



Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej



CEL BADANIA:

próba wykorzystania autokorelacji przestrzennej

do estymacji pośredniej

w badaniu podmiotów gospodarczych DG1

oraz

ocena wpływu uwzględnienia w modelu przestrzeni na

precyzję szacunku



PLAN PREZENTACJI:

- 1. METODOLOGIA I ORGANIZACJA BADANIA DG1***
- 2. ZAŁOŻENIA BADANIA SYMULACYJNEGO***
- 3. WYNIKI BADANIA SYMULACYJNEGO***
- 4. WNIOSKI***

1. METODOLOGIA I ORGANIZACJA BADANIA DG1

CEL BADANIA DG1

- **szybka** informacja o podstawowych miernikach charakteryzujących działalność gospodarczą w **przedsiębiorstwach**
- podstawa opracowywania większości **wskaźników krótkookresowych**

CZĘSTOTLIWOŚĆ

- miesięczna

ZAKRES PODMIOTOWY

osoby prawne i jednostki organizacyjne nie posiadające osobowości prawnej, których liczba pracujących wynosi **10 i więcej osób**

GŁÓWNY URZĄD STATYSTYCZNY al. Niepodległości 208. 00-925 Warszawa www.stat.gov.pl

Nazwa i adres jednostki sprawozdawczej: _____ DG-1

Meldunek o działalności gospodarczej

Numer identyfikacyjny - REGON: _____ za miesiąc 2011 r.

Portal sprawozdawczy GUS www.stat.gov.pl
Urząd Statystyczny
ul. Jana H. Dąbrowskiego 79
60-659 Poznań
Przekazać do 5. dnia roboczego po każdym miesiącu z danymi za miesiąc poprzedni

Obowiązek przekazywania danych statystycznych wynika z art. 30 pkt 3 ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej (Dz. U. Nr 88, poz. 439, z późn. zm.) oraz rozporządzenia Rady Ministrów z dnia 9 listopada 2010 r. w sprawie programu badań statystycznych statystyki publicznej na rok 2011 (Dz. U. Nr 239, poz. 1594).

Symbol rodzaju podstawowej działalności wg PKD 2007: _____

Wypełnić: WIELKIMI LITERAMI

Uwaga: Dane wartościowe należy podawać w tysiącach złotych z jednym znakiem po przecinku, wskaźnik cen - z jednym znakiem po przecinku, pozostałe dane - bez znaku po przecinku; liczbę po przecinku należy wpisać w polu oddzielonym pionową kreską.

Dział 1. Podstawowe dane o działalności gospodarczej*

Wyszczególnienie	Wzrosty	Wzrosty w tym miesiącu	
		1	2
0			
Przychody netto ze sprzedaży produktów (wytworów i usług własnej produkcji) w tys. zł	01		
w tym z działalności budowlano-montażowej wykonanej na terenie kraju i poza granicami kraju, tj. bez podwykonawców ¹⁾	02		
w tym roboty budowlane o charakterze inwestycyjnym ²⁾	03		
Wartość produktów wytworzonych niezależnie od sprzedaży (łącznie z wartością własnych wyrobów przekazanych do własnych punktów sprzedaży detalicznej, własnych hurtowni i własnych placówek gastronomicznych) w tys. zł	04		
Przychody netto ze sprzedaży towarów i materiałów w tys. zł	05		
Pracujący w osobach ³⁾ , stan w ostatnim dniu miesiąca	06		
Przeciętna liczba zatrudnionych ⁴⁾	07		
Czas fabryczny przeliczony w tys. godzin (wypełnić za miesiąc: III, VI, IX, XII - podając dane za okres od początku roku)	08		
Wynagrodzenia brutto osób wykazanych w wierszu 07 w tys. zł	09		
z wiersza 09 w tym: udziały w zysku lub w nadwyżce bilansowej w spółdzielniach	10		
przyjmuje na świadczenia emerytalne, rentowe i chorobowe wypłacone przez ubezpieczonych	11		
Sprzedaż detaliczna (łącznie z podatkiem VAT) w tys. zł	12		
w tym przez Internet	13		
Sprzedaż hurtowa (łącznie z podatkiem VAT) w tys. zł	14		
Podatek akcyzowy od wyrobów własnej produkcji (producent) w tys. zł	15		
Podatek akcyzowy od towarów i materiałów w tys. zł	16		
Dotacje przedmiotowe w tys. zł	17		
Wskaźnik cen produkcji sprzedanej (miesiąc poprzedni = 100) ⁵⁾	18		

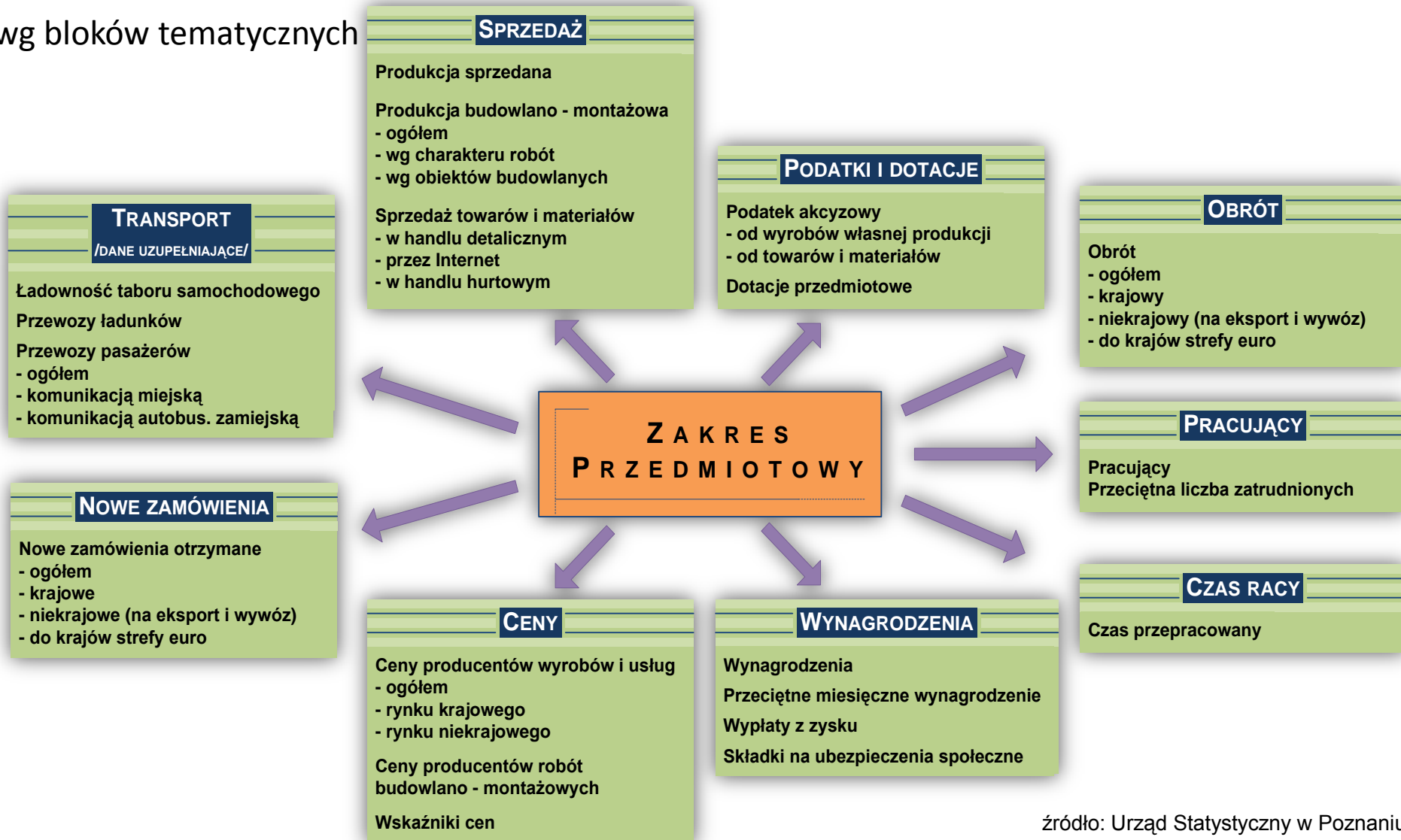




1. METODOLOGIA I ORGANIZACJA BADANIA DG1

ZAKRES PRZEDMIOTOWY BADANIA DG1

wg bloków tematycznych





1. METODOLOGIA I ORGANIZACJA BADANIA DG1



METODA OBSERWACJI EMPIRYCZNEJ

- **badanie częściowe**
- **operat liczy ok. 98 tys. jednostek**
 - **ok. 19 tys. - jednostki duże** (liczba prac. >49),
 - **ok. 80 tys. - jednostki średnie** (10 do 49 pracujących)
- **do próby wybierane są wszystkie jednostki duże**
i 10% jednostek średnich (w praktyce więcej)
- **próba liczy ok. 30 tys. jednostek**

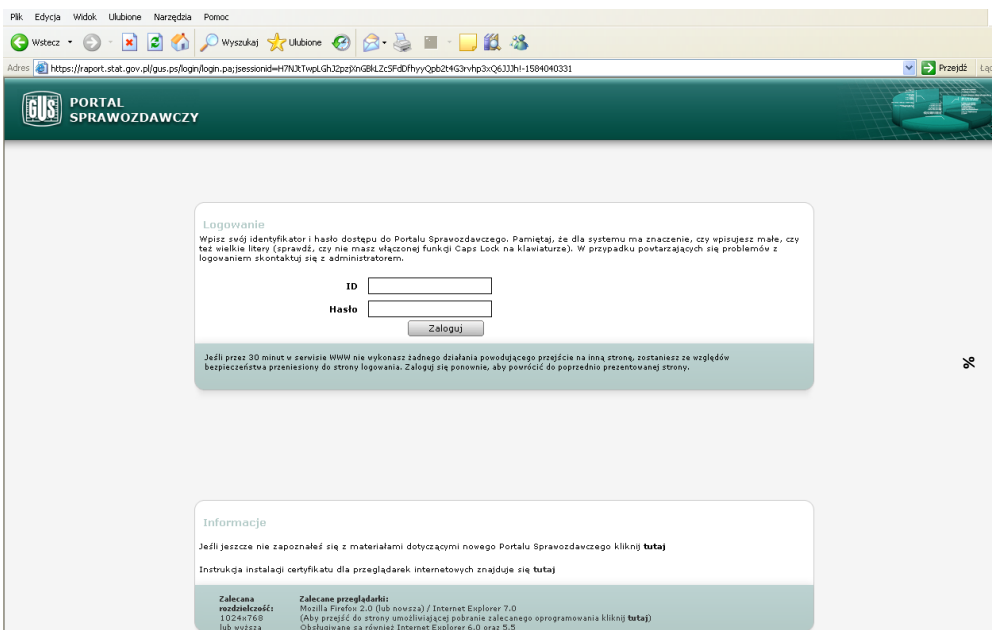


1. METODOLOGIA I ORGANIZACJA BADANIA DG1

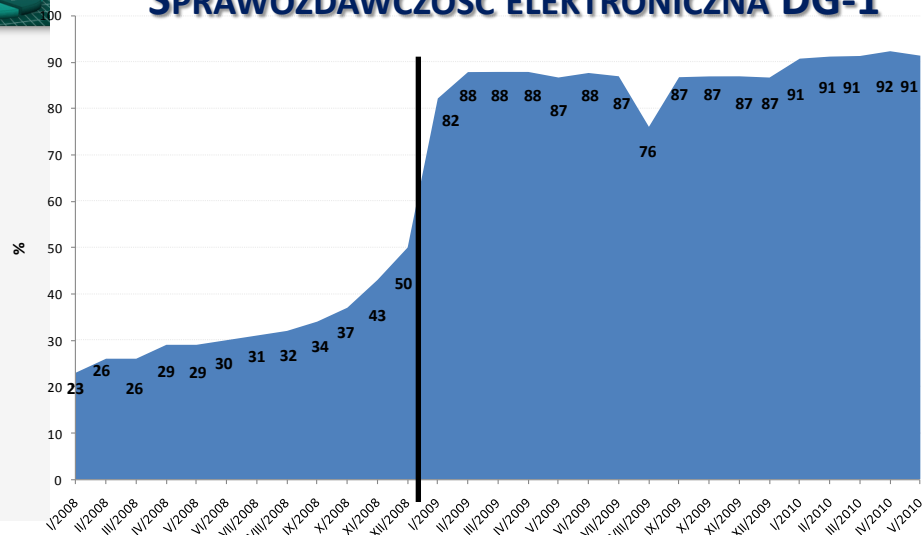


TECHNIKA OBSERWACJI EMPIRYCZNEJ

Formularz elektroniczny wypełniany jest metodą samospisu przez jednostkę sprawozdawczą do 5 dnia roboczego – *opiekun respondenta (ok.200 podmiotów)*



SPRAWOZDAWCZOŚĆ ELEKTRONICZNA DG-1





1. METODOLOGIA I ORGANIZACJA BADANIA DG1

OPRACOWYWANIE DANYCH

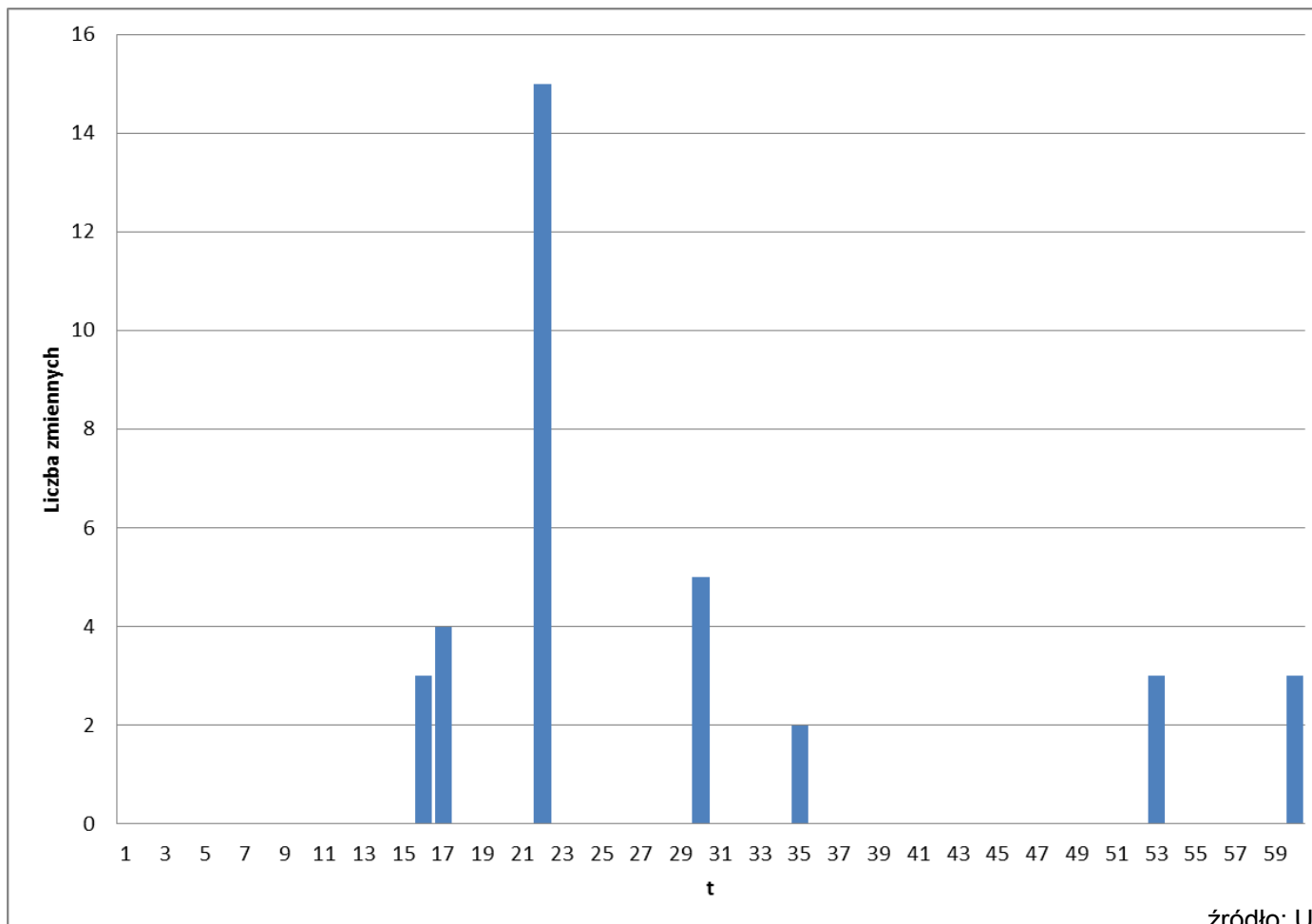
- dane są opracowywane i uogólniane dla jednostek dużych i średnich **łącznie** na poziomie wojewódzkim przez US Poznań, a następnie na poziomie ogólnopolskim w GUS.
- wyniki przesyłane są do GUS **10** dnia roboczego.
- wyniki prezentuje się **16** dnia roboczego.



1. METODOLOGIA I ORGANIZACJA BADANIA DG1



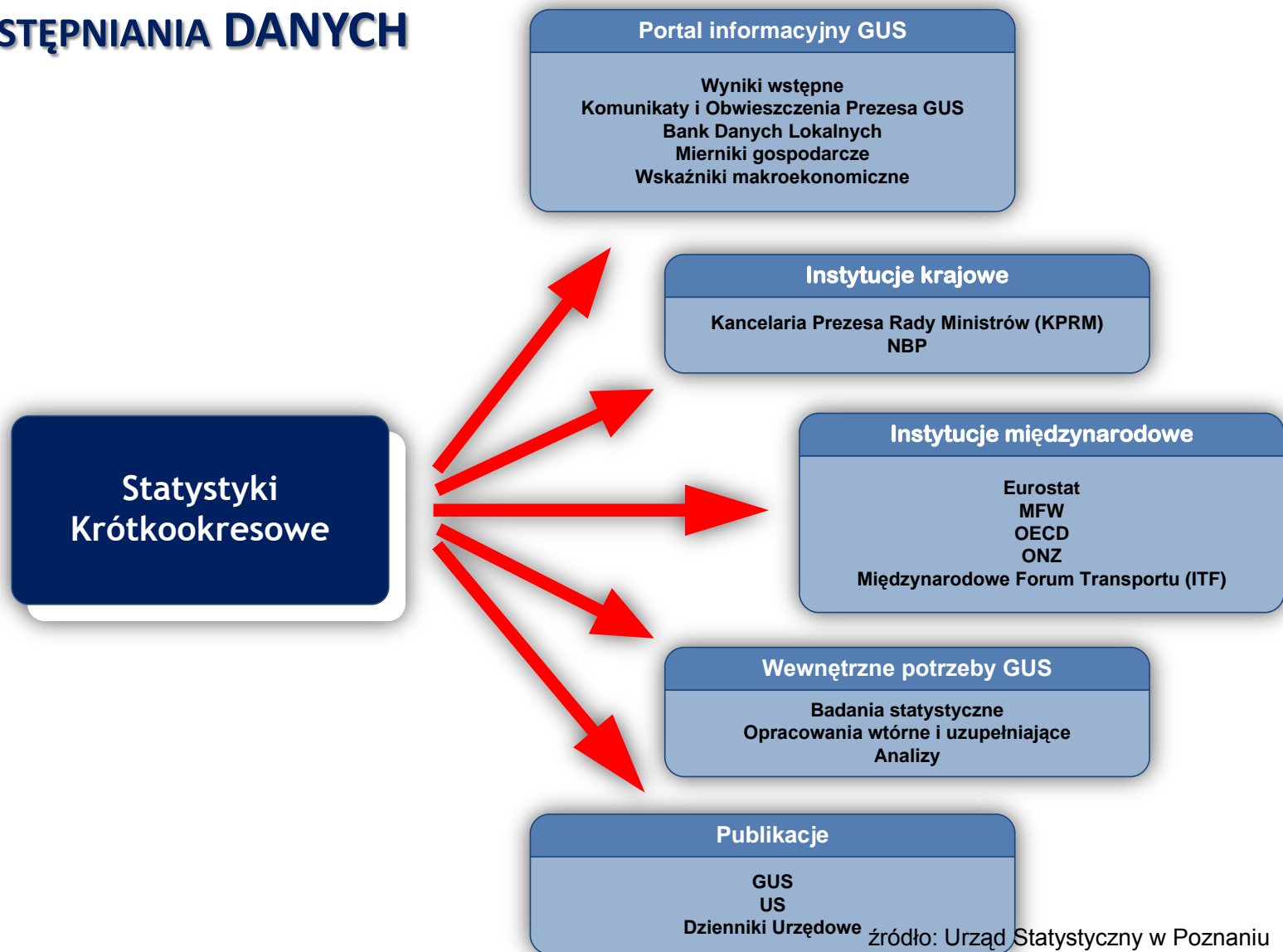
LICZBA ZMIENNYCH PUBLIKOWANYCH NA PODSTAWIE BADANIA DG1





1. METODOLOGIA I ORGANIZACJA BADANIA DG1

OBSZARY UDOSTĘPNIANIA DANYCH





2. ZAŁOŻENIA BADANIA SYMULACYJNEGO

1. **Populacja generalna** - 23962 podmioty gospodarcze (duże i średnie), które wzięły udział w badaniu DG1 w sierpniu 2012 r.
2. **Domena** – przekrój województwo (NTS2) x sekcja
3. **Wybrane sekcje** - rolnictwo (A), przemysł (C), budownictwo (F), handel (G)
4. **Zmienne badane:**
 - liczba pracujących zmienna
 - wynagrodzenia brutto
5. **Zmienne pomocnicze**
 - liczba pracujących stała
 - przychody ze sprzedaży produktów (wyrobów i usług)
 - przychody ze sprzedaży towarów i materiałów
6. **Badanie symulacyjne** - wylosowano 1000 prób 20% (rolnictwo 30%)
7. **Metody estymacji:** - DIRECT - GREG -SYNTH -EBLUP - SEBLUP



2. ZAŁOŻENIA BADANIA SYMULACYJNEGO



CHARAKTERYSTYKA STATYSTYCZNA ROZKŁADU BADANYCH ZMIENNYCH

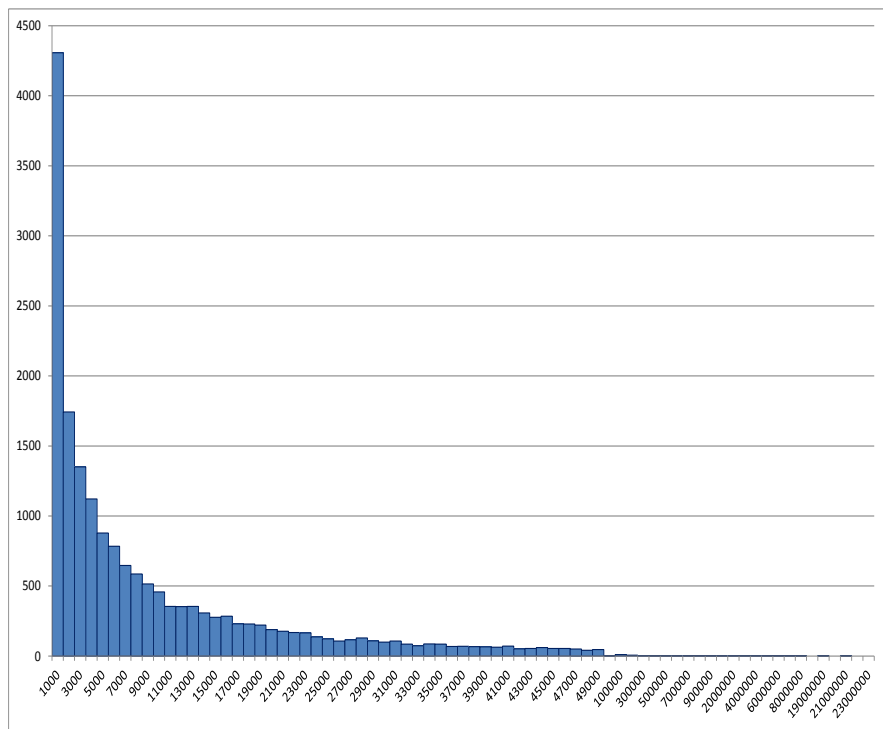
Charakterystyki statystyczne	Przychody netto ze sprzedaży produktów (wyr. i usług własnej produkcji) w tys. zł	Przychody netto ze sprzedaży towarów i materiałów w tys. zł	Przychody netto ze sprzedaży ogółem w tys. zł	Wynagrodzenia brutto w tys. zł
min	0	0	0	0
max	3918065	5468647	94368,3	9386712
Q_1	72,2	0	64	357
Q_2	417	56,7	149,7	1080,1
Q_3	1613	800	351,2	3342,8
średnia	3578	3116	442	6694
$s(x)$	37284	46298	1606	73740
$V_{s(x)}$	1042	1486	364	1102



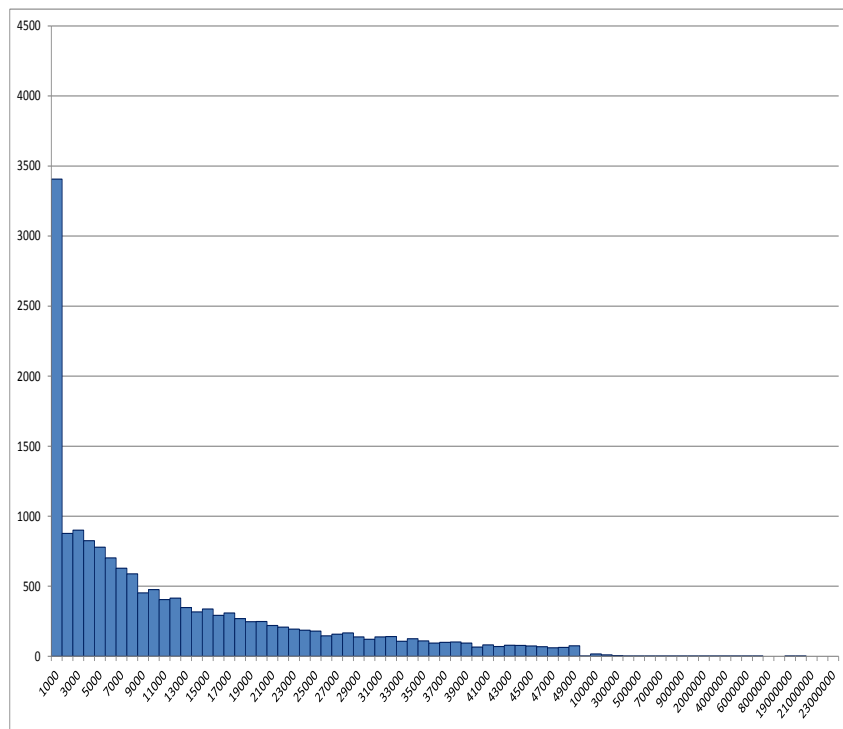
2. ZAŁOŻENIA BADANIA SYMULACYJNEGO

OCENA ROZKŁADÓW SZACOWANYCH ZMIENNYCH

- ROZKŁAD PODMIOTÓW WEDŁUG WARTOŚCI ROCZNEGO PRZYCHODU



Źródło: Opracowanie własne na podstawie badania DG1, 2008



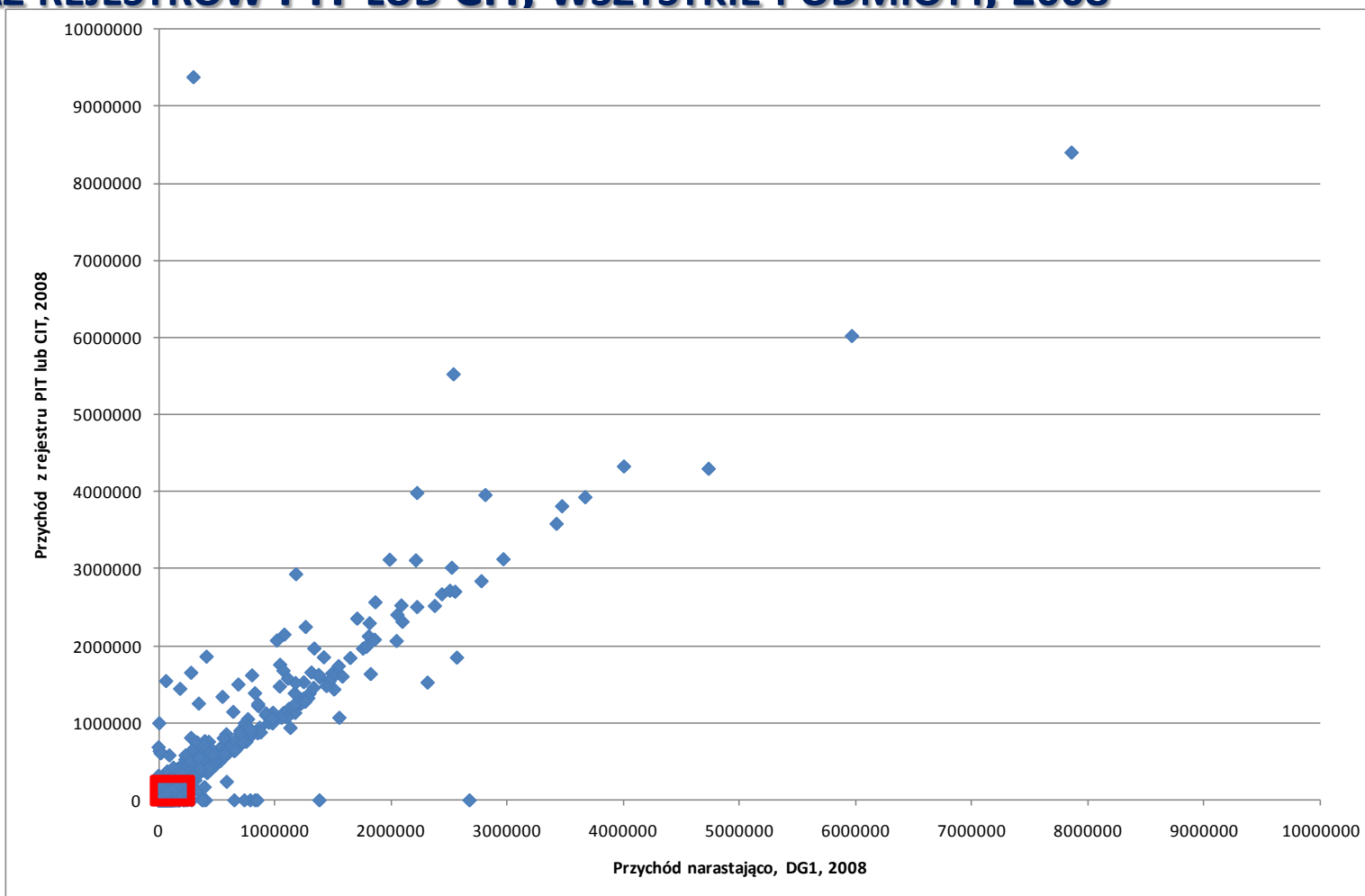
Źródło: Opracowanie własne na podstawie rejestru PIT lub CIT, 2008



2. ZAŁOŻENIA BADANIA SYMULACYJNEGO



RELACJA MIĘDZY WARTOŚCIAMI ZMIENNEJ PRZYCHÓD Z BADANIA DG1 ORAZ REJESTRÓW PIT LUB CIT, WSZYSTKIE PODMIOTY, 2008

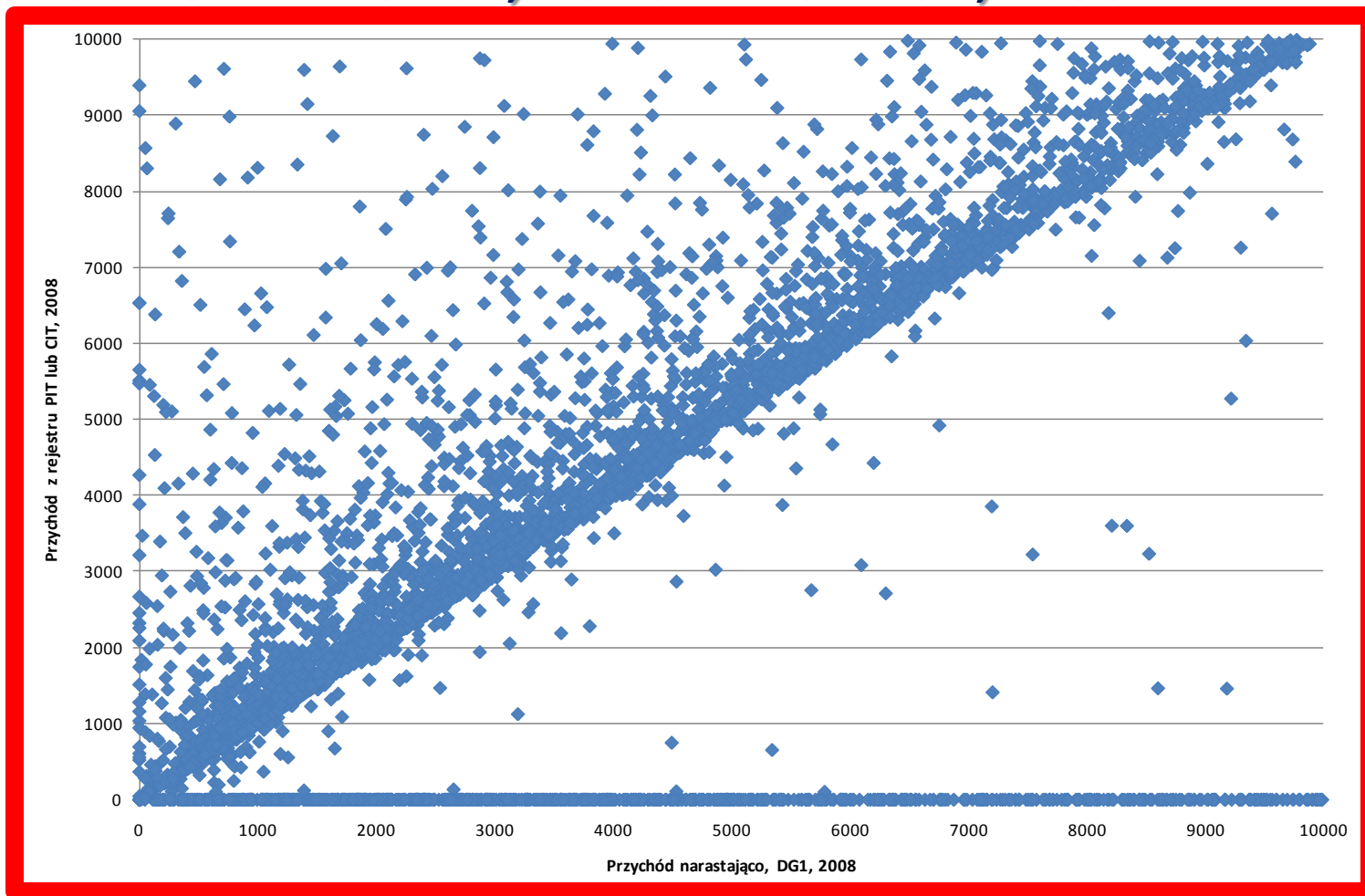


Skala osi uwzględniająca jednostki o najwyższych przychodach (w ograniczeniu do 10 000 000)



2. ZAŁOŻENIA BADANIA SYMULACYJNEGO

RELACJA MIĘDZY WARTOŚCIAMI ZMIENNEJ PRZYCHÓD Z BADANIA DG1 ORAZ REJESTRÓW PIT LUB CIT, WSZYSTKIE PODMIOTY, 2008

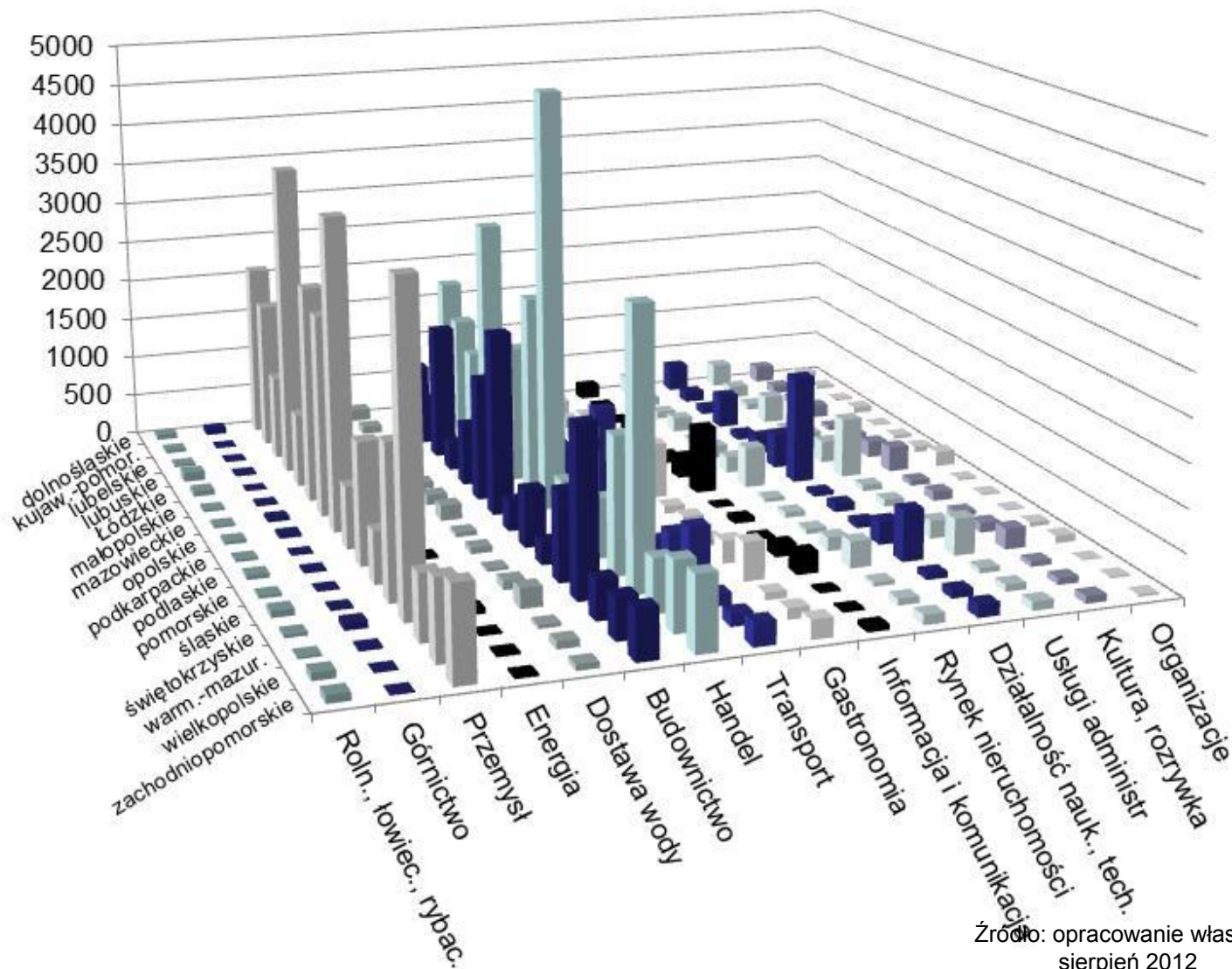


Skala osi nieuwzględniająca jednostek o najwyższych przychodach (w ograniczeniu do 10 000)



2. ZAŁOŻENIA BADANIA SYMULACYJNEGO

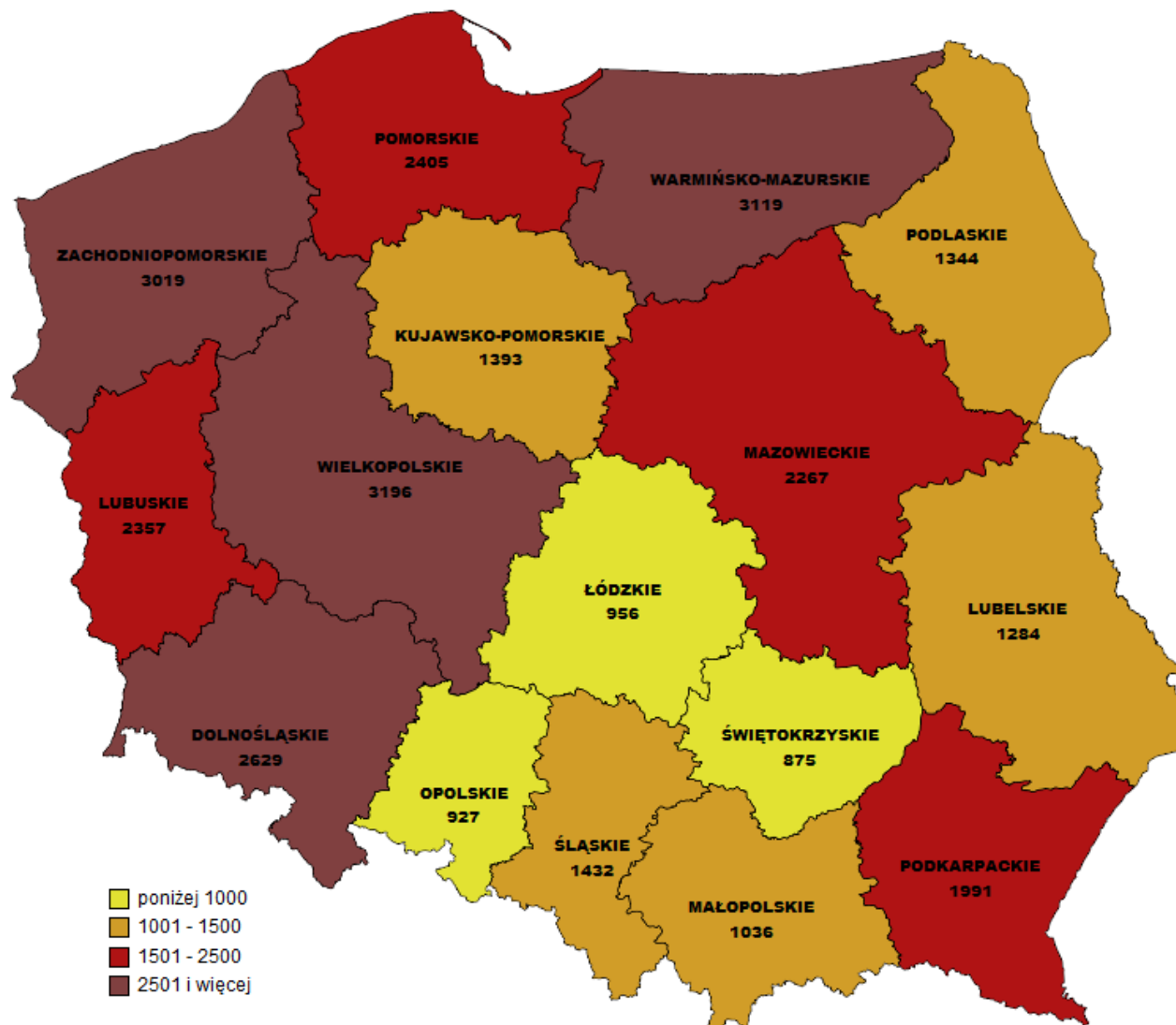
Liczba podmiotów wg województw i sekcji PKD 2007, sierpień 2012





3. WYNIKI BADANIA SYMULACYJNEGO

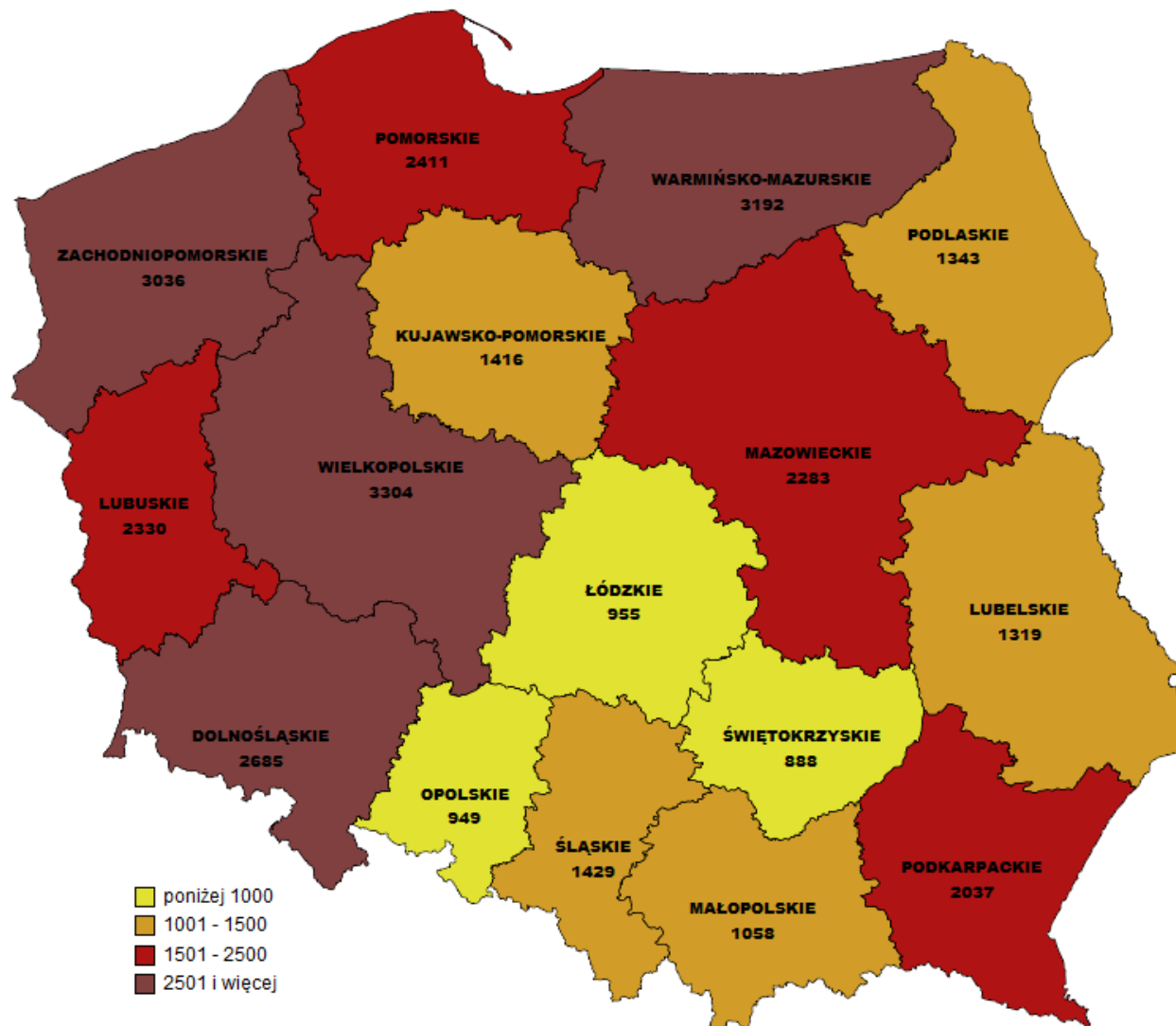
Liczba pracujących zmienna – sekcja A (Rolnictwo)





3. WYNIKI BADANIA SYMULACYJNEGO

Liczba pracujących stała – sekcja A (Rolnictwo)

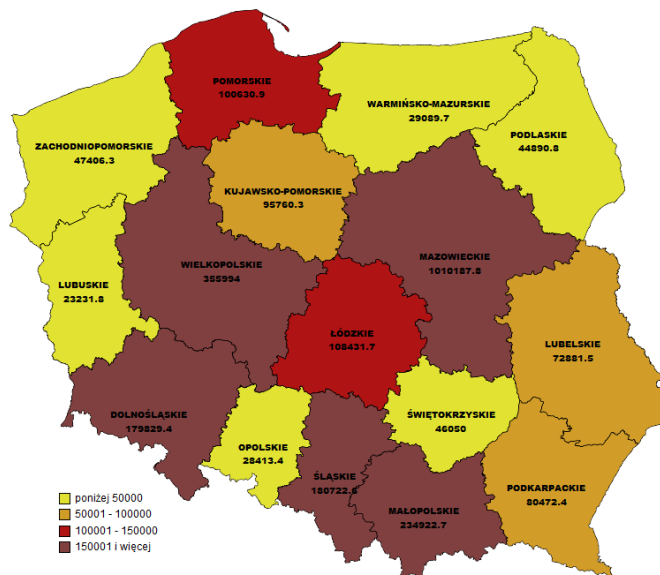
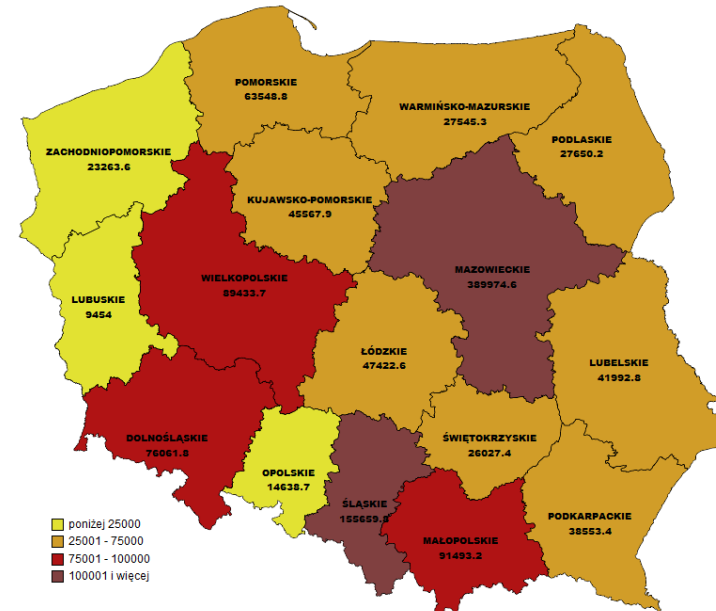
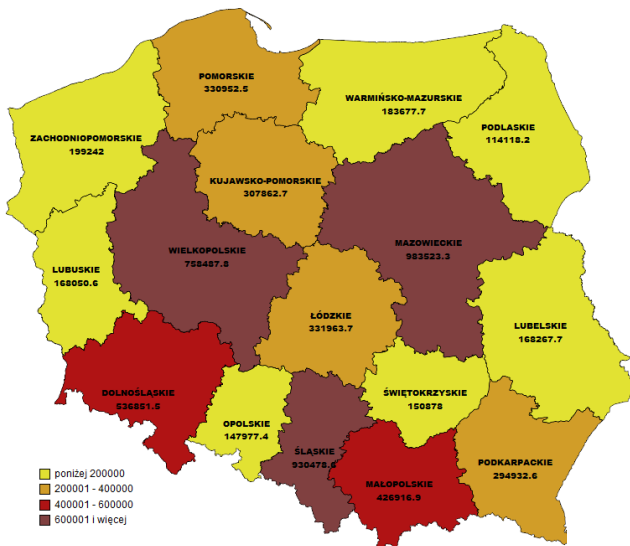




3. WYNIKI BADANIA SYMULACYJNEGO



Wynagrodzenia – sekcje C, F, G (Przemysł, Budownictwo, Handel)

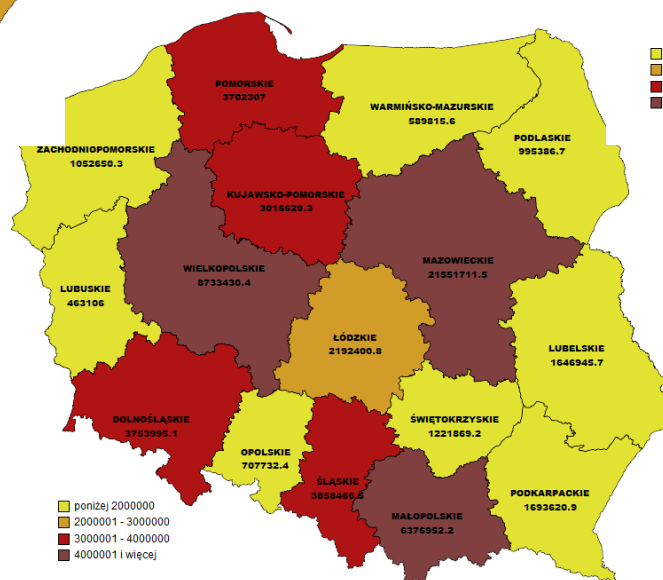
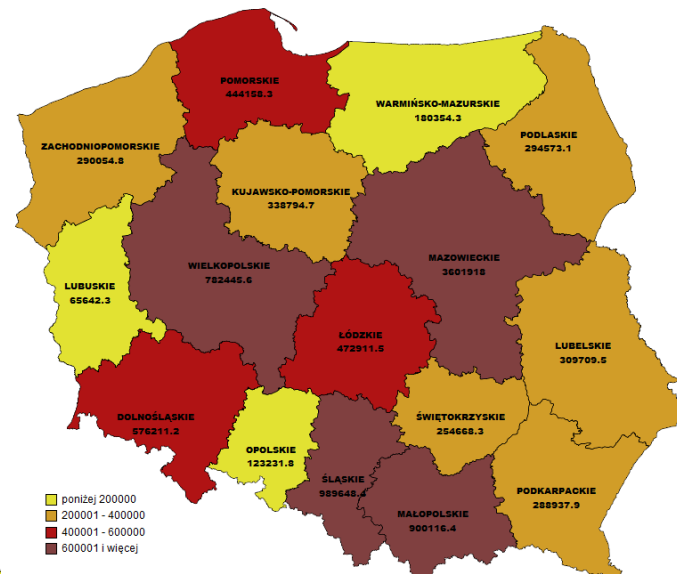




3. WYNIKI BADANIA SYMULACYJNEGO



Przychody ze sprzedaży – sekcje C, F, G (Przemysł, Budownictwo, Handel)





3. WYNIKI BADANIA SYMULACYJNEGO

Motywacje

- **Charakterystyki gospodarstwa rolnego (SKAD 2010)**
- **Rynek pracy (projekt ESSnet on Small Area Estimation 2011)**
- **Rynek nieruchomości mieszkaniowych (SKAD 2012)**

Wykorzystanie estymacji pośredniej wykorzystujących autokorelację przestrzenną

Zastosowanie estymatora EBLUP uwzględniającego korelację przestrzenną wynika z naturalnej konsekwencji uchylenia, krępującego założenia, o niezależności efektów związanych z każdą parą obszarów (por. A. Saei, R. Chambers, 2004).

Oprócz efektów losowych związanych z obszarem, model może zostać wzmocniony informacjami na temat odległości między obszarami w przestrzeni geograficznej. Jak wskazuje się w literaturze, naturalnym wyborem jest uwzględnienie w procedurach estymacyjnych współrzędnych x i y dla każdego małego obszaru. Najczęściej jako punkt charakterystyczny małego obszaru wybiera się jego środek – tzw. centroidę. Następnie oblicza się, w oparciu o wyznaczone centroidy, macierz odległości z uwzględnieniem modelu wykładniczego lub potęgowego.



3. WYNIKI BADANIA SYMULACYJNEGO

Estymator EBLUP

Liniowy model o skorelowanych przestrzennie efektach losowych dla małych obszarów jest dany następującym wzorem (por. A. Saei, R. Chambers, 2004, s. 1):

$$y_{di} = \beta_0 + X_{di}^T \beta + \sum_{d=1}^D u_d + e_{di}$$

gdzie:

β_0 jest wyrazem wolnym modelu regresji;

β jest współczynnikiem regresji;

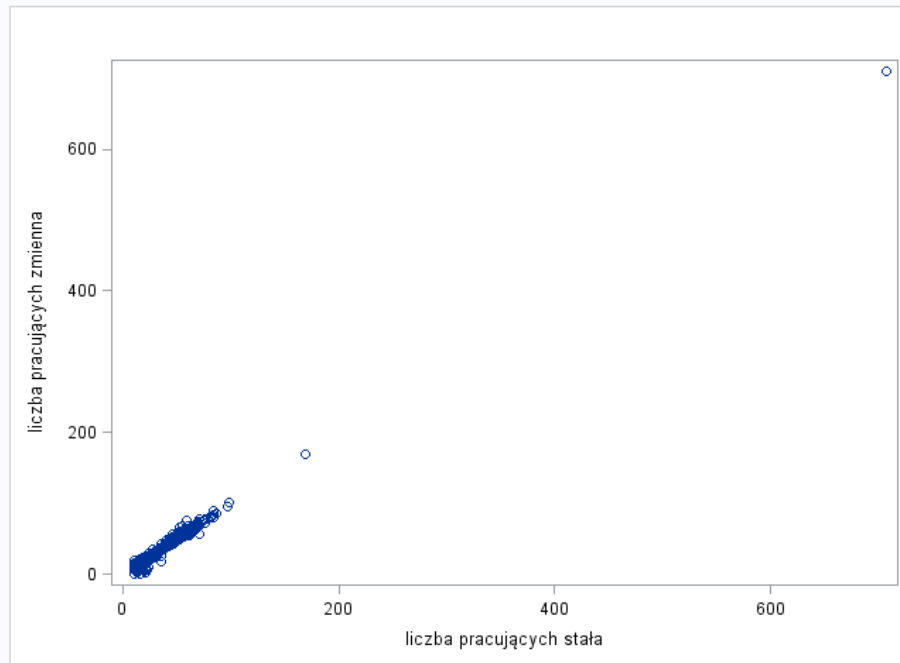
d jest subskryptem oznaczającym mały obszar, $d = 1, 2, \dots, D$;

u_d , e_{di} są efektami losowymi związanymi odpowiednio z obszarami i obserwacjami, o których zakłada się, że są niezależne i mają rozkłady normalne o wartości oczekiwanej zero i pewnej stałej wariancji.

3. WYNIKI BADANIA SYMULACYJNEGO

Charakterystyka modeli

- dla szacowania liczby pracujących (zmiennej) w sekcji A



Root MSE	3.15388	R-Square	0.9901
Dependent Mean	36.03099	Adj R-Sq	0.9901
Coeff Var	8.75325		

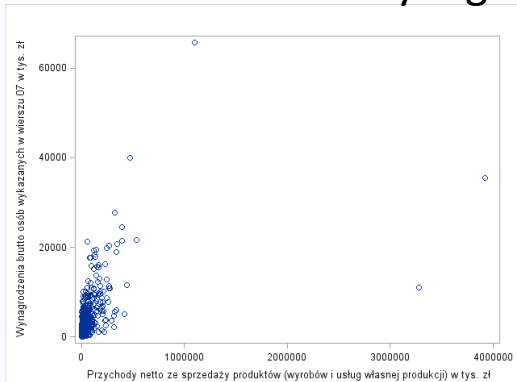
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-1.19772	0.16844	-7.11	<.0001
X1	liczba pracujących stała	1	1.01958	0.00352	289.68	<.0001



3. WYNIKI BADANIA SYMULACYJNEGO

Charakterystyka modeli

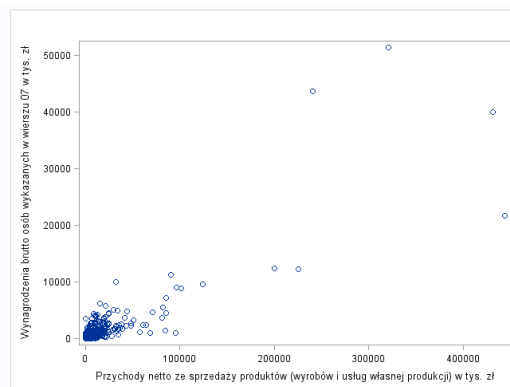
- dla szacowania wynagrodzeń w sekcjach C, F i G



Root MSE	1269.78171	R-Square	0.2852
Dependent Mean	523.48236	Adj R-Sq	0.2852
Coeff Var	242.56437		

Parameter Estimates

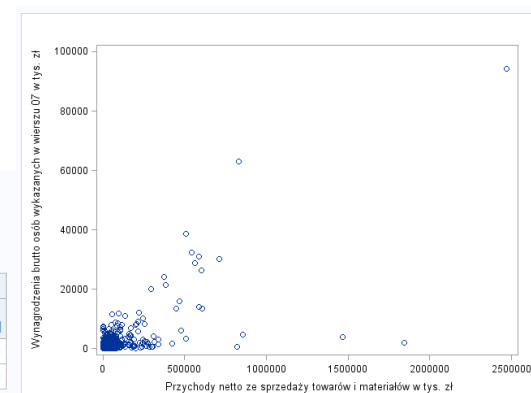
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	432.03889	11.90352	36.30	<.0001
X3	Przychody netto ze sprzedaży produktów (wyrobów i usług własnej produkcji) w tys. zł	1	0.01519	0.00022395	67.81	<.0001



Root MSE	784.81474	R-Square	0.7483
Dependent Mean	348.95096	Adj R-Sq	0.7482
Coeff Var	224.90689		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	88.01285	13.81361	6.37	<.0001
X3	Przychody netto ze sprzedaży produktów (wyrobów i usług własnej produkcji) w tys. zł	1	0.08813	0.00088359	99.74	<.0001



Root MSE	1226.40812	R-Square	0.5244
Dependent Mean	344.64089	Adj R-Sq	0.5243
Coeff Var	355.85102		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	146.48882	14.18039	10.33	<.0001
X4	Przychody netto ze sprzedaży towarów i materiałów w tys. zł	1	0.02465	0.00026829	91.87	<.0001



3. WYNIKI BADANIA SYMULACYJNEGO



Testowanie autokorelacji globalnej i lokalnej – statystyka I Morana

Dodatnia i istotna wartość **statystyki globalnej I Morana** wskazuje, że istnieje dodatnia autokorelacja przestrzenna, czyli podobieństwo badanych obiektów w sąsiedztwie.

Statystyka lokalna wskazuje czy dany region jest otoczony przez regiony sąsiedzkie o podobnych lub różnych wartościach badanej cechy w stosunku do losowego rozmieszczenia tych wartości w przestrzeni.

Może być wykorzystywany do wykrywania lokalnych klastrów i tzw. regionów nietypowych.

Dodatnia i istotna wartość statystyki lokalnej Morana że dany region jest otoczony przez regiony o podobnych wartościach cechy. Natomiast w przypadku ujemnych istotnych wartości oznacza to sytuację w której dany region jest otoczony przez regiony o znacząco różnych wartościach danej cechy.

```
Moran's I test under randomisation

data: Dataset$Y
weights: wojewod.B.listw

Moran I statistic standard deviate = -2.4121, p-value = 0.00793
alternative hypothesis: less
sample estimates:
Moran I statistic      Expectation      Variance
-0.272060592         -0.066666667      0.007250594
```

Sekcja	> Zmienna analizowana	Statystyka Morana	p-value
A	Liczba pracujących zmienna	0,23	0,02
C	Wynagrodzenia brutto	-0,35	0,02
F	Wynagrodzenia brutto	-0,25	0,02
G	Wynagrodzenia brutto	-0,27	0,01

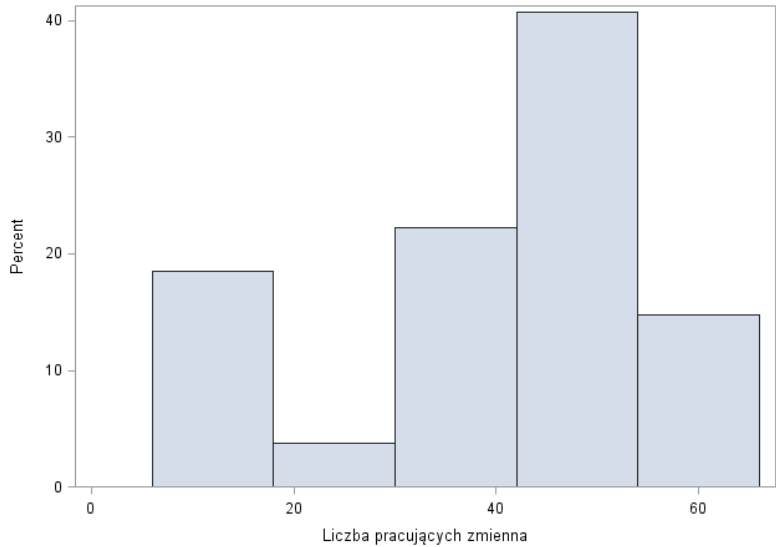


3. WYNIKI BADANIA SYMULACYJNEGO

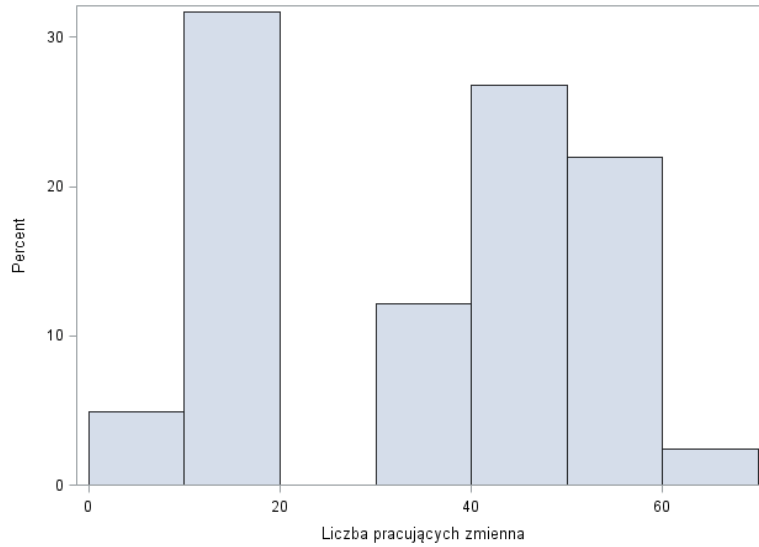
sekcja A - ROLNICTWO



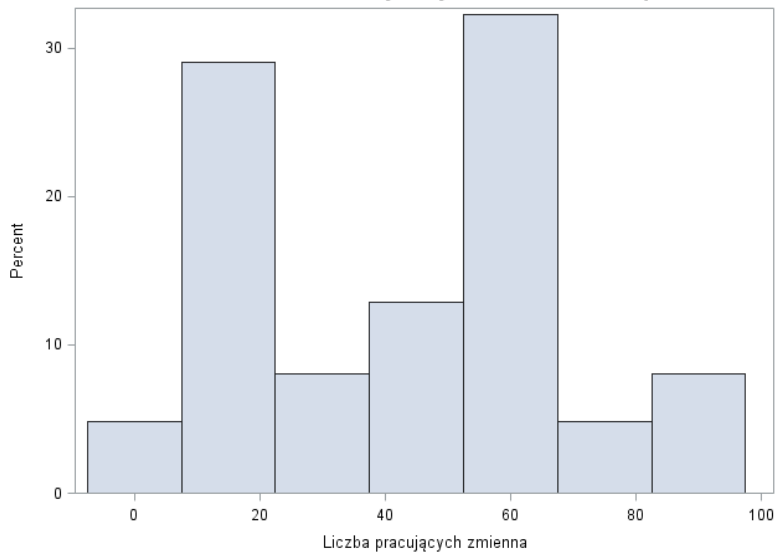
Rozkład wartości zmiennej w województwie MAŁOPOLSKIE



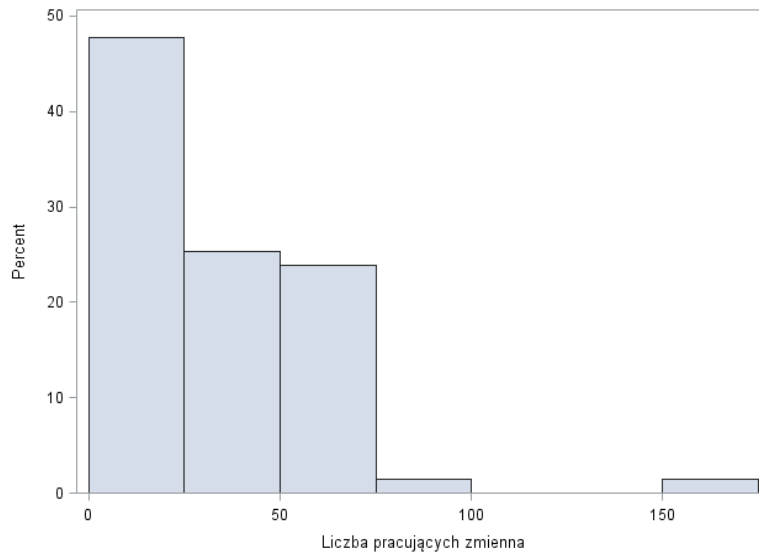
Rozkład wartości zmiennej w województwie KUJAWSKO-POMORSKIE



Rozkład wartości zmiennej w województwie DOLNOŚLĄSKIE



Rozkład wartości zmiennej w województwie LUBUSKIE



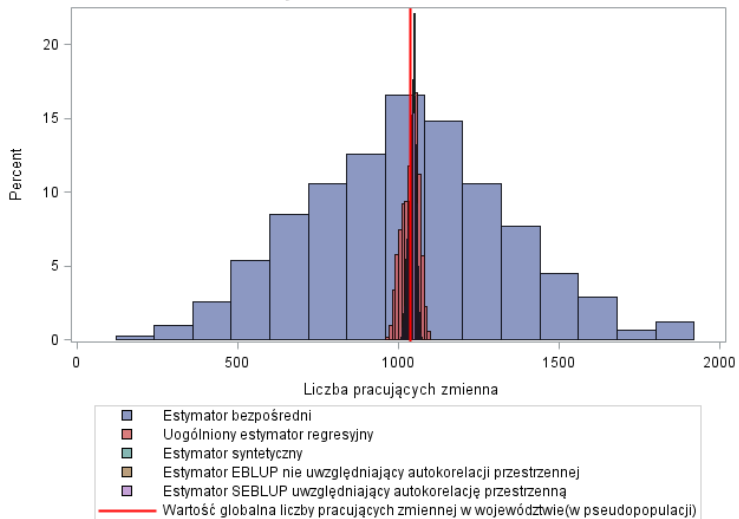


3. WYNIKI BADANIA SYMULACYJNEGO

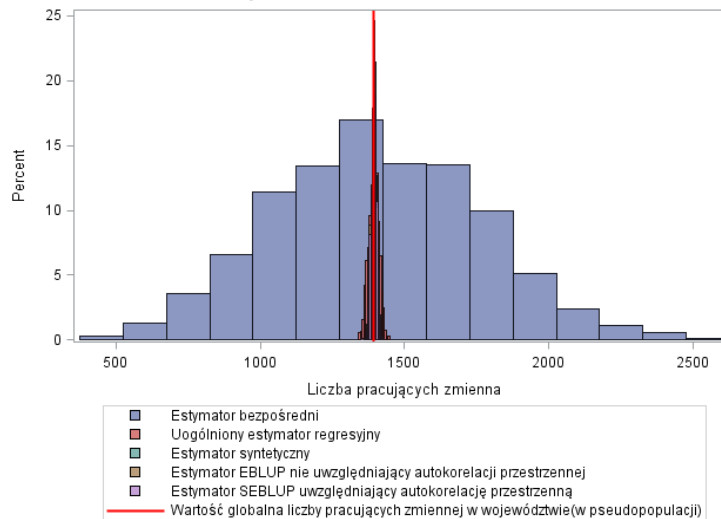
sekcja A - ROLNICTWO



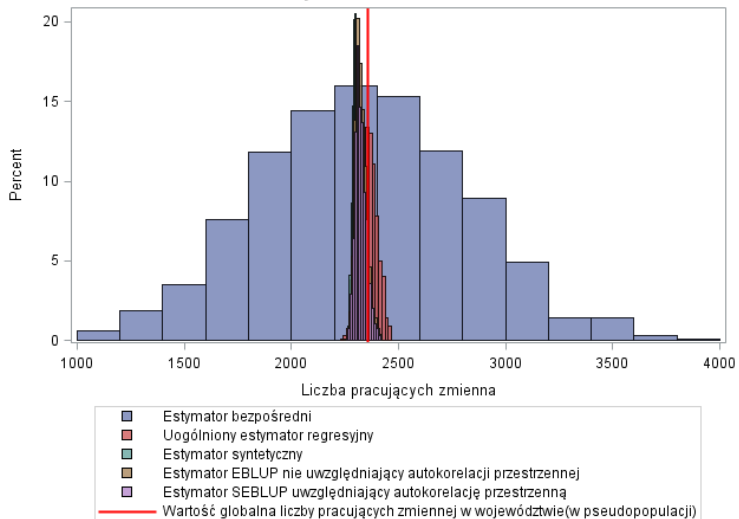
Rozkład ocen estymatorów - WOJEWÓDZTWO MAŁOPOLSKIE



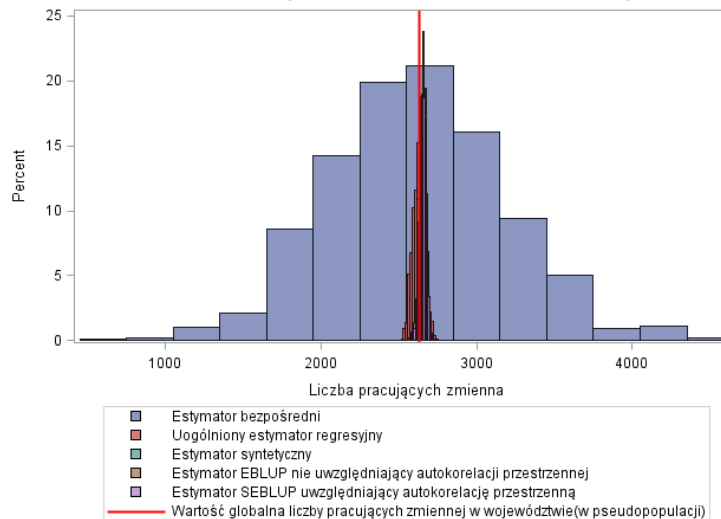
Rozkład ocen estymatorów - WOJEWÓDZTWO KUJAWSKO-POMORSKIE



Rozkład ocen estymatorów - WOJEWÓDZTWO LUBUSKIE

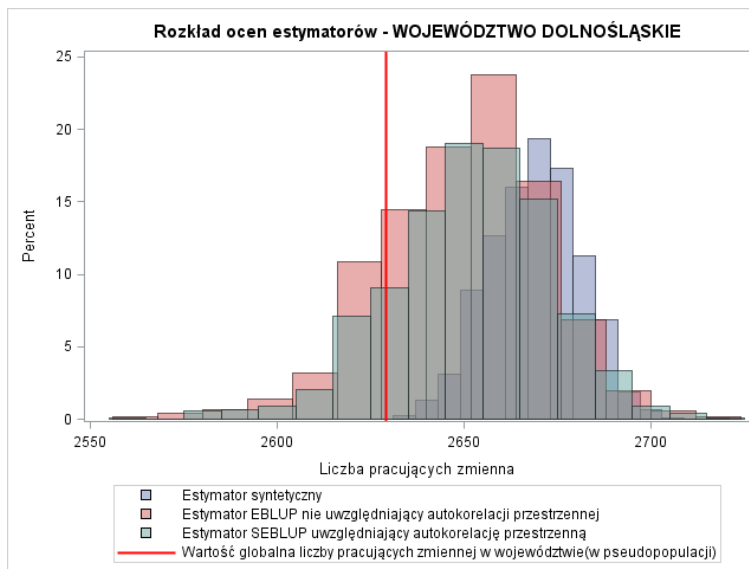
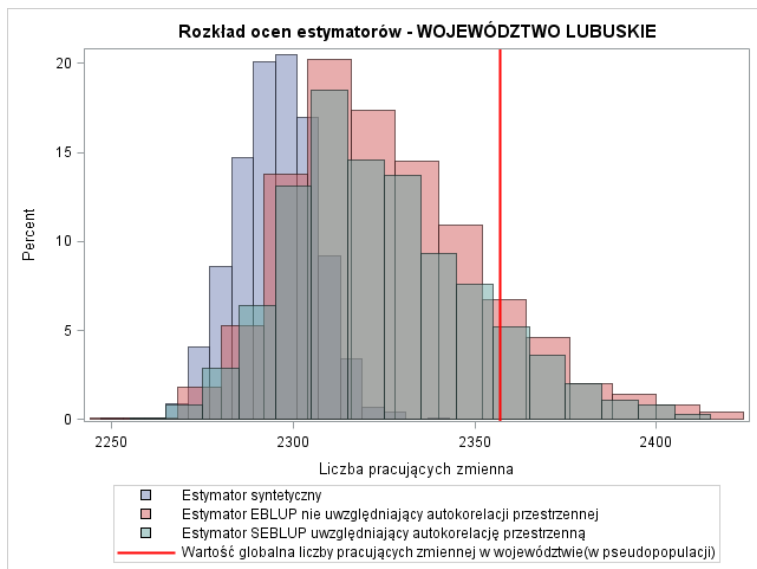
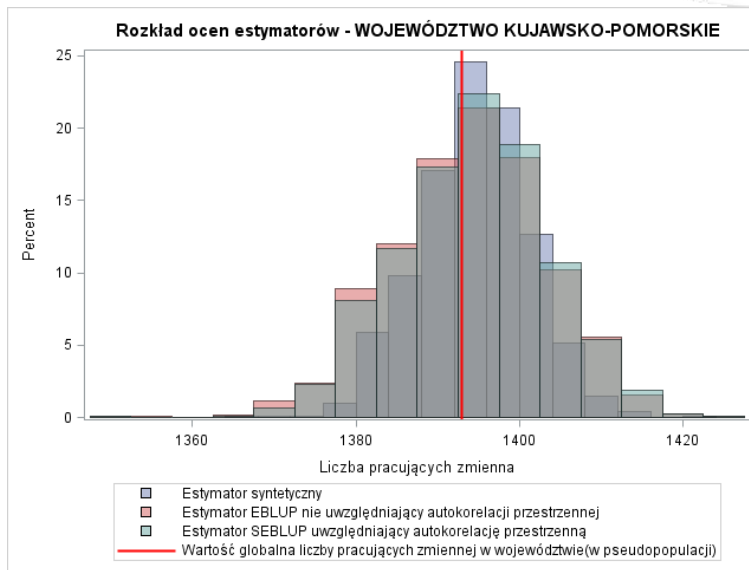
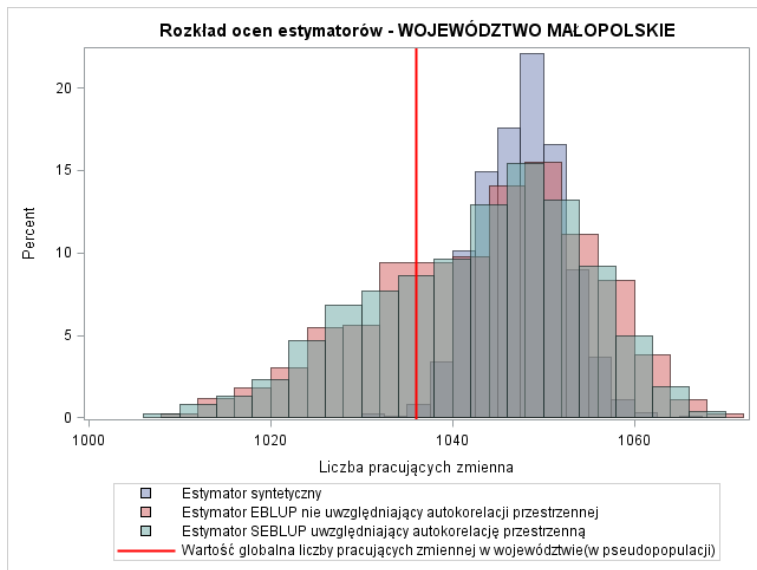


Rozkład ocen estymatorów - WOJEWÓDZTWO DOLNOŚLĄSKIE



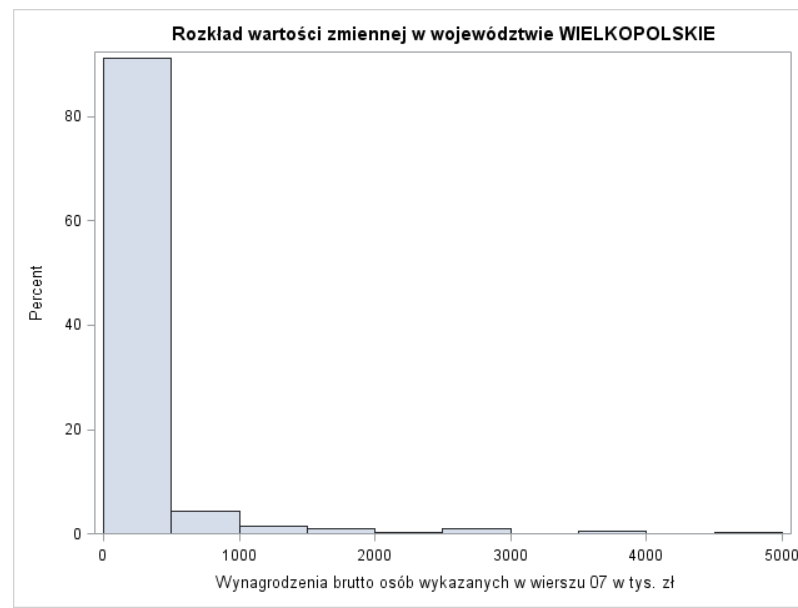
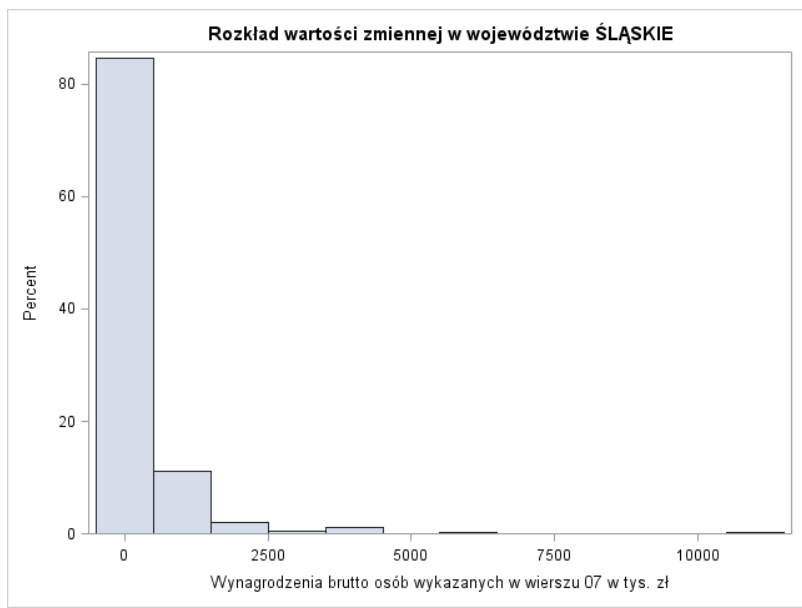
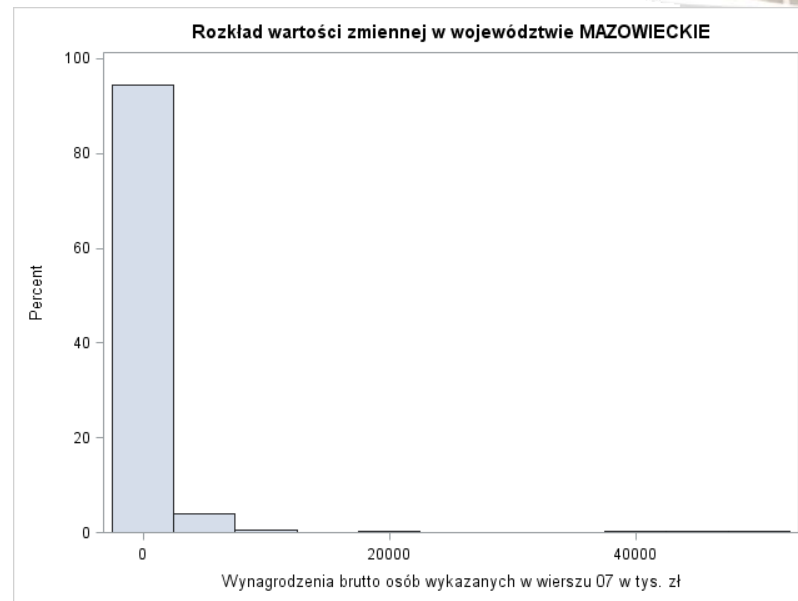
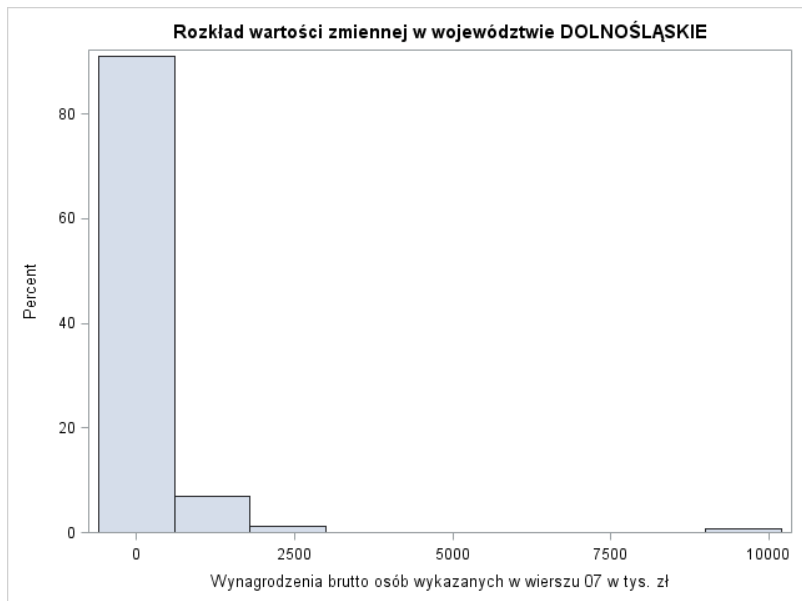
3. WYNIKI BADANIA SYMULACYJNEGO

sekcja A - ROLNICTWO



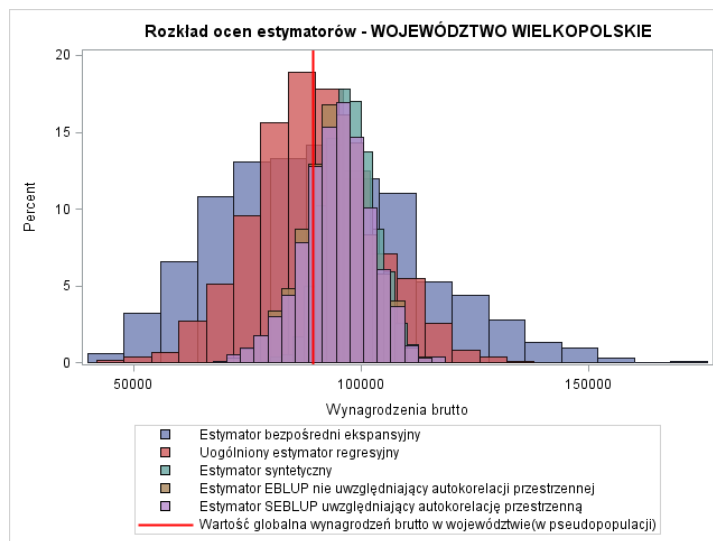
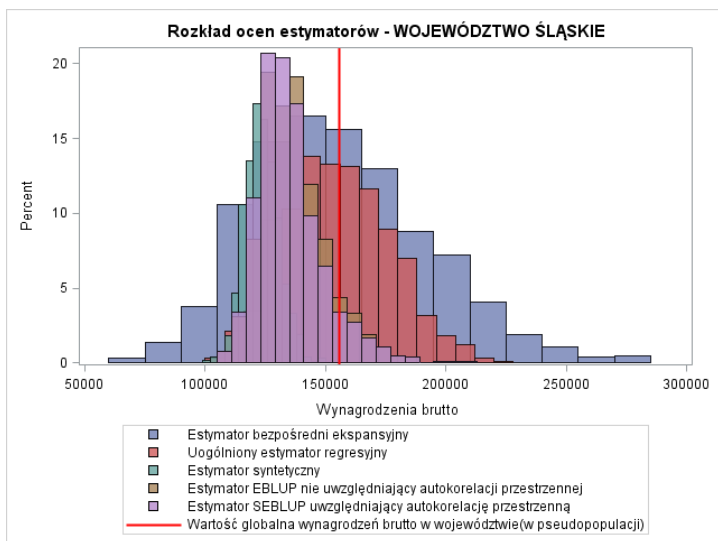
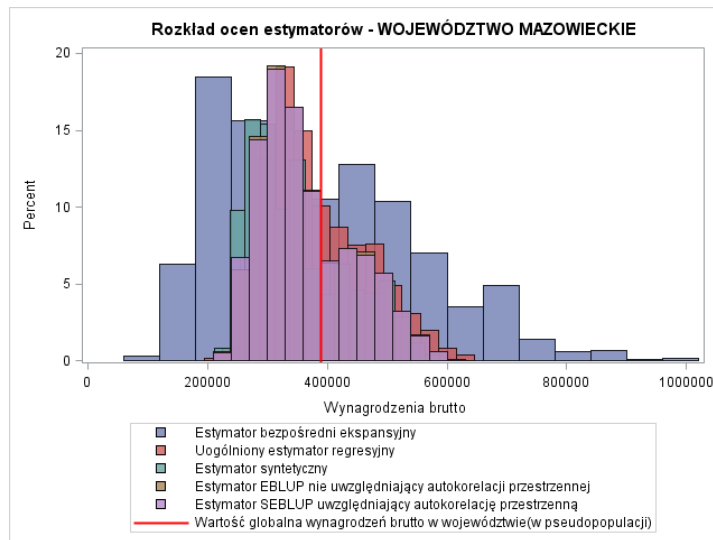
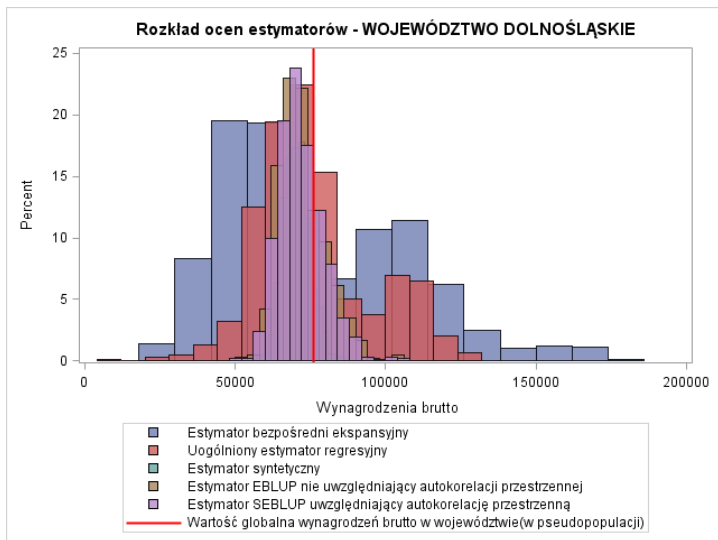
3. WYNIKI BADANIA SYMULACYJNEGO

sekcja F - BUDOWNICTWO



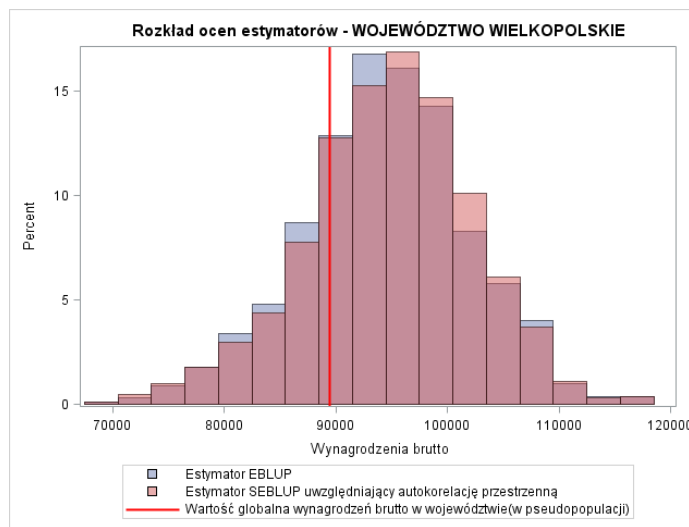
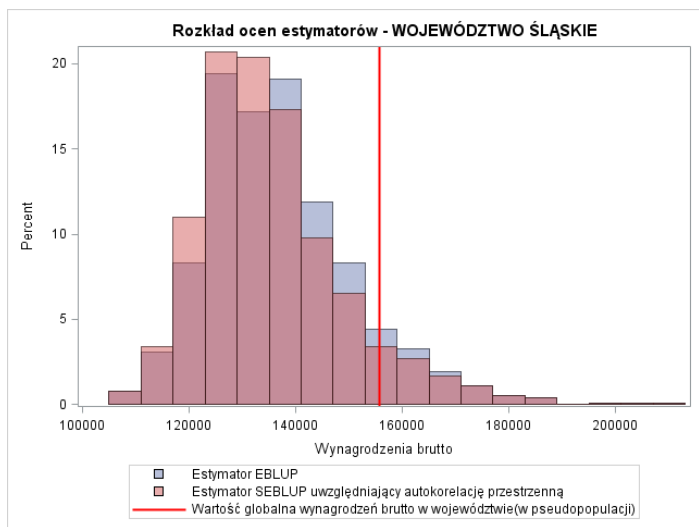
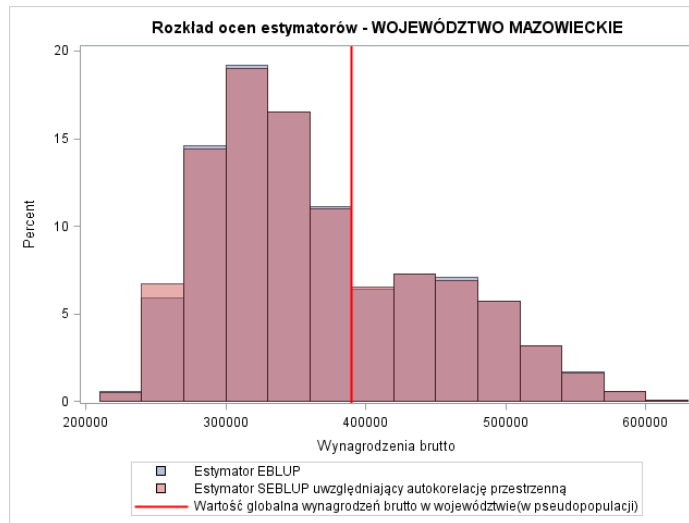
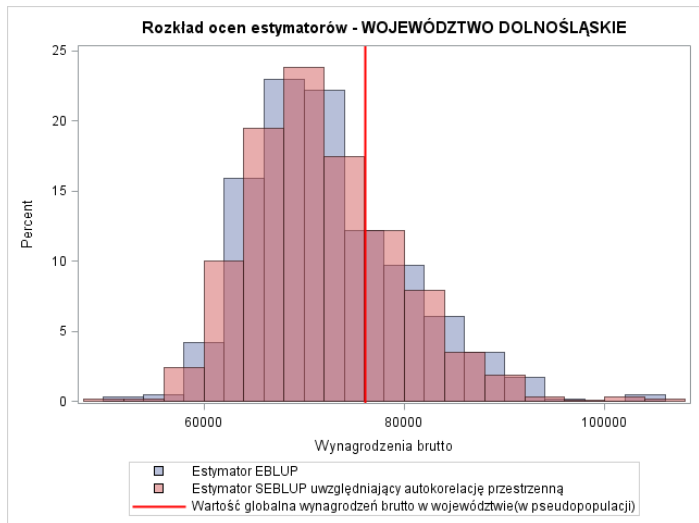
3. WYNIKI BADANIA SYMULACYJNEGO

sekcja F - BUDOWNICTWO



3. WYNIKI BADANIA SYMULACYJNEGO

sekcja F - BUDOWNICTWO





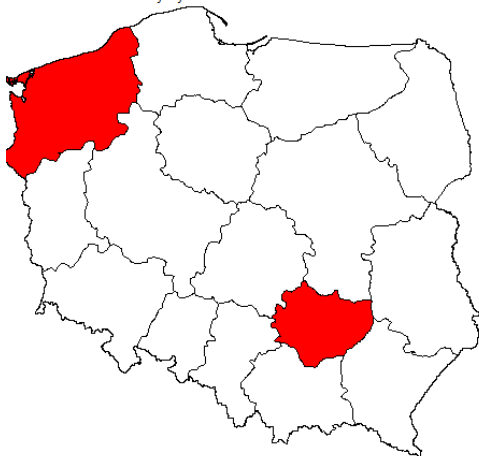
3. WYNIKI BADANIA SYMULACYJNEGO



TESTOWANIE AUTOKORELACJI LOKALNEJ – ISTOTNE STATYSTYKI I MORANA

Sekcja A - Rolnictwo

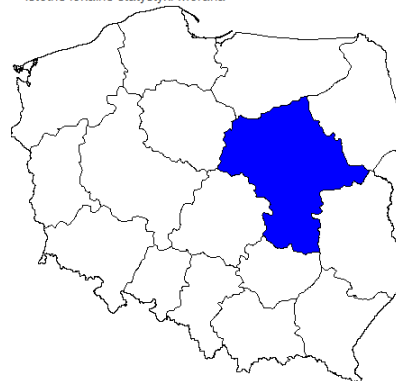
Istotne lokalne statystyki Morana



województwo	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z > 0)
Łódzkie	1.7975406	-0.4000000	5.899404	0.904758810	0.182796589
Świętokrzyskie	3.7154384	-0.4000000	5.899404	1.694384689	0.045096104
wielkopolskie	0.9295656	-0.4666667	6.861667	0.533019212	0.297010141
Kujawsko-Pomorskie	-1.9146685	-0.3333333	4.928253	-0.712323788	0.761867852
Małopolskie	1.8068308	-0.2000000	2.959283	1.166587986	0.121688399
Dolnośląskie	0.9279362	-0.2000000	2.959283	0.655679019	0.256015335
Lubelskie	1.0108032	-0.2666667	3.948212	0.642910360	0.260141133
Lubuskie	2.2949778	-0.2000000	2.959283	1.450352066	0.073480183
Mazowieckie	-1.3802417	-0.4000000	5.899404	-0.403579511	0.656739023
Opolskie	-0.9748424	-0.2666667	3.948212	-0.356402525	0.639230431
Podlaskie	-0.8443571	-0.2000000	2.959283	-0.374570298	0.646009967
Pomorskie	2.5254693	-0.2666667	3.948212	1.405194073	0.079981788
Śląskie	2.6600005	-0.2666667	3.948212	1.472899375	0.070389085
Podkarpackie	-0.3883864	-0.2000000	2.959283	-0.109510625	0.543601253
warmińsko-Mazurskie	-0.2821737	-0.2666667	3.948212	-0.007804193	0.503113391
Zachodniopomorskie	3.9972696	-0.2000000	2.959283	2.439908924	0.007345483

Sekcje:
C – Przemysł
F – Budownictwo
G – Handel

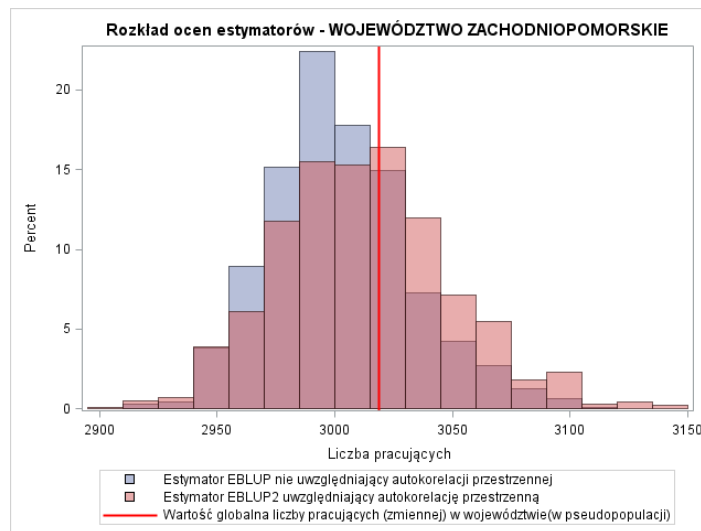
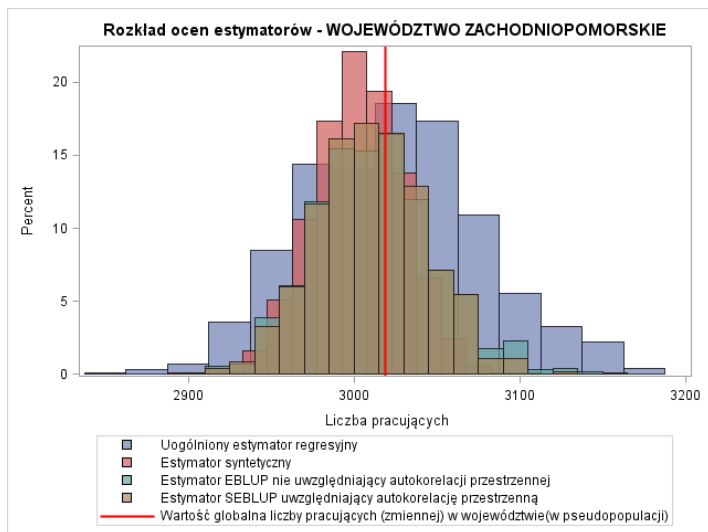
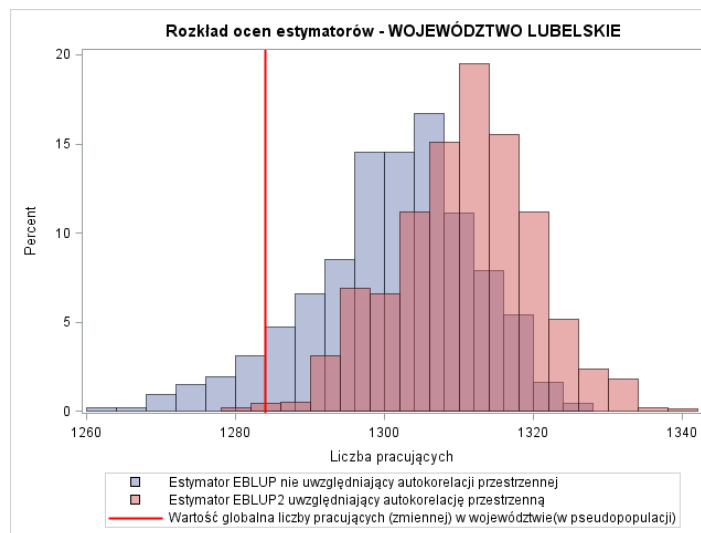
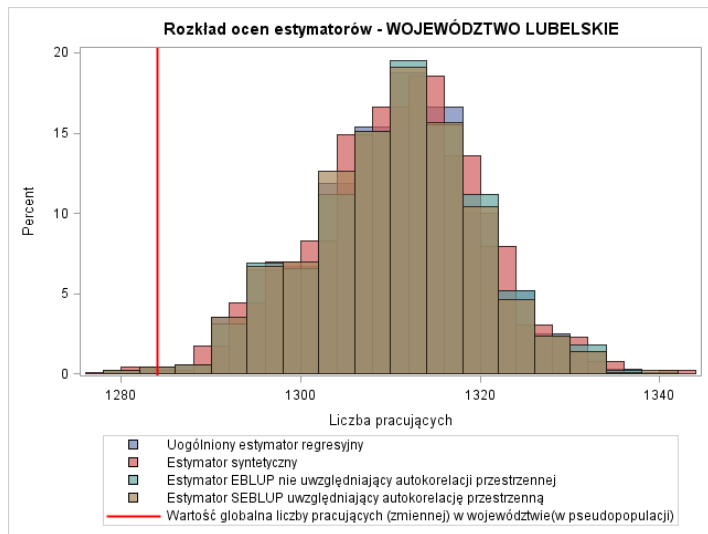
Istotne lokalne statystyki Morana



3. WYNIKI BADANIA SYMULACYJNEGO

PODEJŚCIE Z WYKORZYSTANIEM ISTOTNEJ AUTOKORELACJI LOKALNEJ

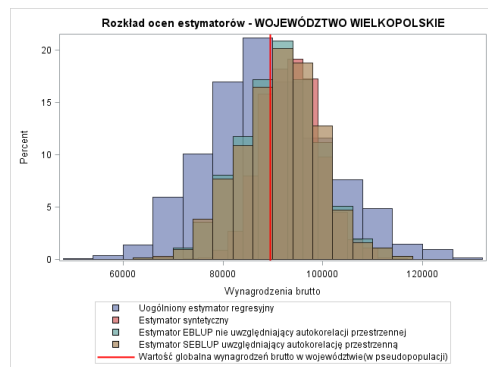
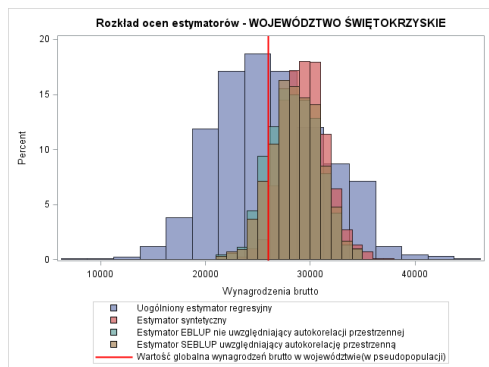
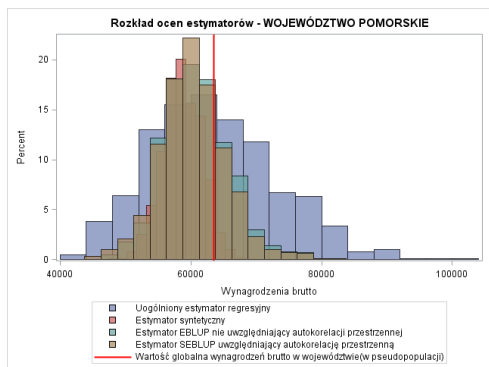
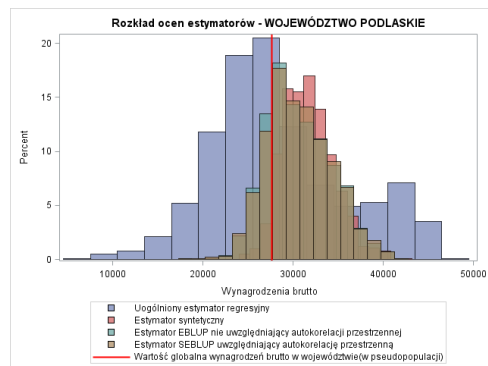
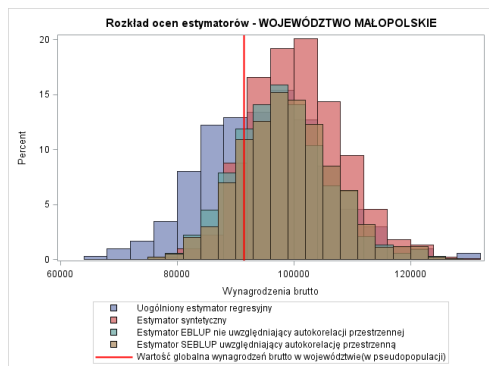
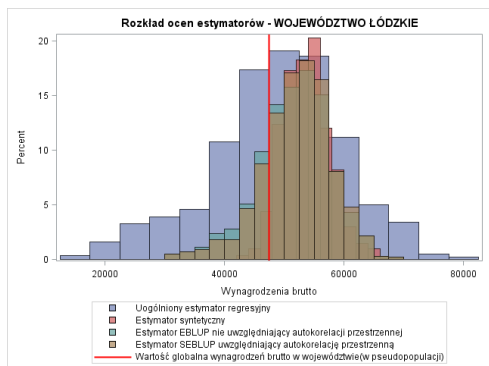
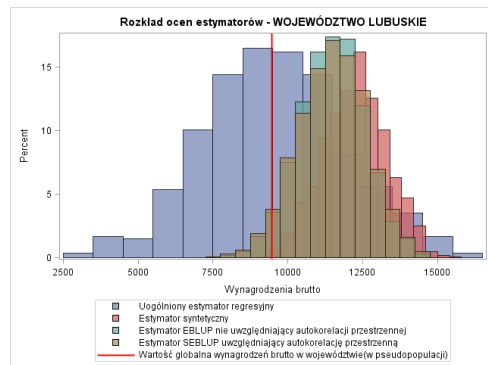
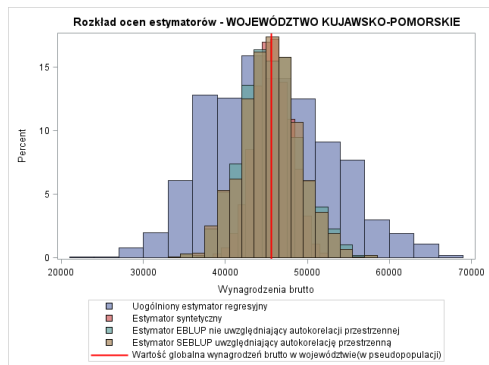
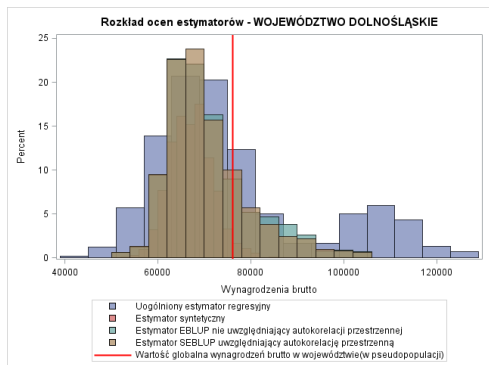
sekcja A - ROLNICTWO





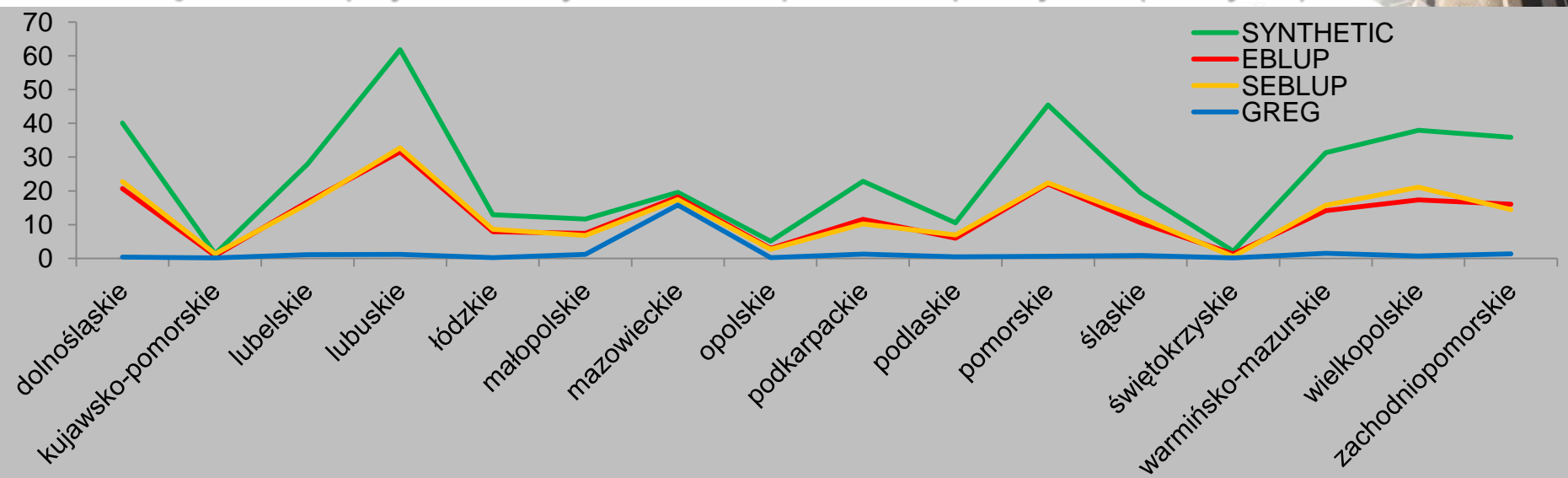
3. WYNIKI BADANIA SYMULACYJNEGO

sekcja F - BUDOWNICTWO

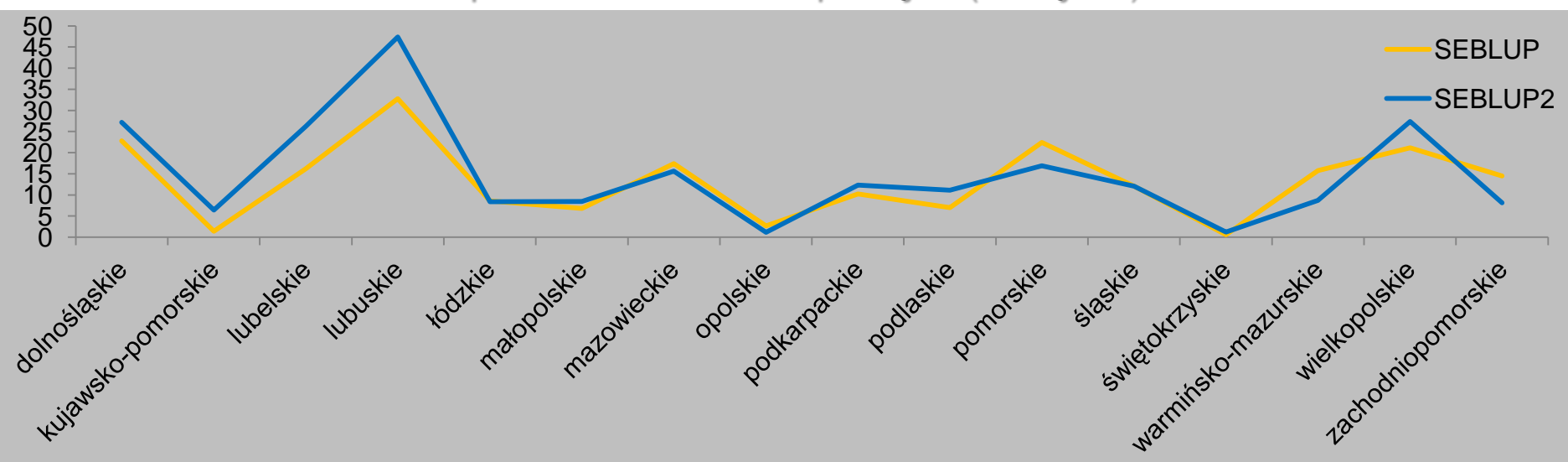




Obciążenie empiryczne estymatorów – pierwsze podejście (sekcja A)

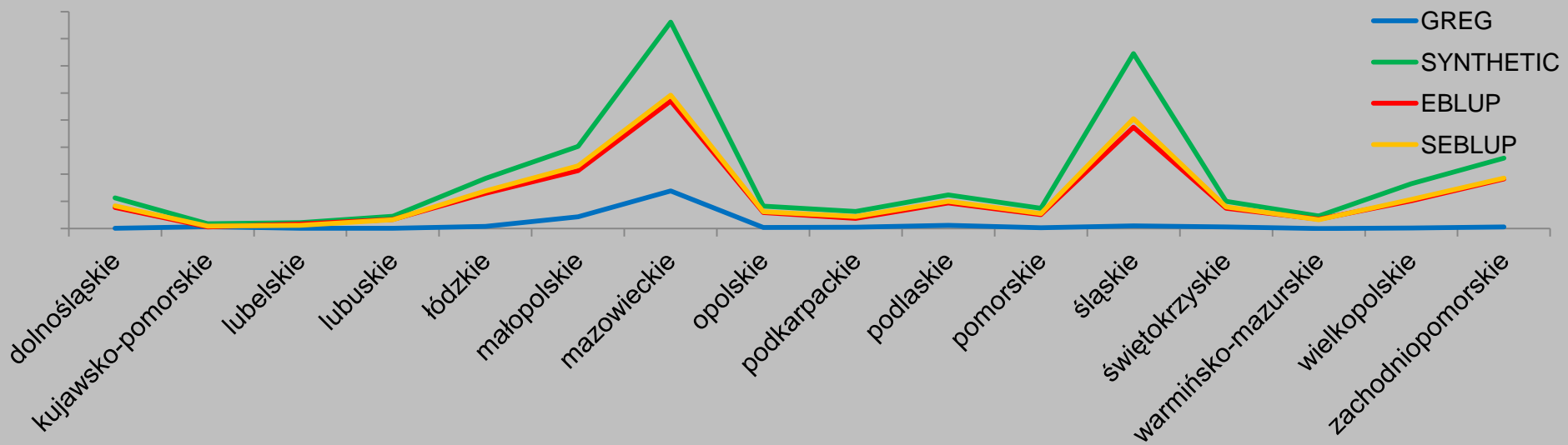


Obciążenie empiryczne estymatorów wykorzystujących korelację przestrzenną – porównanie dwóch podejść (sekcja A)

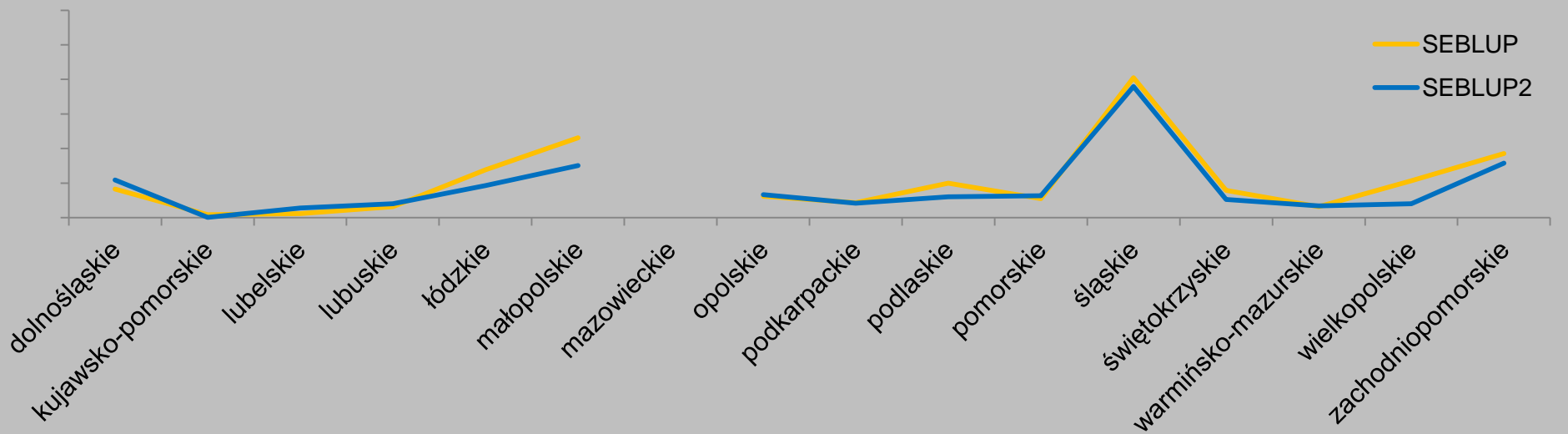




Obciążenie empiryczne estymatorów – pierwsze podejście (sekcja F)

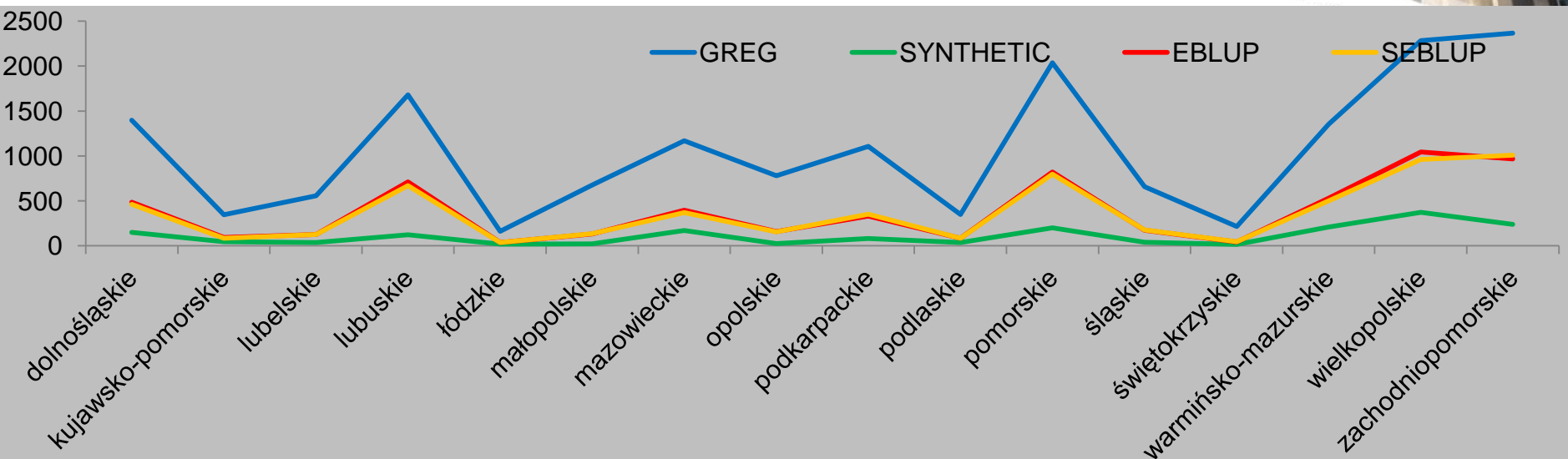


Obciążenie empiryczne estymatorów wykorzystujących korelację przestrzenną – porównanie dwóch podejść (sekcja F)

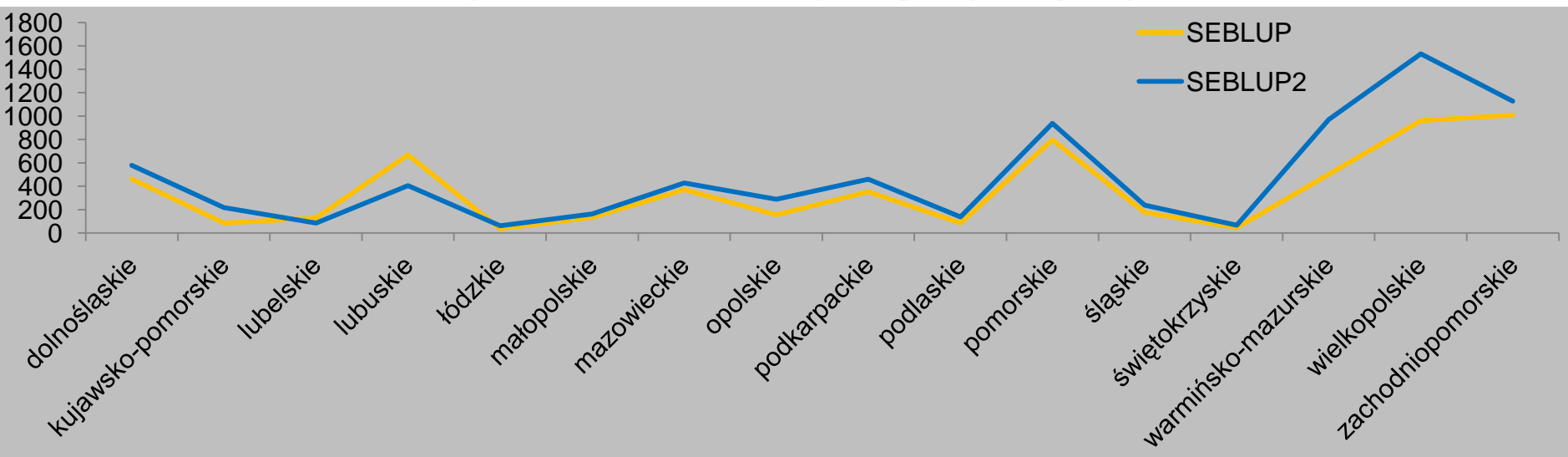




Wariancja empiryczna estymatorów – pierwsze podejście (sekcja A)

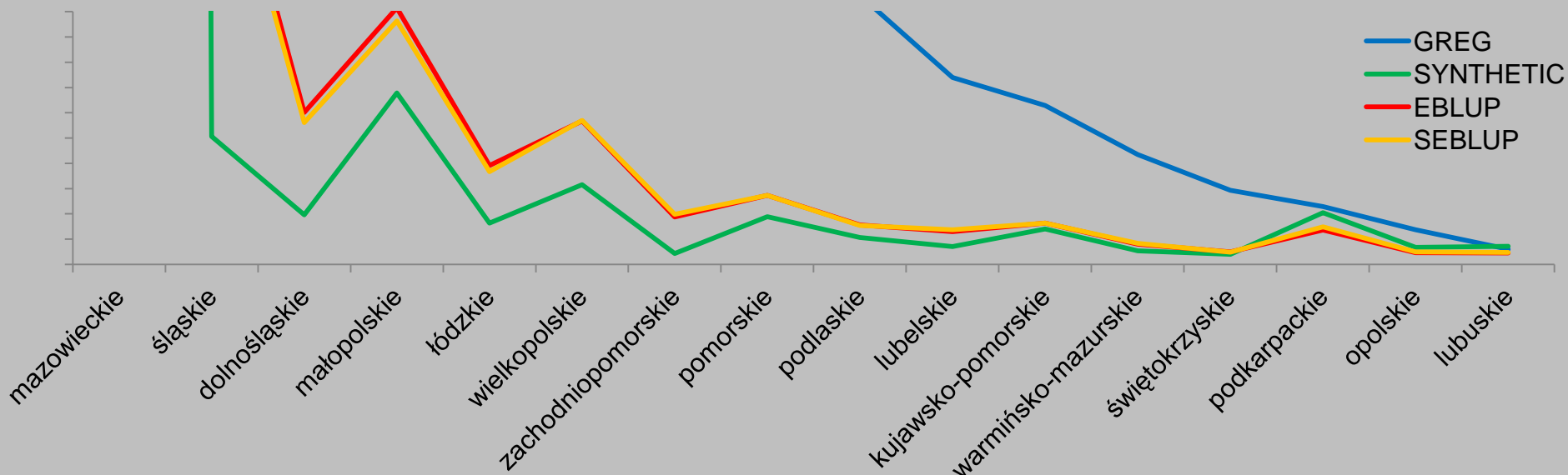


Wariancja empiryczna estymatorów wykorzystujących korelację przestrzenną – porównanie dwóch podejść (sekcja A)

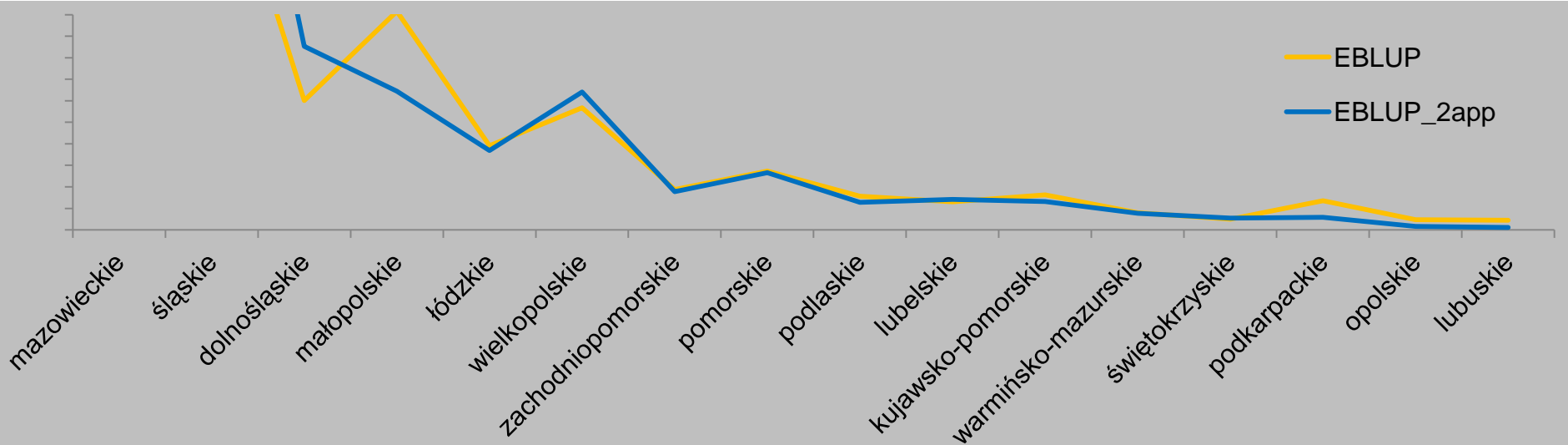




Wariancja empiryczna estymatorów – pierwsze podejście (sekcja F)



Wariancja empiryczna estymatorów wykorzystujących korelację przestrzenną – porównanie dwóch podejść (sekcja F)





4. WNIOSKI



1) Estymator bezpośredni, chociaż nieobciążony, charakteryzuje się trzema własnościami w przypadku estymacji dla małych domen:

- ma nieakceptowalną wariancję, a w związku z tym także błąd szacunku,
- jeżeli próba jest dostatecznie liczna, to własności estymatora bezpośredniego mogą być lepsze niż estymatorów opartych na modelach,
- w przypadku zerowej próby w domenie nie można wyznaczyć oceny estymatora.

2) Uogólniony estymator regresyjny, chociaż umożliwia uzyskanie oceny estymatora w przypadku zerowych prób, to jednak charakteryzuje się równie dużą wariancją co estymator ekspansyjny.

3) Estymatory syntetyczne i uwzględniające autokorelację przestrzenną charakteryzują się niewielką wariancją (w przypadku dobrze wyspecyfikowanego modelu). W porównaniu do estymatorów bezpośrednich, są one jednak obciążone.

4) Analiza przestrzennego rozkładu estymatora uwzględniającego autokorelację przestrzenną sugeruje, że może być on dobrym narzędziem do detekcji przestrzennych wzorców kształtowania zjawisk gospodarczych oraz ich estymacji.



Dziękujemy za uwagę