Urząd Statystyczny
w Poznaniu

Imputacja danych w krótkookresowym badaniu przedsiębiorstw

**Paweł Lańduch, Magdalena Homenko,
Katarzyna Buhłak**

Ośrodek Statystyki Krótkookresowej Dział Metodologii i Programowania

4.07.2023 Warszawa

Cechy badań przedsiębiorstw

- Asymetria rozkładu cech,
- Mniejsza liczba zmiennych w stosunku do badań społecznych,
- Proces odpowiedzi – bardziej złożony niż w przypadku badań społecznych,
- Nacisk kładziony jest na dostępność wyników, co może mieć wpływ na ich jakość,
- Złożoność jednostki,
- Obowiązkowy charakter badań,
- Złożone, techniczne definicje,
- Kwestionariusz wypełniany w formie samospisu.

Longitudinalne badania przedsiębiorstw

- Długofalowy charakter badań może zniechęcać do zmian metodologicznych,
- Trzeba uwzględnić dynamiczne zmiany w populacji przedsiębiorstw, takie jak „wejścia” i „wyjścia”,
- Obciążenia przedsiębiorstw,

Rozkład populacji przedsiębiorstw według wielkości opartej na liczbie osób pracujących.

Wyszczególnienie	Ogólna liczba przedsiębiorstw	Rozkład przedsiębiorstw według klasy wielkości				
		mikro	małe	małe	średnie	duże
		< 10	<10;49>	<20;49>	<50;249>	>=250
		%	%	%	%	%
UE-27	23 382 451*	93,5	#	1,9	0,9*	0,2*
Polska	2 066 209	95,1	2,6	1,5	0,7	0,2

* — dane szacunkowe, # – dane poufne, Źródło: (Eurostat, 2020)

Krótkookresowe badanie przedsiębiorstw – DG-1

- celem jest szybkie uzyskanie mierników ekonomicznych i gospodarczych w celu oceny trendów w głównych dziedzinach gospodarki, tzn. przemysłu, budownictwa, handlu i usługach.
- Dotyczy sektora przedsiębiorstw,
- Metodologia badania DG-1 rozróżnia jednostki:
  - duże, dla których liczba pracujących jest równa lub większa niż 50 osób
  - średnie, tzn. takie w których liczba pracujących wynosi od 10 do 49 osób.
  - małe, z liczbą pracujących mniejszą niż 10,
- Do badania włączone są:
  - Wszystkie jednostki duże,
  - 10 % jednostek średnich,
  - Nie ma jednostek małych.

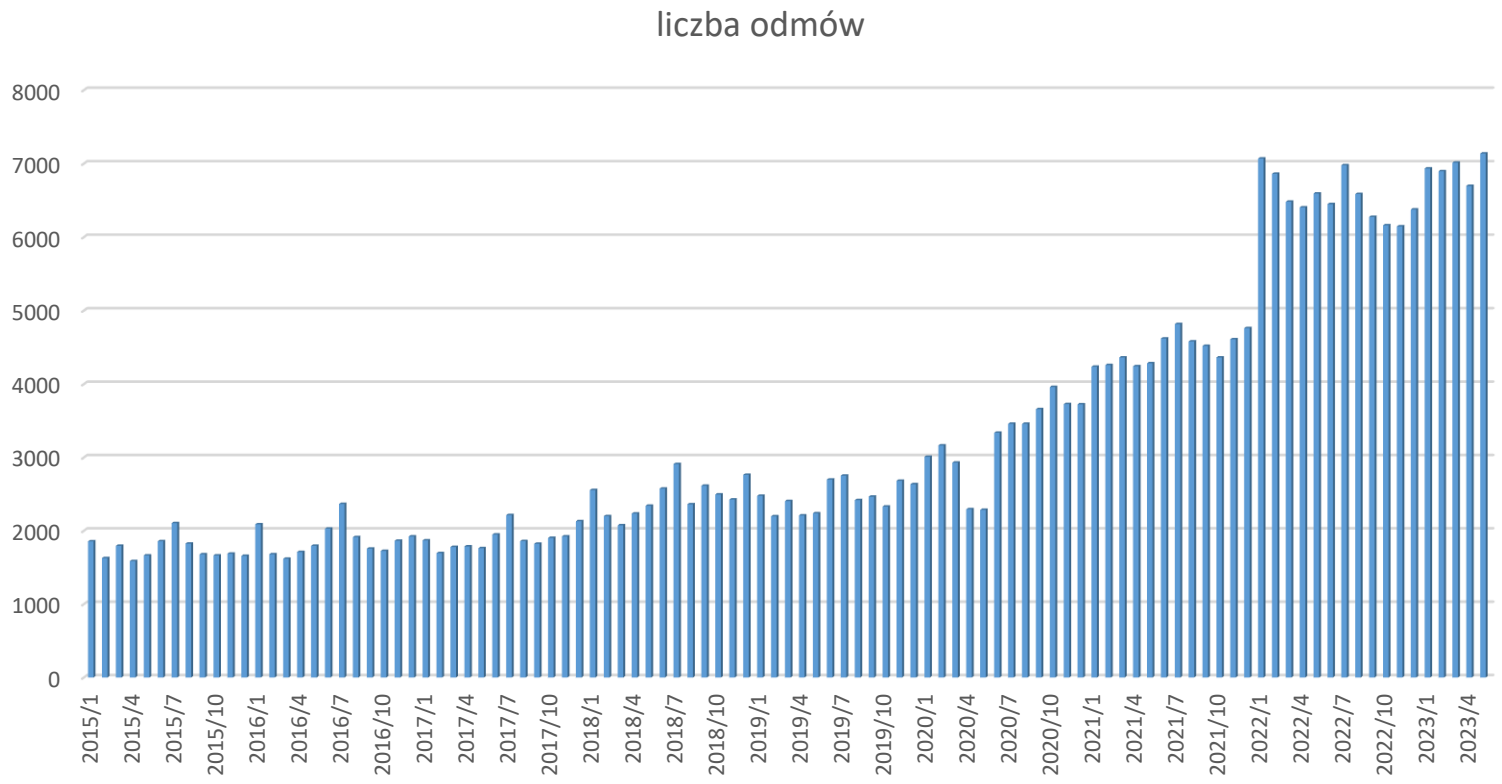
Krótkookresowe badanie przedsiębiorstw – DG-1

Operat badania liczy około 108 000 jednostek z czego:
- około 20 000 to podmioty o liczbie pracujących 50 i więcej osób,
- 88 000 to podmioty o liczbie pracujących 10-49 osób.

Obowiązkiem sprawozdawczym objętych jest około 33 000 jednostek, z czego 13 000 to podmioty o liczbie pracujących 10-49 osób.

Formularz składa się z trzech części:
Dział 1 – Podstawowe dane o działalności gospodarczej, tzn. dane o przychodach ze sprzedaży, liczbie osób pracujących, zatrudnienia, wynagrodzeń i czasu pracy,
Dział 2 - Dane uzupełniające dla jednostek przemysłowych, dane o przychodach ze sprzedaży na eksport i wywóz oraz o nowych zamówieniach,
Dział 3 - Dane uzupełniające dla jednostek świadczących usługi transportowe

Problem braków danych w badaniu



Problem braków danych w badaniu

Braki danych dotyczą jednostek

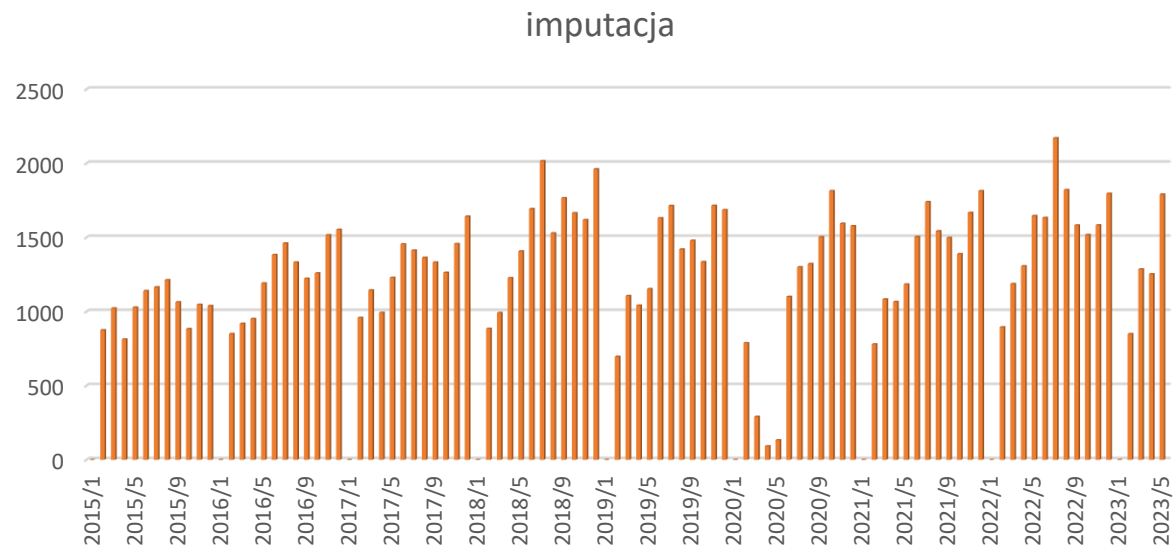
Braki pozycyjne występują sporadycznie

Dwie metody postępowanie:

- Imputacja,
- Uwzględnienie przy uogólnianiu wyników

Metoda imputacji – obecna

- Przeniesienie danych z poprzedniego okresu,



Metoda uzupełniająca – model ETS

Pegels, C.C., 1969. Exponential Forecasting : Some New Variations. Management Science. 15, 311–315.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with Exponential Smoothing. Springer Berlin Heidelberg.

ETS (ERROR-TREND-SEASONALITY)

1. Błąd – A – addytywny, M – multiplikatywny,
2. Trend – N – brak, A – addytywny, Ad – addytywny tłumiony, M – multiplikatywny, Md – multiplikatywny tłumiony
3. Sezonowość – N – brak, A – addytywny M – multiplikatywny

Metoda – przyjęte parametry

- Szereg czasowy zawierający dane dla 24 okresów.
- Braki w szeregu czasowym nie mogą przekraczać 3 okresów, tzn. uzupełnieniami z poprzednich okresów.
- Prognoza wykonana za pomocą pakietu forecast środowiska R.
- Na podstawie błędu RMSE z modelu wyliczono wskaźnik prognozy, tzn.:

$\left(\frac{RMSE}{x}\right) * 100$, gdzie x jest prognozowaną wartością z modelu.

Przyjęto, że jeżeli wskaźnik:

Jest <= 15 % - prognoza wiarygodna, tzw. „flaga zielona”

<15 % i <= 30 % – prognoza budząca wątpliwość, „flaga żółta”

> 30 % - prognoza złej jakości, „flaga czerwona”

Zakres wskaźnika od 0 do 100 %.

Metoda – zmienne

- Zmienne:

- Sw_1b - Przychody netto ze sprzedaży produktów (wyrobów i usług własnej produkcji) w tys. zł (wiersz pierwszy formularza)
- Sh_1b - Przychody netto ze sprzedaży towarów i materiałów w tys. zł. (wiersz piąty formularza)

On_1b=Sw_1b+ Sh_1b obrót w cenach bieżących

Wyniki

Porównanie metod imputacji dla jednostek, które w wymaganym terminie nie złożyły sprawozdania, a dla których udało się pozyskać dane w okresie przetwarzania w danym miesiącu

różnica_mod_ets=|wartość z modelu - wartość rzeczywista|

różnica_pop_okres=|wartość z poprzedniego okresu - wartość rzeczywista|

Zmienna Sw_1b

rok	miesiąc	Liczba jednostek	różnica_mod_ets = różnica_pop_okres	różnica_mod_ets < różnica_pop_okres	różnica_mod_ets > różnica_pop_okres
2023	styczeń	1188	16,92%	48,32%	34,76%
2023	luty	1202	14,73%	40,10%	45,17%
2023	marzec	1163	16,42%	39,81%	43,77%
2023	kwiecień	931	16,97%	51,77%	31,26%
2023	maj	916	14,30%	42,25%	43,45%

Wyniki

Porównanie metod imptutacji dla jednostek, które w wymaganym terminie nie złożyły sprawozdania, a dla których udało się pozyskać dane w okresie przetwarzania w danym miesiącu

różnica_mod_ets=|wartość z modelu - wartość rzeczywista|

różnica_pop_okres=|wartość z poprzedniego okresu - wartość rzeczywista|

Zmienna Sh_1b

rok	miesiac	Liczba jednostek	różnica_mod_ets = różnica_pop_okres	różnica_mod_ets < różnica_pop_okres	różnica_mod_ets > różnica_pop_okres
2023	styczeń	1188	60,35%	25,34%	14,31%
2023	luty	1202	61,23%	15,14%	23,63%
2023	marzec	1163	61,05%	23,39%	15,56%
2023	kwiecień	931	60,79%	25,78%	13,43%
2023	maj	916	63,54%	19,32%	17,14%

Metoda – zmienne

Dodatkowo (tylko dla potrzeb analizy):

- Sz_1b - Wartość produktów wytworzonych niezaliczonych do sprzedaży (łącznie z wartością własnych wyrobów przekazanych do własnych punktów sprzedaży detalicznej, własnych hurtowni i własnych placówek gastronomicznych) w tys. zł (wiersz czwarty formularza)

Sa_1b = Sw_1b+Sz_1b-Da_1b+Dp_1b, gdzie Sa_1b - sprzedaż wyr. i usług w bieżących cenach bazowych, Da_1b – to podatek akcyzowy, Dp_1b dotacje przedmiotowe

- Pz_1b - Przeciętna liczba zatrudnionych (wiersz siódmy formularza)
- Wb_1b - Wynagrodzenia brutto osób wykazanych w wierszu 07 w tys. zł (wiersz dziewiąty formularza)

Łącznie analizowano wszystkich rekordów było 16493

Metoda – zmienne

- Ogólna charakterystyka wartości globalnych zmiennych w analizowanym zbiorze
- Jednostki średnie

Zmienna	Minimum	1. kwartyl	Mediana	Średnia	3. kwartyl	Współ. zm.
Sw_1b	0	51,17	342,7	1031,83	887,07	592%
Sz_1b	0	0	0	4,81	0	2046,52%
Sh_1b	0	0	0	672,4	128,1	470,4%
Pz_1b	0					48,03%
Wb_1b	0					68,66%

- Jednostki duże

Zmienna	Minimum	1. kwartyl	Mediana	Średnia	3. kwartyl	Współ. zm.
Sw_1b	0	881	2925	14628	8481	649,84%
Sz_1b	0	0	0	305,5	0	6819,29%
Sh_1b	0	0	33	10399	1659	1168,88%
Pz_1b	0					416,07%
Wb_1b	0					400,25%

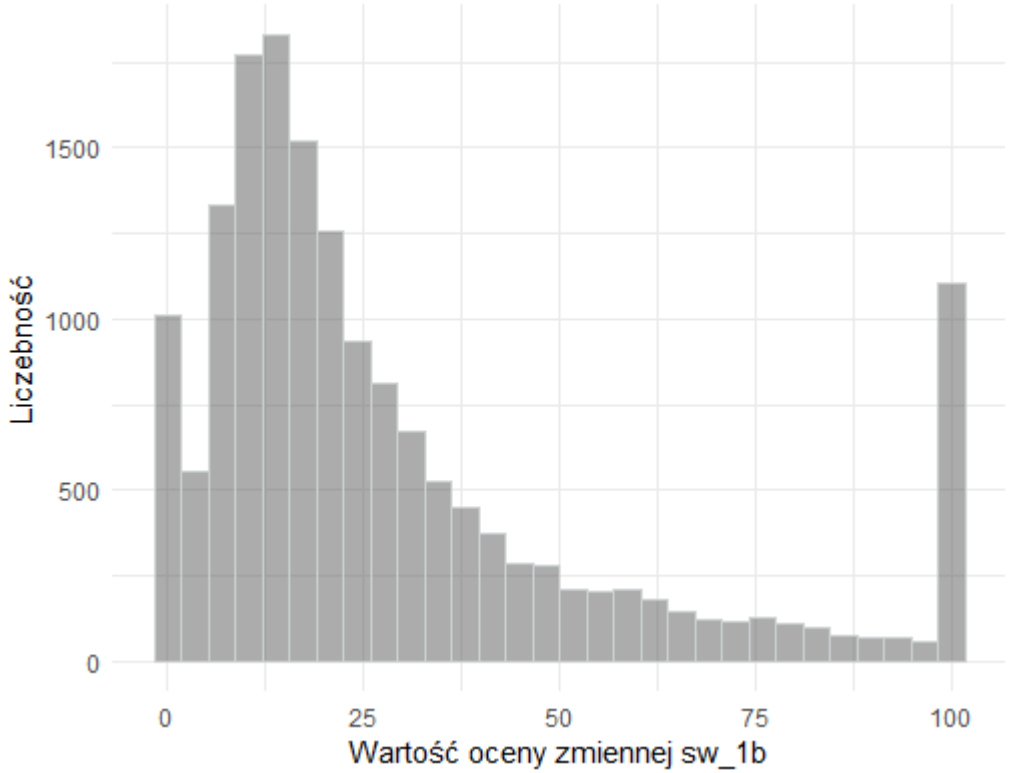
Wyniki – dodatkowa analiza

		Zmienna				
l.p.	model	sw_1b	sz_1b	sh_1b	pz_1b	wb_1b
1	ETS(A,A,A)	1	0	0	2	0
2	ETS(A,A,N)	572	51	213	1238	1210
3	ETS(A,Ad,N)	44	12	60	289	135
4	ETS(A,N,A)	7	2	7	3	20
5	ETS(A,N,N)	7220	802	6688	7284	7280
6	ETS(M,A,A)	0	0	0	1	0
7	ETS(M,A,M)	1	0	0	0	1
8	ETS(M,A,N)	1428	149	907	1155	2437
9	ETS(M,Ad,N)	91	14	60	311	177
10	ETS(M,N,A)	7	1	5	2	44
11	ETS(M,N,M)	117	11	102	1	207
12	ETS(M,N,N)	6022	272	3238	6145	4975
	razem	15510	1314	11280	16431	16486

Łącznie wszystkich rekordów było 16493

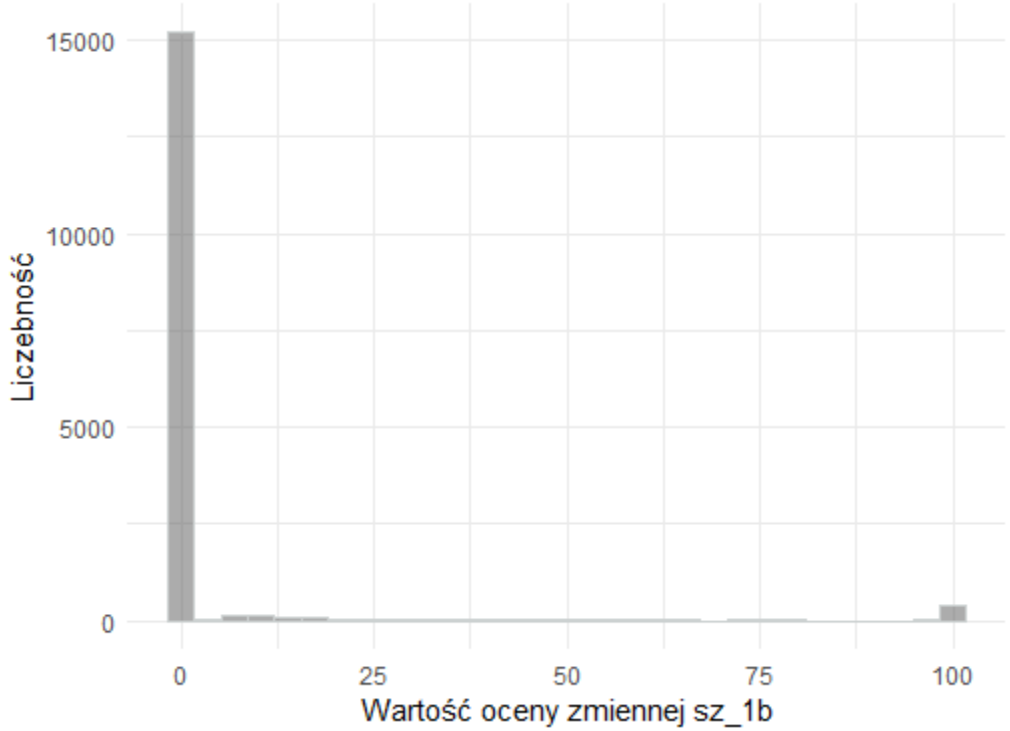
Wyniki – dodatkowa analiza

Histogramy dla wskaźników oceny prognozy



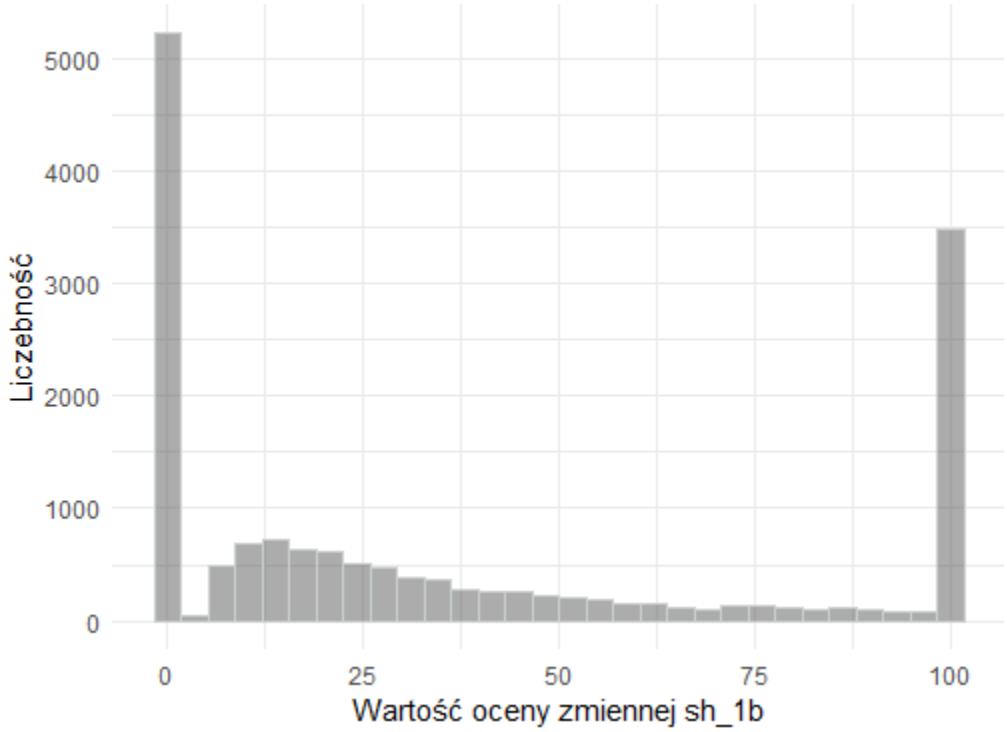
Wyniki – dodatkowa analiza

Histogramy dla wskaźników oceny prognozy



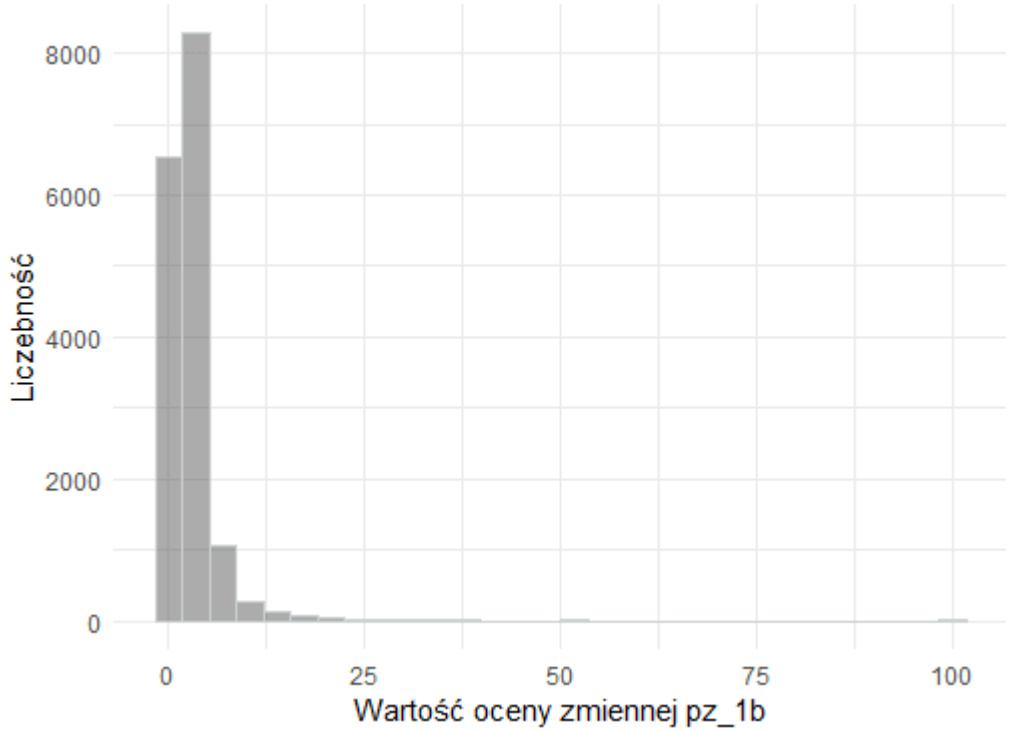
Wyniki – dodatkowa analiza

Histogramy dla wskaźników oceny prognozy



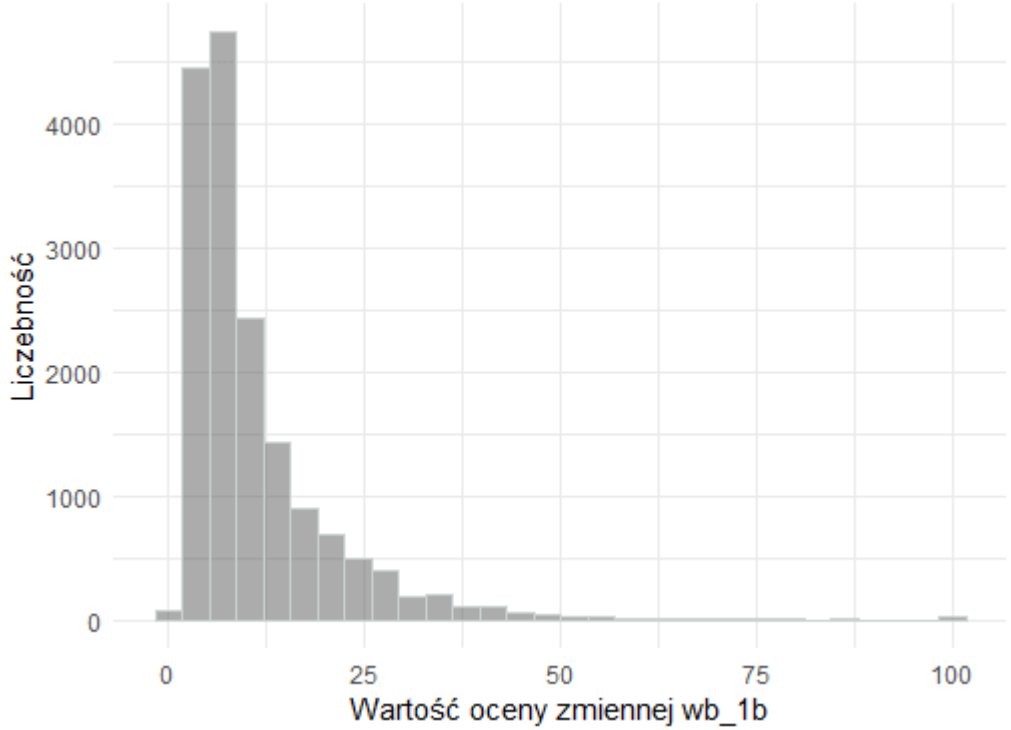
Wyniki – dodatkowa analiza

Histogramy dla wskaźników oceny prognozy

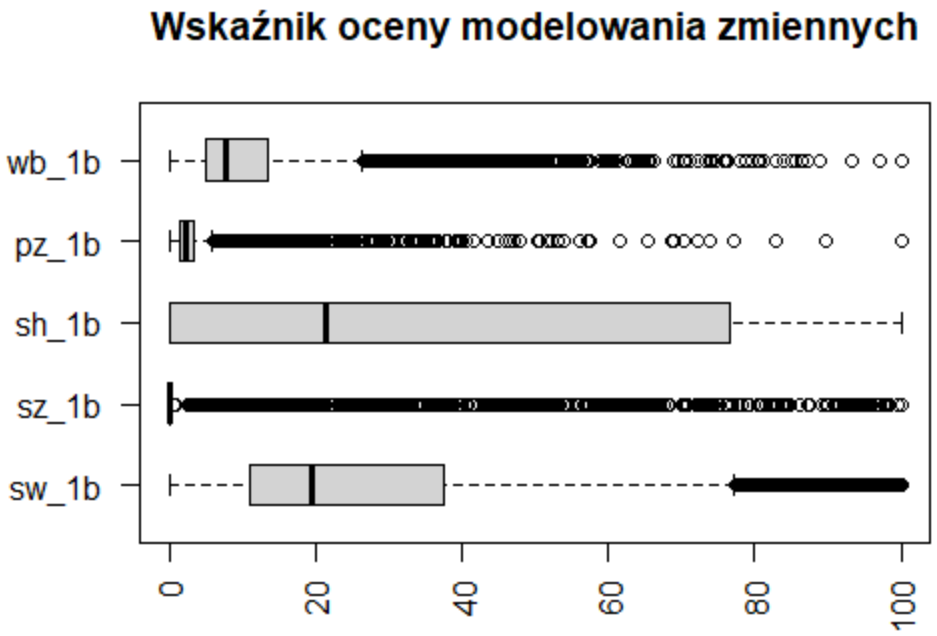


Wyniki – dodatkowa analiza

Histogramy dla wskaźników oceny prognozy



Wyniki – dodatkowa analiza



Wnioski

- Długość szeregu czasowego może być zbyt krótka,
- Wykorzystanie innego modelu,
- Imputacja masowa,
- Inna metoda niż imputacja,
- Oparcie się na rejestrach.

Dziękuję za uwagę!