Urząd Statystyczny w Kielcach

Warszawa, 2-4 lipca 2024 r.
IV Kongres Statystyki Polskiej

Determinanty odmowy odpowiedzi w badaniu CBSG/01: próba identyfikacji

Urząd Statystyczny w Kielcach
Mariola Chrzanowska
Mateusz Kolmaga

Plan prezentacji

- Wprowadzenie do badania CBSG/01
- Problem badawczy
- Wstępna analizy danych za lata 2017-2022: analiza częstości, tabele krzyżowe, mapy ciepła, test niezależności chi-kwadrat
- Drzewo klasyfikacyjne CHAID
- Analiza log-liniowa
- Wnioski

Badanie CBSG/01

„Badanie podmiotów o liczbie pracujących do 49 osób”

- Warunki prowadzenia działalności gospodarczej i opinia respondentów nt. otoczenia prowadzonej działalności gospodarczej
- Próba zawiera jednostki o liczbie pracujących do 49 osób, w tym tzw. mikroprzedsiębiorstwa o liczbie pracujących do 9 osób. Wielkość próby w określonej edycji zawierała się między ok. 28 tys. a 39 tys. podmiotów
- Ankieta zawiera 18 pytań z czego 10 pytań to pytania otwarte
- Forma pozyskiwania odpowiedzi: internetowa przez Portal Sprawozdawczy, telefoniczna, listowna

Problem badawczy

„Co badacz może powiedzieć o odmowie w związku z geograficzną i branżową charakterystyką respondenta?”

ANKIETY NIEZREALIZOWANE

Zawieszenie (RA = 13)	Likwidacja (RA = 15)	Odmowa (RA = 22)

- Odmowa złożenia sprawozdania: status RA = 22
- Siedziby podmiotów: województwa i makroregiony
- Branże, w których działają podmioty: sekcje i grupy sekcji

Przygotowanie do badania

- Na potrzeby badania pozostałe zmienne uwzględnione w badaniu zakodowano w następujący sposób:

Sposób kodowania	Makroregion (MREG_NUM)	Grupa sekcji (GS_NUM)
1	WOJ. MAZOWIECKIEGO	PRZEMYSŁ
2	CENTRALNY	BUDOWNICTWO
3	PÓŁNOCNY	HANDEL
4	POŁUDNIOWO-ZACHODNI	TRANSPORT
5	WSCHODNI	OBSŁUGA NIERUCHOMOŚCI I FIRM
6	PÓŁNOCNO-ZACHODNI	POZOSTAŁE SEKCJE PKD
7	POŁUDNIOWY	

Badanie CBSG/01

„Badanie podmiotów o liczbie pracujących do 49 osób”

Odsetki statusów sprawozdań w badaniu CBSG/01 (2017-2022)				
Edycja badania	RA = 01	**RA = 22**	Pozostałe RA	Ogółem
2017 r.	46,5	**49,2**	4,3	100
2018 r.	50,5	**44,6**	4,8	100
2019 r.	44,6	**53,4**	2	100
2020 r.	46,2	**48,9**	4,9	100
2021 r.	46,1	**45,4**	8,5	100
2022 r.	48,3	**47,8**	3,9	100

Jednostki o statusie „01” i „22” (odmowa złożenia sprawozdania) stanowią przeważającą część podmiotów we wszystkich omawianych edycjach.

Tabele krzyżowe 1

- Odsetki odmów w poszczególnych latach (edycjach badania CBSG/01)
- Ilorazy liczebności obserwowanych do liczebności oczekiwanych
- Punktem odniesienia jest zbiór składający się z podmiotów, których sprawozdania zostały zatwierdzone i tych, które odmówiły dopełnienia obowiązku sprawozdawczego
- Istotność dwustronna testu niezależności chi-kwadrat w każdym przypadku wynosi < 0,05 (0,000)
- Wyższe odsetki odmów i ilorazy większe od 1 (więcej odmów danej kategorii niż sugerowałaby to liczebność oczekiwana) oznaczono odcieniami czerwieni, niższe odsetki i ilorazy mniejsze niż 1 - zieleni

Tabele krzyżowe 2: pola działalności

Sekcja PKD	2017	2018	2019	2020	2021	2022
A	48,2	30,8	34,6	42,1	37,1	50,9
B	43,6	52,6	29,4	30,8	56,5	43,3
C	51,4	49,6	54,6	49,8	49,8	49,4
D	45,5	47,4	75,0	68,8	84,6	53,6
E	61,5	45,5	54,9	67,8	52,9	61,8
F	54,1	48,0	58,2	56,1	50,6	50,7
G	50,0	45,9	52,3	48,1	47,3	51,5
H	48,4	45,8	51,3	48,7	48,9	46,9
I	47,7	45,5	46,1	46,3	48,3	46,7
J	48,9	46,5	56,1	49,4	48,5	48,6
L	49,2	49,9	59,0	49,9	53,9	49,1
M	60,8	49,9	58,2	58,2	54,3	51,9
N	(-)	(-)	62,9	62,4	53,1	48,5
P	49,9	44,6	55,6	46,3	54,2	46,7
Q	46,8	41,8	45,3	42,9	46,1	45,5
R	44,0	49,7	56,1	46,4	42,8	44,6
S	49,3	47,1	55,8	46,3	45,9	47,1

Odsetek odmów (RA=22) w sekcjach PKD w edycjach z lat 2017-2022

Sekcja PKD	2017	2018	2019	2020	2021	2022
A	0,94	0,66	0,63	0,82	0,75	1,02
B	0,85	1,12	0,54	0,60	1,14	0,87
C	1,00	1,06	1,00	0,97	1,00	0,99
D	0,89	1,01	1,38	1,34	1,70	1,08
E	1,20	0,97	1,01	1,32	1,07	1,24
F	1,06	1,02	1,07	1,09	1,02	1,02
G	0,98	0,98	0,96	0,93	0,95	1,04
H	0,94	0,98	0,94	0,95	0,99	0,94
I	0,93	0,97	0,85	0,90	0,97	0,94
J	0,95	0,99	1,03	0,96	0,98	0,98
L	0,96	1,06	1,08	0,97	1,09	0,99
M	1,19	1,06	1,07	1,13	1,09	1,04
N	(-)	(-)	1,15	1,21	1,07	0,98
P	0,97	0,95	1,02	0,90	1,09	0,94
Q	0,91	0,89	0,83	0,83	0,93	0,92
R	0,86	1,06	1,03	0,90	0,86	0,90
S	0,96	1,00	1,02	0,90	0,93	0,95

Proporcja liczebności odmów: wartości obserwowane do oczekiwanych

Tabele krzyżowe 3: województwa

Województwo	2017	2018	2019	2020	2021	2022
02 Dolnośląskie	50,2	42,9	52,6	49,5	42,0	49,2
04 Małopolskie	41,5	47,5	47,6	50,7	49,6	47,8
06 Lubelskie	55,8	51,2	63,4	53,0	37,5	43,8
08 Lubuskie	57,2	50,9	58,0	54,7	56,4	55,2
10 Łódzkie	54,8	53,2	59,8	48,9	52,1	52,3
12 Małopolskie	53,4	48,0	56,0	49,7	45,7	48,9
14 Mazowieckie	51,7	50,0	56,1	53,7	54,1	52,6
16 Opolskie	53,3	40,5	45,7	50,6	53,3	39,5
18 Podkarpackie	50,3	46,4	49,1	51,5	51,5	48,2
20 Podlaskie	50,4	47,0	46,0	45,8	45,2	48,3
22 Pomorskie	47,8	46,1	55,3	51,8	50,5	53,6
24 Śląskie	52,6	48,0	52,9	50,6	52,1	45,2
26 Świętokrzyskie	48,7	45,3	38,7	42,5	44,6	47,6
28 Warm.-mazurskie	51,6	43,1	46,4	48,2	45,6	46,9
30 Wielkopolskie	53,2	49,5	57,5	56,8	51,9	51,4
32 Zachodniopomorskie	45,3	27,8	52,7	50,2	46,7	51,0

Odsetek odmów (RA=22)
w województwach w edycjach
z lat 2017-2022

	2017	2018	2019	2020	2021	2022
02	0,98	0,91	0,96	0,96	0,85	0,99
04	0,81	1,01	0,87	0,99	1,00	0,96
06	1,09	1,09	1,16	1,03	0,76	0,88
08	1,12	1,08	1,06	1,06	1,14	1,11
10	1,07	1,13	1,10	0,95	1,05	1,05
12	1,04	1,02	1,03	0,97	0,92	0,98
14	1,01	1,06	1,03	1,04	1,09	1,06
16	1,04	0,86	0,84	0,98	1,07	0,79
18	0,98	0,99	0,90	1,00	1,04	0,97
20	0,98	1,00	0,84	0,89	0,91	0,97
22	0,93	0,98	1,01	1,01	1,02	1,08
24	1,03	1,02	0,97	0,98	1,05	0,91
26	0,95	0,97	0,71	0,83	0,90	0,96
28	1,01	0,92	0,85	0,94	0,92	0,94
30	1,04	1,05	1,05	1,10	1,05	1,03
32	0,88	0,59	0,97	0,98	0,94	1,03

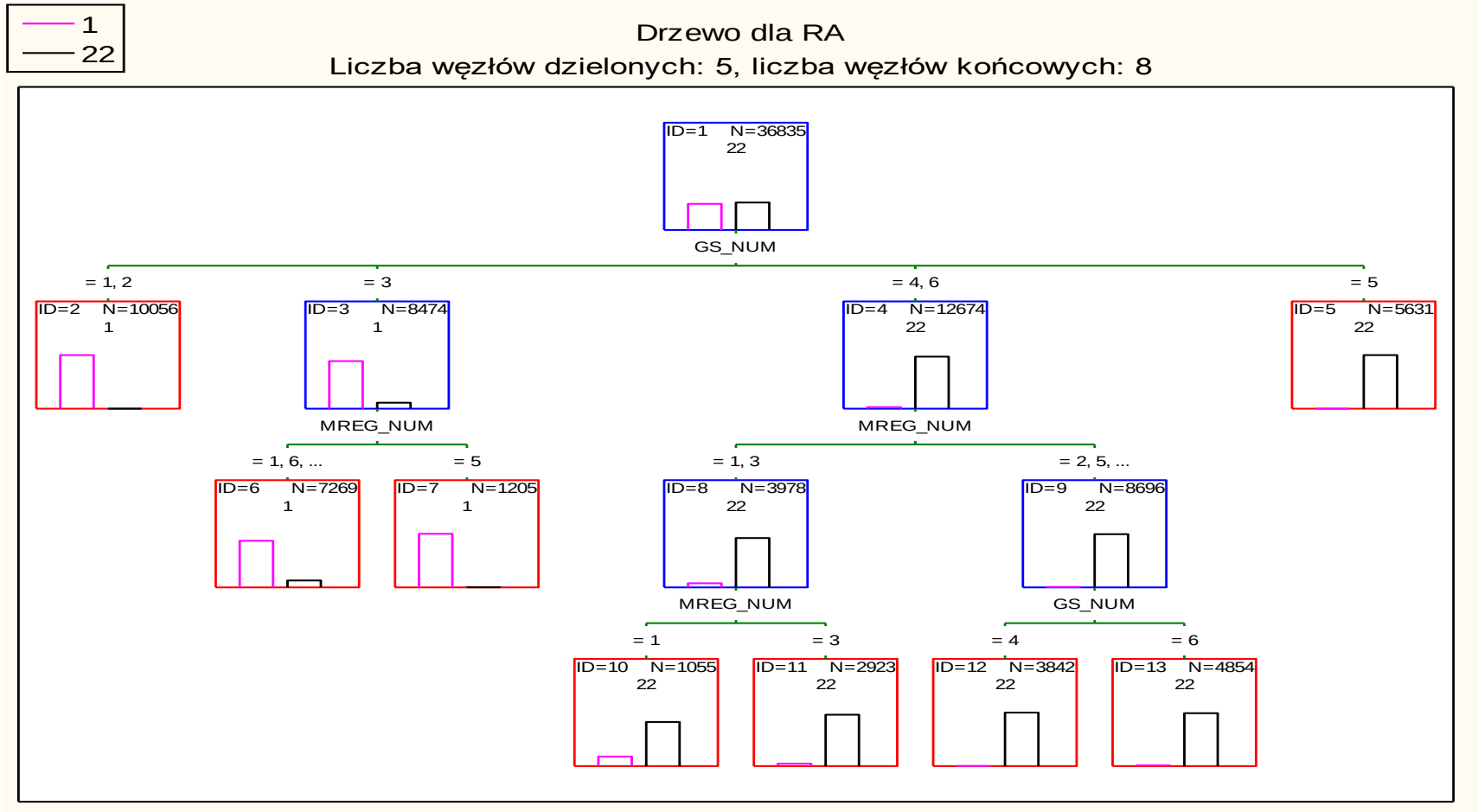
Proporcja liczebności odmów:
wartości obserwowane do oczekiwanych

Drzewo klasyfikacyjne CHAID 1

- Jedna z głównych metod stosowanych w „data mining”
- Cel stosowania: w oparciu o zmienne objaśniające (predyktory) określić przynależność przypadków do klas jakościowej zmiennej zależnej
- Na tle innych metod analitycznych, wymagających często trudnych do spełnienia założeń, drzewa klasyfikacyjne wyróżniają się małą restrykcyjnością
- CHAID jest akronimem od „Chi-square Automatic Interaction Detection”

Drzewo klasyfikacyjne CHAID 2

- Interpretacja wyników (dla grup sekcji i makroregionów)



Drzewo CHAID 3: podsumowanie

- Można zauważyć, że jako odmawiający wypełnienia ankiety zostali (za pomocą algorytmu) zakwalifikowani przedsiębiorcy którzy:
  - zamieszkują makroregion: województwa mazowieckiego, centralny oraz wschodni
  - posiadają firmę z grup sekcji: Transport, Obsługa nieruchomości i firm oraz Pozostałe sekcje PKD

Analiza log-liniowa 1

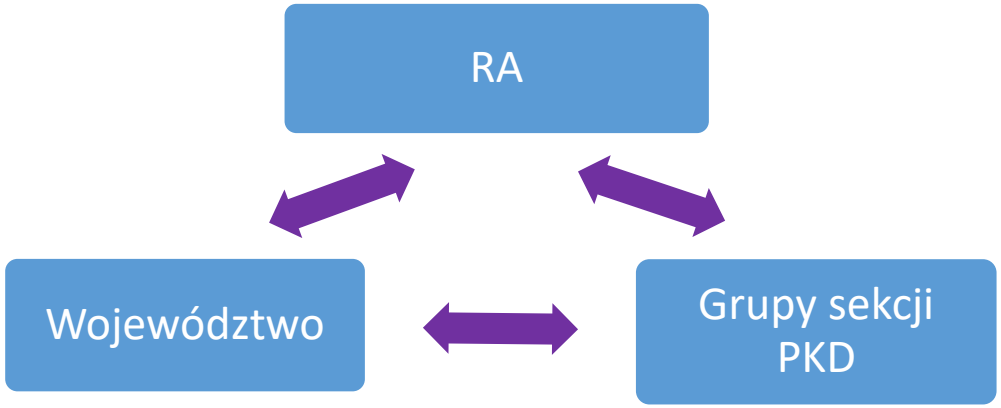
- Analiza logarytmiczno-liniowa należy do zaawansowanych metod badawczych dotyczących związków (interakcji) ukrytych w tabelach kontyngencji
- Umożliwia badanie nieznanych interakcji w danych, do których stosuje się najsłabszy (nominalny) poziom pomiaru
- Pozwala zweryfikować statystyczną istotność wykrytych relacji
- Jakość modeli powstałych w wyniku analizy log-liniowej można zweryfikować za pomocą typowych miar stosowanych w KMNK (kryterium AIC i BIC)
- Pozwala skonstruować modele ilościowo opisujące interakcje pomiędzy badanymi czynnikami i ich kategoriami

Analiza log-liniowa 2

- Poprawnie zbudowany model log-liniowy umożliwia najlepszą predykcję liczebności, przy uwzględnieniu w modelu jak najmniejszej liczby interakcji
- Uwzględnienie wszystkich czynników i ich interakcji daje tzw. model nasycony, który choć najlepiej dopasowany, nie zawsze jest wygodny w użyciu (wpływ niektórych czynników i interakcji może być niewielki, w porównaniu z pozostałymi składnikami modelu)

Analiza log-liniowa 3

Zastosowanie w wykrywaniu determinant odmów odpowiedzi



- W omawianym przypadku predyktorami (zmiennymi niezależnymi) są województwo i grupy sekcji PKD.

Analiza log-liniowa 4

- W analizie log-linowej wykorzystano model Poissona. Jako zmienne do modelu wybrano jak poprzednio: RA, MREG_NUM, GS_NUM
- Na etapie wyboru modelu okazało się, że najlepszym modelem jest model zawierający stałą oraz interakcje 3 rzędu (RA*MREG_NUM*GS_NUM)
- Z uwagi na wielkość modelu 84 interakcje 3 rzędu (zmienne zawierają odpowiednio: 2, 6 oraz 7 kategorii) w dalszych analizach skoncentrowano się główne na 42 przypadkach dla których zanotowano odmowę wypełnienia ankiety
- Na kolejnym slajdzie zaprezentowano te komponenty dla których NIE zaobserwowano statystycznej interakcji pomiędzy kategoriami (w przypadku odmowy wypełnienia ankiety)
- Dodatkowo na kolejnym slajdzie zaprezentowano kategorie wyodrębnione przez drzewo CHAID

Analiza log-liniowa 5: wybrane wyniki

- Analiza log-liniowa –brak ukrytych interakcji między kategoriami (dla grup sekcji i makroregionów)

Predyktor	Oszacowanie	SE	Z	p
[RA = 22] * [MREG_NUM = 2] * [GS_NUM = 2]	0,168	0,09	1,85	0,06
[RA = 22] * [MREG_NUM = 2] * [GS_NUM = 4]	-0,064	0,10	-0,67	0,50
[RA = 22] * [MREG_NUM = 5] * [GS_NUM = 1]	0,102	0,09	1,10	0,27
[RA = 22] * [MREG_NUM = 6] * [GS_NUM = 2]	-0,036	0,10	-0,38	0,70
[RA = 22] * [MREG_NUM = 6] * [GS_NUM = 4]	0,089	0,09	0,97	0,33

Kategoria wskazana przez algorytm CHAID

Analiza log-liniowa 6

- Wyniki dla grup sekcji 5 - ONiF

Predyktor	Oszacowanie	SE	Z	p
[RA = 22] * [MREG _NUM = 1] * [GS_NUM = 5]	0,298	0,088	3,385	<,001
[RA = 22] * [MREG_NUM = 2] * [GS_NUM = 5]	0,298	0,088	3,385	<,001
[RA = 22] * [MREG_NUM = 3] * [GS_NUM = 5]	0,623	0,083	7,528	<,001
[RA = 22] * [MREG_NUM = 4] * [GS_NUM = 5]	0,771	0,081	9,557	<,001
[RA = 22] * [MREG_NUM = 5] * [GS_NUM = 5]	0,987	0,078	12,625	<,001
[RA = 22] * [MREG_NUM = 6] * [GS_NUM = 5]	0,613	0,083	7,4	<,001
[RA = 22] * [MREG_NUM = 7] * [GS_NUM = 5]	0,934	0,079	11,863	<,001

Analiza log-liniowa 7

- Wyniki dla pozostałych kategorii wyodrębnionych algorytmem CHAID

Predyktor	Oszacowanie	SE	Z	p
[RA = 22] * [MREG_NUM = 2] * [GS_NUM = 6]	0,543	0,084	6,474	<,001
[RA = 22] * [MREG_NUM = 3] * [GS_NUM = 4]	1,409	0,074	18,923	<,001
[RA = 22] * [MREG_NUM = 4] * [GS_NUM = 4]	1,258	0,076	16,626	<,001
[RA = 22] * [MREG_NUM = 4] * [GS_NUM = 6]	1,145	0,077	14,943	<,001
[RA = 22] * [MREG_NUM = 5] * [GS_NUM = 4]	0,49	0,085	5,783	<,001
[RA = 22] * [MREG_NUM = 5] * [GS_NUM = 6]	1,174	0,076	15,379	<,001
[RA = 22] * [MREG_NUM = 6] * [GS_NUM = 6]	1,029	0,078	13,24	<,001

Wnioski 1

- W prezentowanej analizie podjęto próbę wyodrębnienia determinant odmowy wypełnienia ankiety CBSG/01
- Przeprowadzone analizy potwierdziły, że można zauważyć istotne interakcje pomiędzy grupą sekcji a makroregionami

Wnioski 2

- Na podstawie algorytmu CHAiD wyodrębniono makroregiony (województwa mazowieckiego, centralny oraz wschodni), w których przedsiębiorcy skłaniają się do odmowy wzięcia udziału w badaniu
- Ten etap badań wskazał również na grupę sekcji, w których nasila się tendencja odmowy wypełnienia ankiety. Są to: Transport, Obsługa nieruchomości i firm oraz Pozostałe sekcje PKD. W tych kategoriach (według algorytmu) należy wdrożyć działania motywujące przedsiębiorców do udziału w badaniu

Wnioski 3

- Analiza log-liniowa potwierdziła występowanie ogólnej interakcji pomiędzy częstotliwością występowania odmów a makroregionem i grupą sekcji
- Dla większości interakcji w kategorii ODMOWA (RA=22) odnotowano statystyczną istotność
- Nie udało się udowodnić statystycznie występowania ukrytych interakcji pomiędzy następującymi kategoriami:
  - Odmowa*budownictwo*region centralny
  - Odmowa*budownictwo*region południowo-zachodni
  - Odmowa*przemysł*region wschodni
  - Odmowa*transport*region południowo-zachodni

Wnioski 4

- Bazując na wynikach przeprowadzonych analiz szczególnym monitem należałoby objąć podmioty grupy sekcji Transport w regionie Centralnym
- Ta grupa została wyodrębniona w algorytmie ChAID natomiast analiza log-liniowa wykazała brak istotnych interakcji w tym module. Warto zatem zwiększyć monit dla tej kategorii

Bibliografia

- Małgorzata Półtorak, „Modele log-liniowe i ich zastosowania w psychologii”. Przegląd Psychologiczny, 2007, Tom 50, Nr 1, s. 25-44.
- Małgorzata Rószkiewicz, „Próba diagnozy uwarunkowań poziomu wskaźnika braku odpowiedzi w środowisku polskich gospodarstw domowych”. Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 2015, nr 385, s. 228-237.
- Małgorzata Rószkiewicz, „Identyfikacja determinant braku odpowiedzi w badaniu polskich gospodarstw domowych”. Przegląd Statystyczny, R. LXII - Zeszyt 4 – 2015, s. 361-378.
- Mariusz Dacko, Ireneusz Kruszyna, „Zastosowanie modelu drzewa klasyfikacyjnego w ocenie ryzyka kredytowego”. Roczniki naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu• 2023 • Vol. XXV • No. (1).
- Leo Goodman, „Simple models for the analysis of association in cross-classifications having ordered categories”. Journal of the American Statistical Association, 74, s. 537–552.

Urząd Statystyczny
w Kielcach

Dziękujemy za uwagę

M.Kolmaga@stat.gov.pl
M.Chrzanowska@stat.gov.pl