}	...	y^A_{aQ}	x^A_{a1}	...	x^A_{aP}	$\tilde{z}^B_{b1}$	...	$\tilde{z}^B_{bR}$
...	...	...	...	...	...	...	...	...
$y^A_{n_A1}$	...	$y^A_{n_AQ}$	$x^A_{n_A1}$	...	$x^A_{n_AP}$	$\tilde{z}^B_{n_A1}$	...	$\tilde{z}^B_{n_AR}$
$\tilde{y}^A_{n_A+1,1}$	...	$\tilde{y}^A_{n_A+1,Q}$	$x^B_{n_A+1,1}$	...	$x^B_{n_A+1,P}$	$z^B_{n_A+1,1}$	...	$z^B_{n_A+1,R}$
...	...	...	...	...	...	...	...	...
$\tilde{y}^A_{n_A+b,1}$	...	$\tilde{y}^A_{aQ}$	$x^B_{n_A+b,1}$	...	$x^B_{n_A+b,P}$	$z^B_{n_A+b,1}$	...	$z^B_{n_A+b,R}$
...	...	...	...	...	...	...	...	...
$\tilde{y}^A_{n_A+n_B,1}$	...	$\tilde{y}^A_{n_A+n_B,Q}$	$x^B_{n_A+n_B,1}$	...	$x^B_{n_A+n_B,P}$	$z^B_{n_A+n_B,1}$	...	$z^B_{n_A+n_B,R}$

Źródło: opracowanie własne

Parowanie statystyczne

Ponieważ zmienne $\boldsymbol{Y}$ oraz $\boldsymbol{Z}$ nie są łącznie obserwowane w żadnym ze źródeł, w procesie estymacji związków pomiędzy tymi cechami zwykle przyjmuje się założenie, że zmienne $\boldsymbol{Y}$ i $\boldsymbol{Z}$ są warunkowo niezależne przy danym $\boldsymbol{X}$ – założenie o warunkowej niezależności (*conditional independence assumption, CIA*):

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})$$

Przetestowanie prawdziwości CIA jest niemożliwe przy wykorzystaniu informacji $A \cup B$. Wyróżnia się tutaj trzy możliwości postepowania:

- wykorzystanie dodatkowych źródeł informacji (np. wcześniejszych doświadczeń lub za przeprowadzenia pomocniczego badania), które potwierdzą prawdziwość CIA,
- użycie dodatkowych źródeł informacji w toku integracji,
- w przypadku braku dodatkowych informacji o związkach między ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) – rozważenie tzw. niepewności (*uncertainty*) dla właściwości modelu integracji.

Parowanie statystyczne

Zasadniczo wyróżnia się **trzy metody łączenia** zbiorów danych:

1. Parametryczne

2. Nieparametryczne

3. Mieszane

Parowanie statystyczne

Metody parametryczne:

- **Imputacja regresyjna**

Do zbioru A imputowane są wartości teoretyczne wynikające z modelu:
$\hat{z}_a^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a,\ a = 1,2,\dots,n_A.$
Do zbioru *B* imputowane są wartości teoretyczne wynikające z modelu:
$\hat{y}_b^{(B)} = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b,\ b = 1,2,\dots,n_B.$
Konkatenacja zbiorów A i B : $S = A \cup B;\ n_S = n_A + n_B.$

- **Stochastyczna imputacja regresyjna** - do wartości teoretycznych wynikających z modeli regresji dolosowane są wartości z określonego rozkładu

$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a + e_a$
gdzie $e_a \sim N\left(0, \hat{\sigma}_{Z|X}\right)$
$\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b + e_b$
gdzie $e_b \sim N\left(0, \hat{\sigma}_{Y|X}\right).$

Parowanie statystyczne

Metody parametryczne:

- **Wielokrotna imputacja**
  - Każdy brak danych zastępowany jest za pomocą wielu (*m*) wartości.
  - Te *m* wartości są uporządkowane w takim sensie, że pierwszy zestaw wartości tworzy pierwszy zbiór danych itd.
  - Oznacza to, że dla *m* wartości tworzonych jest *m* kompletnych zbiorów danych.
  - Każdy z tych zbiorów jest analizowany za pomocą standardowych procedur wykorzystujących informację pełną w taki sposób, jakby wartości imputowane były prawdziwe.

Dla każdego wykorzystywanego modelu imputowane są **co najmniej dwie** wartości, co ma odzwierciedlać **niepewność** co do wartości imputowanych braków.

Parowanie statystyczne

Metody parametryczne:

- **Wielokrotna imputacja**

Estymatorem punktowym wielokrotnej imputacji jest średnia arytmetyczna z m podstawień:

$$\hat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}^{(t)}$$

Wariancja międzygrupowa wyraża się wzorem: $B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2$

wariancję wewnątrzgrupową można zapisać jako wyrażenie:

$$W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}\left(\hat{\theta}^{(t)}\right)$$

Wariancja ogólna jest sumą wariancji wewnątrz- i międzygrupowej zmodyfikowanym o składnik $\frac{m+1}{m}$:

$$T = W + \frac{m+1}{m}B$$

Estymacji przedziałowej w wielokrotnej imputacji dokonuje się szacując przedział ufności:

$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$,

gdzie **liczba stopni swobody** $v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2$

Parowanie statystyczne

Metody nieparametryczne:

- **Losowa**

Polega na losowym doborze zmiennej $\boldsymbol{Z}$ ze zbioru dawcy do zbioru biorcy. By zachować jak największą zgodność dołączanych wartości, zbiory A i B dzielone są na jak największą liczbę homogenicznych grup (na podstawie wartości wybranych zmiennych, najlepiej jakościowych, $\boldsymbol{X}$) - $\boldsymbol{X_G}$. Grup takich powinno być możliwie dużo. Losowe dołączanie przebiega wtedy w obrębie wyznaczonych grup.

- **Najbliższego sąsiada**

Polega na wybraniu dla każdego rekordu ze zbioru A najbardziej podobnego rekordu ze zbioru B. „Podobieństwo” to mierzone jest odległością między wartościami zmiennych parujących wybranych z wektora zmiennych wspólnych $\boldsymbol{X}$ $\boldsymbol{(X_M \subseteq X)}$:

$d_{ab} = (x_{M.a}, x_{M.b}) = min,\ b = 1,2, \dots, n_b.$

Wartość $\boldsymbol{Z}$ jest następnie imputowana w A.

Parowanie statystyczne

Metody mieszane:

1. Konstrukcję modelu regresji na podstawie informacji ze zbioru dawcy (na potrzeby rozważań można przyjąć B) $Z = g(x; \theta)$. Estymacji parametrów θ. Na podstawie oszacowanego modelu obliczane są wartości teoretyczne $\tilde{z}_a$ w zbiorze A.

2. Dla każdego rekordu w zbiorze biorcy wyszukiwany jest „najbliższy sąsiad” w zbiorze dawcy na podstawie odległości między wartościami teoretycznymi w A i empirycznymi w B: $d_{ab}(\tilde{z}_a, z_b) = min$.

Parowanie statystyczne

Ocena jakości zintegrowanych źródeł:
1. Reprodukcja nieznanych wartości Z w pliku biorcy – prawdziwe, nieznane wartości wektora zmiennych Z w pliku biorcy są reprodukowane. Jeżeli w efekcie otrzymujemy prawdziwą wartość, sytuacje taką określa się „trafieniem” (*hit* - dla każdej jednostki zbioru biorcy). Można obliczyć „współczynnik trafień” (*hit ratio*).
2: **Zachowanie łącznego rozkładu** – prawdziwy łączny rozkład zmiennych $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ jest odzwierciedlony w zintegrowanym zbiorze.
3: Struktura korelacji zmiennych jest zachowana w zintegrowanym pliku: $\widetilde{cov}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = cov(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Poprawnie odwzorowane również są rozkłady brzegowe: $\tilde{f}_{\boldsymbol{XY}} = f_{\boldsymbol{X}}$ oraz $\tilde{f}_{\boldsymbol{XZ}} = f_{\boldsymbol{XZ}}$.
4: Po przeprowadzeniu parowania statystycznego **brzegowy i łączny rozkład zmiennych** w pliku dawcy powinien zostać zachowany w zintegrowanym zbiorze. Wtedy należy się spodziewać, że spełnione zostaną równości $\tilde{f}_Z = f_Z$ oraz $\tilde{f}_{XZ} = f_{XZ}$ jeżeli $\boldsymbol{Z}$ jest imputowane do próby $(\boldsymbol{X}, \boldsymbol{Y})$.

Parowanie statystyczne

Ocena jakości zintegrowanych źródeł:

Niemieckie Stowarzyszenie Analiz Medialnych wystosowało postulaty dotyczące kontroli jakości zintegrowanych repozytoriów:

- najpierw porównywane są empiryczne rozkłady zmiennych wspólnych X w pliku dawcy i biorcy w celu oceny zgodności,
- następnie porównywany jest rozkład empiryczny dołączonych zmiennych Z w pliku biorcy i dawcy,
- w ostatnim etapie porównuje się łączny rozkład $f_{X,Z}$ obserwowany w pliku dawcy z rozkładem łącznym $\tilde{f}_{X,Z}$ obserwowanym w pliku zintegrowanym.

*„W sytuacji, gdy badacze społeczni tak chciwie pragną bogatych w informacje zbiorów danych, parowanie statystyczne może wydawać się ogromnie atrakcyjną procedurą tworzenia zbiorów zawierających logiczne powiązania zmiennych znajdujących się w oddzielnych źródłach […]. Chciałbym najpierw zobaczyć rzetelną ocenę takich łącznych rozkładów zanim zdjąłbym z procedury tabliczkę: „**UWAGA! NIEBEZPIECZEŃSTWO! STOSOWAĆ Z ZACHOWANIEM OSTROŻNOŚCI!**”.*

Ivan Fellegi

Badanie empiryczne

Badanie empiryczne

Charakterystyka	Badanie Budżetów Gospodarstw Domowych	Badanie Dochodów i Warunków Życia
Czas realizacji	cały rok 2005	2 maja – 19 czerwca 2006
Zbiorowość badania	gospodarstwa domowe w Polsce	gospodarstwa domowe w Polsce
Metoda doboru próby	reprezentacyjna	reprezentacyjna
Schemat losowania	dwustopniowy, warstwowy	dwustopniowy, warstwowy
Przedmiot badania	— budżet gospodarstwa domowego (zestawienie różnych źródeł dochodów i wydatków) — wyposażenie gospodarstwa domowego — wielkość spożycia produktów i usług	— sytuacja dochodowa — wyposażenie gospodarstwa domowego — ubóstwo — różne aspekty warunków życia ludności
Zakładana liczebność populacji generalnej (suma wag analitycznych)	13 332 605	13 300 839
Wielkość próby (roczna)	34 767	14 914 (zakładana próba 18 494)

Badanie empiryczne

Na potrzeby badania empirycznego **postanowiono dołączyć**:

- do zbioru EU-SILC – wydatki ogółem gospodarstwa domowego,
- do zbioru BBGD – dochody głowy gospodarstwa domowego **.**

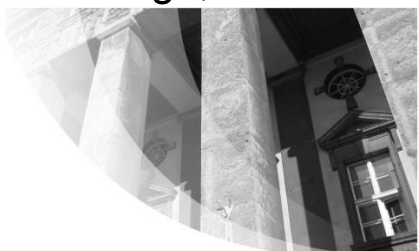
Na podstawie informacji ze zmiennych:

wydatki gospodarstw domowych:

- czy jest łazienka,
- czy jest ustęp spłukiwany,
- czy gospodarstwo posiada samochód,
- liczba pokoi,
- rodzaj budynku,
- ekwiwalentny dochód do dyspozycji ,
- dochód do dyspozycji ,
- wielkość gospodarstwa domowego;

dochody głów gospodarstw domowych:

- czy jest ustęp spłukiwany,
- czy gospodarstwo posiada pralkę,
- czy gospodarstwo posiada samochód,
- czy gospodarstwo posiada tv,
- liczba pokoi,
- rodzaj budynku,
- tytuł prawny do zajmowanego mieszkania,
- ekwiwalentny dochód do dyspozycji,
- dochód do dyspozycji,
- wielkość gospodarstwa domowego;



Badanie empiryczne

Dla celów integracji wykorzystano następujące metody:

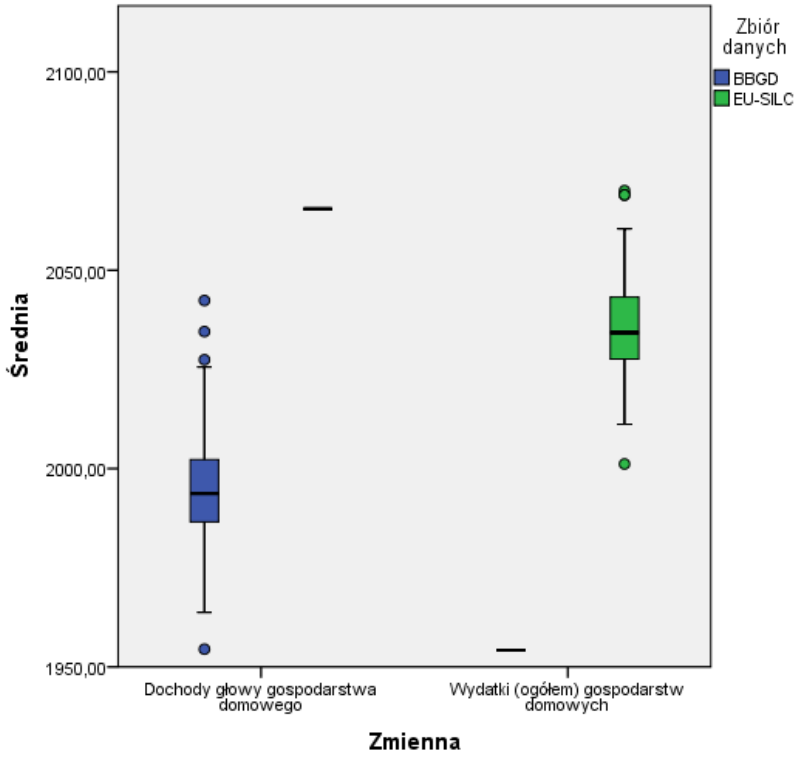
1. Nieparametryczne
   a) Losowa
   b) Najbliższego sąsiada

2. Parametryczna
   a) Wielokrotna imputacja

3. Mieszana

Badanie empiryczne

Metoda losowa
Wykonano 100 iteracji



Współczynnik korelacji między zmiennymi dołączanymi a wybranymi zmiennymi parującymi

Zmienna	dochody głowy GD		wydatki GD	
	Średnia	Empiryczna	Średnia	Empiryczna
Ekwiwalentny dochód GD	0,0006	0,8537	-0,0019	0,484
Dochód rozporządzalny GD	0,0005	0,8865	-0,0019	0,588
Liczba osób w GD	0,0005	0,1541	0,0005	0,2685
Ekwiwalentna wielkość GD	0,0005	0,1536	0,0003	0,279

Badanie empiryczne

Metoda najbliższego sąsiada

Zmienna	Statystyka	Zbiór			$\frac{\hat{\theta}_{int}}{\hat{\theta}_{emp}}$
		EU-SILC	BBGD	Zintegrowany	
dochód głowy GD	Średnia	2 065,48	1 868,64	1 967,01	0,952
	Wariancja	3 144 682,13	2 027 856,49	2 595 665,26	0,825
	Odchylenie standardowe	1 773,33	1 424,03	1 611,11	0,909
wydatki GD	Średnia	1 954,20	1 982,21	1 968,20	1,007
	Wariancja	2 271 514,07	2 086 521,57	2 179 261,95	0,959
	Odchylenie standardowe	1 507,15	1 444,48	1 476,23	0,979

Współczynnik korelacji między zmiennymi dołączanymi a wybranymi zmiennymi parującymi

Zmienna dołączana	Zmienna parująca	Zbiór danych			$\frac{\rho_{int}}{\rho_{emp}}$
		BBGD	EU-SILC	Zintegrowany	
wydatki GD	Liczba osób	0,2685	0,2865	0,2773	1,0326
	Ekwiwalentna wielkość	0,279	0,292	0,2853	1,0228
	Dochód rozporządzalny	0,588	0,6693	0,6271	1,0666
	Dochód ekwiwalenty	0,484	0,5812	0,5292	1,0934
dochód głowy GD	Liczba osób	0,1701	0,1541	0,1601	1,0392
	Ekwiwalentna wielkość	0,1738	0,1536	0,1618	1,053
	Dochód rozporządzalny	0,7609	0,8865	0,8236	0,929
	Dochód ekwiwalenty	0,706	0,8537	0,7756	0,9084

Badanie empiryczne

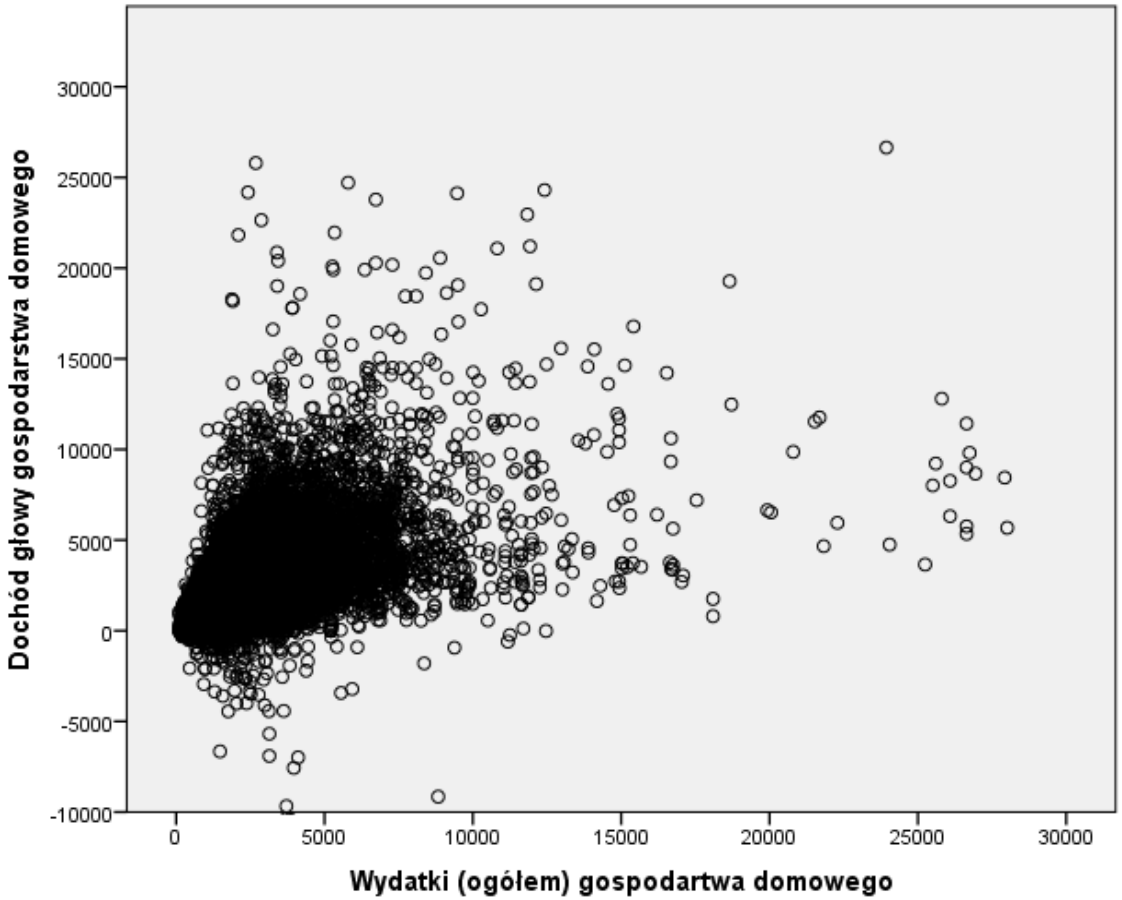
Metoda wielokrotnej imputacji i mieszana

Zmienna dołączana	Statystyka	Imputacja stochastyczna	Model mieszany
Wydatki (ogółem) gospodarstwa domowego	B	191,76	936,66
	W	0,0003	0,0004
	T	193,68	946,03
	$\sqrt{T}$	13,92	30,76
	v	112,32	187,22
	$t_{v,\frac{\alpha}{2}}$	2,271935	2,2595942
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 426,69	1 516,20
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 489,93	1 655,20
	szerokość przedziału	63,24	139
Dochody głowy gospodarstwa domowego	B	29,85	16,45
	W	0,0006	0,0005
	T	30,15	16,62
	$\sqrt{T}$	5,49	4,08
	v	102,46	100,53
	$t_{v,\frac{\alpha}{2}}$	2,2749712	2,2756524
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 890,58	1 696,98
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 915,57	1 715,54
	szerokość przedziału	24,98	18,55

Zmienna dołączana	Statystyka	Imputacja stochastyczna	Model mieszany
Korelacja $z(\hat{\rho}^{(t)})$	B	0,00003	0,00013
	W	2,00E-15	2,00E-15
	T	0,00003	0,00013
	$\sqrt{T}$	0,01	0,01
	v	99	99
	$t_{v,\frac{\alpha}{2}}$	2,2760035	2,2760035
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	0,5661	0,5463
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	0,5824	0,5817
	szerokość przedziału	0,0164	0,0354

Badanie empiryczne

Łączna obserwacja cech nieobserwowanych łącznie



Badanie empiryczne

Ocena precyzji estymatorów w zintegrowanym zbiorze

Zmienna	Region	BBGD			EU-SILC			zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Wydatki (ogółem) gospodarstwa domowego	centralny	1916	16,40	0,86	2130	29,48	1,38	2024	14,95	0,74	0,088
	południowy	2008	16,44	0,82	2072	27,59	1,33	2040	14,45	0,71	0,121
	wschodni	1970	19,27	0,98	2040	29,12	1,43	2005	16,19	0,81	0,160
	północno-zachodni	1973	18,20	0,92	2147	37,72	1,76	2061	18,11	0,88	0,005
	południowo-zachodni	2138	27,10	1,27	2121	44,00	2,07	2130	23,27	1,09	0,141
	północny	2058	21,28	1,03	1994	27,64	1,39	2026	16,51	0,82	0,224
Dochód głowy gospodarstwa domowego	centralny	1831	18,40	1,00	2270	34,05	1,50	2052	17,16	0,84	0,168
	południowy	2005	21,04	1,05	2097	42,49	2,03	2051	20,63	1,01	0,041
	wschodni	1928	18,61	0,97	2017	26,84	1,33	1972	15,27	0,77	0,198
	północno-zachodni	2000	20,86	1,04	2092	38,55	1,84	2046	19,31	0,94	0,095
	południowo-zachodni	2111	34,24	1,62	2134	37,40	1,75	2122	24,96	1,18	0,275
	północny	2108	25,16	1,19	1990	30,51	1,53	2049	18,99	0,93	0,223

Uwaga, kolory:

szary – szacunki na podstawie wartości dołączanych,

zielony – zysk na jakości szacunków w zbiorze zintegrowanym względem danego zbioru wejściowego.

Miara zysku definiowana jest jako $1-\frac{s_{int}(\bar{x})}{s_{BBGD}(\bar{x})}$ lub $1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$.

Źródło: opracowanie własne

Badanie empiryczne

Ocena precyzji estymatorów w zintegrowanym zbiorze

Zmienna	Województwo	BBGD			EU-SILC			zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Wydatki (ogółem) gospodarstwa domowego	dolnośląskie	2112	28,65	1,36	2123	50,67	2,39	2117	25,61	1,21	0,106
	kujawsko-pomorskie	2100	36,62	1,74	1946	44,69	2,30	2022	27,75	1,37	0,242
	lubelskie	1952	35,47	1,82	2097	56,94	2,72	2024	30,88	1,53	0,130
	lubuskie	1999	42,23	2,11	2113	78,22	3,70	2056	39,16	1,90	0,073
	łódzkie	1925	26,84	1,39	2058	44,50	2,16	1992	23,36	1,17	0,130
	małopolskie	1982	25,96	1,31	2186	48,23	2,21	2085	24,22	1,16	0,067
	mazowieckie	1912	20,68	1,08	2169	38,52	1,78	2042	19,27	0,94	0,068
	opolskie	2218	65,54	2,95	2116	87,94	4,16	2166	51,70	2,39	0,211
	podkarpackie	2006	34,96	1,74	1977	40,41	2,04	1991	26,04	1,31	0,255
	podlaskie	1994	47,54	2,38	2057	54,86	2,67	2024	35,41	1,75	0,255
	pomorskie	2026	34,40	1,70	2055	47,17	2,29	2040	27,25	1,34	0,208
	śląskie	2024	21,21	1,05	2002	33,16	1,66	2013	17,96	0,89	0,153
	świętokrzyskie	1924	37,10	1,93	2019	81,87	4,05	1971	38,97	1,98	-0,050
	warmińsko-mazurskie	2046	39,76	1,94	1978	52,57	2,66	2011	31,18	1,55	0,216
	wielkopolskie	1959	24,51	1,25	2107	45,23	2,15	2034	22,72	1,12	0,073
	zachodniopomorskie	1983	35,24	1,78	2241	88,03	3,93	2111	39,72	1,88	-0,127

Badanie empiryczne

Ocena precyzji estymatorów w zintegrowanym zbiorze

Zmienna	Województwo	BBGD			EU-SILC			zintegrowany			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Dochód głowy gospodarstwa domowego	dolnośląskie	2086	31,35	1,50	2090	41,00	1,96	2088	24,38	1,17	0,405
	kujawsko-pomorskie	2181	41,34	1,90	1851	39,79	2,15	2014	29,02	1,44	0,271
	lubelskie	1930	32,61	1,69	2093	51,76	2,47	2011	28,23	1,40	0,455
	lubuskie	1957	47,73	2,44	2168	99,12	4,57	2063	47,55	2,31	0,520
	łódzkie	1842	31,69	1,72	2109	47,85	2,27	1977	26,36	1,33	0,449
	małopolskie	1967	35,27	1,79	2252	55,58	2,47	2110	30,05	1,42	0,459
	mazowieckie	1825	22,57	1,24	2358	45,57	1,93	2094	22,27	1,06	0,511
	opolskie	2186	98,49	4,51	2252	81,96	3,64	2220	65,72	2,96	0,198
	podkarpackie	1942	34,01	1,75	1927	43,65	2,27	1935	26,46	1,37	0,394
	podlaskie	1957	45,84	2,34	2123	61,91	2,92	2038	36,41	1,79	0,412
	pomorskie	2041	37,23	1,82	2169	61,63	2,84	2104	32,22	1,53	0,477
	śląskie	2028	26,16	1,29	2002	59,38	2,97	2015	27,70	1,37	0,533
	świętokrzyskie	1875	39,95	2,13	1913	58,43	3,05	1894	32,92	1,74	0,437
	warmińsko-mazurskie	2104	56,08	2,67	1943	54,41	2,80	2020	38,95	1,93	0,284
	wielkopolskie	1981	27,52	1,39	1962	45,11	2,30	1971	23,83	1,21	0,472
	zachodniopomorskie	2060	41,97	2,04	2281	84,39	3,70	2170	40,85	1,88	0,516

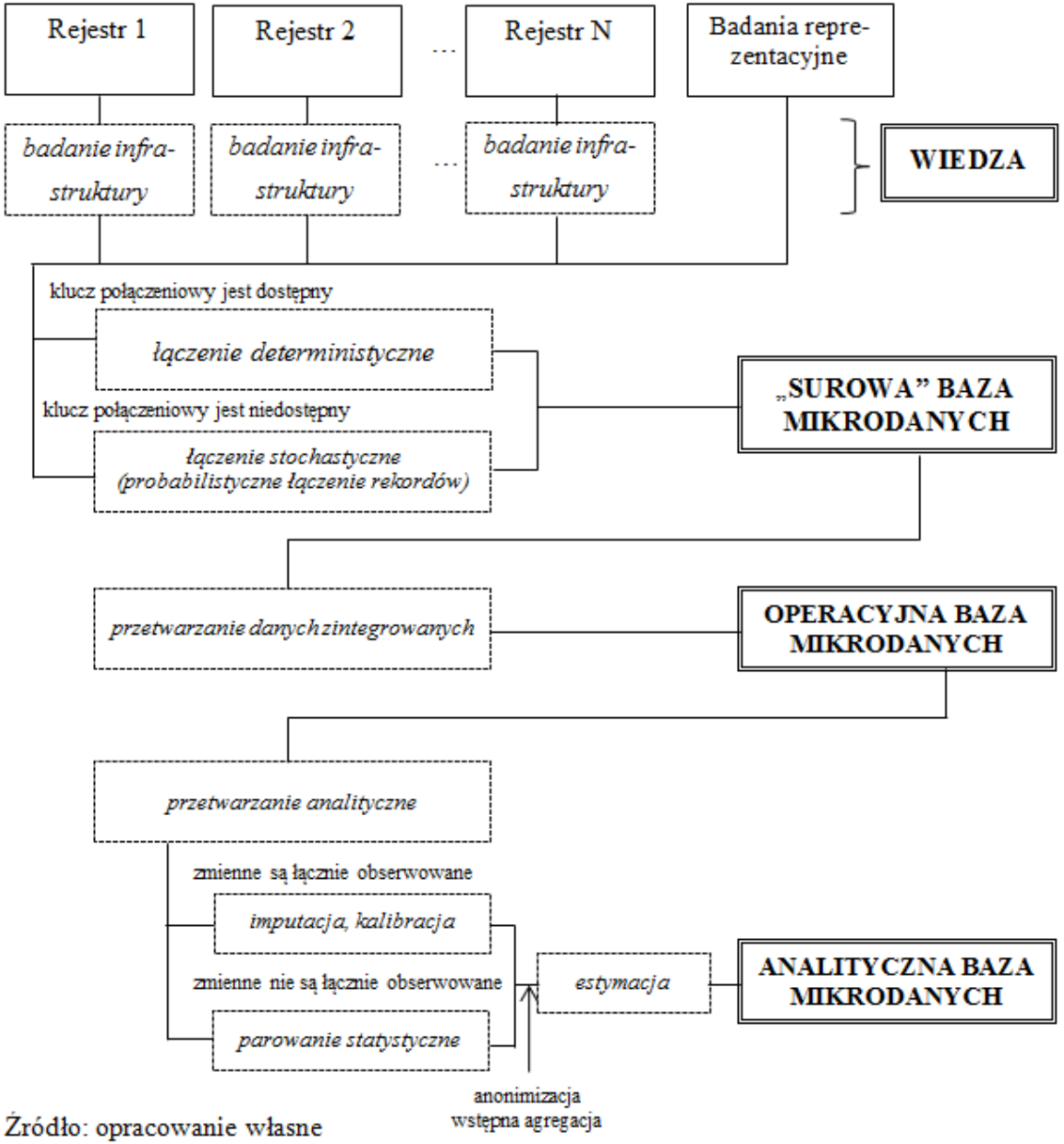
Podsumowanie

Podsumowanie

- Integracja danych jest wykorzystywana w statystyce na coraz większą skalę.
- Wykorzystanie istniejących źródeł danych umożliwia obniżenie kosztów badań i obciążenie respondentów.
- Wielość źródeł danych, przy rosnącym zapotrzebowaniu na rzetelną, aktualną, precyzyjną, wielowymiarową informację, będzie powodowała dalszy rozwój metod integracji danych.

- Statystyczna integracja danych umożliwia łączną obserwację cech nieobserwowanych łącznie w przypadku niedostępności klucza.
- Konkatenacja baz danych pochodzących z badań reprezentacyjnych umożliwia poprawę jakości oszacowań.

Podsumowanie



Źródło: opracowanie własne

Ważna zagadnienia pominięte w wykładzie

- Ocena jakości istniejących źródeł
- Harmonizacja źródeł danych przed integracją
  - Populacje
  - Definicje zmiennych
- Wybór zmiennych parujących
- Analiza porównawcza różnych metod integracji

Wybrane pozycje literaturowe

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

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

Data Integration Manual, (2006), praca zbiorowa Statistics New Zealand , Wellington

D'Orazio M., Di Zio M., Scanu M., (2006), *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England

Dygaszewicz J., (2010), *Integracja rejestrów publicznych*, Główny Urząd Statystyczny, Warszawa

Gill L., (2001), *Methods for Automatic Record Matching and Linkage and their use in National Statistics*, National Statistics Methodological Series No 25, National Statistics, United Kingdom.

Hardling A., Kelly S., Percival R., Keegan M., (2009), *Population Ageing and Government Age Pension Outlays*, ESRI International Collaboration Project, NATSEM, University of Canberras

Linder F., (2004), *The use of administrative registers and sample surveys in the Dutch Census* of 2001 [w:] *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen

Nordholdt E.S., (2004), *Introduction to the Dutch Virtual Census of 2001* [w:] *The Dutch Virtual Census of 2001. Analysis and Methodology*, Statistics Netherlands, Voorburg/Heerlen

Penneck S., (2007), *Using administrative data for statistical purposes*, Economic & Labour Market Review

Raessler S., (2002), *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA

Swiss Federal Statistical Office, (2008), *The Swiss Census 2010: Moving towards a comprehensive system of household and person statistics*, Federal Statistical Office.

van der Laan P., (2000), *Integrating administrative registers and household surveys*, „Netherlands Official Statistics", vol. 15, Summer 2000, Special issue*: Integrating administrative registers and household surveys*, Statistics Netherlands, Voorburg/Heerlen.

Wallgren A., Wallgren B., (2007), *Register-based Statistics. Administrative Data for Statistical Purposes*, John Wiley and Sons Ltd.

Zhang L-C., (2012), *Micro calibration for data integration*, referat wygłoszony na Kongresie Statystyki Polskiej, Poznań