

**Statystyczna analiza
publicznych rejestrów wyborczych**

**Tomasz Przechlewski
Powiślańska Szkoła Wyższa
Kwidzyn, 12 października 2019**



Plan wykładu

Najpierw nieco teorii

A potem dużo obrazków

Art. 107. Ustawy z dnia 5 stycznia 2011r. Kodeks Wyborczy

§ 1. W dniu głosowania oraz na 24 godziny przed tym dniem prowadzenie agitacji wyborczej, w tym zwoływanie zgromadzeń, organizowanie pochodów i manifestacji, wygłaszaanie przemówień oraz rozpowszechnianie materiałów wyborczych jest zabronione.

Dane

Koń jaki jest każdy widzi: dane (treść) vs informacje (treść+forma) vs reklama (forma)

Jakość danych (wiarygodność+użyteczność+dostępność): kompletne | poprawne | niesprzeczne | relevantne nieulotne | trwałe pierwotne | nieprzetworzone | opisane | dostosowane (zgodne ze standardem) | machine processable (dokument vs zbiór danych)

Pradoks jakości: lepsze dane więcej kosztują, ale gorsze dane niekoniecznie kosztują mniej.

High quality data are accurate, available, complete, conformant, consistent, credible, processable, relevant and timely. <https://www.europeandataportal.eu/>

Dane publiczne

https://en.wikipedia.org/wiki/Open_data Open data is the idea that some data should be freely available to everyone to use and republish as they wish.

https://pl.wikipedia.org/wiki/Open_Government_Data udostępnienia informacji tworzonych przez organy administracji publicznej Słowo open oznacza możliwość (ponownego) wykorzystania danych bez ponoszenia kosztów

<http://www.oecd.org/gov/digital-government/open-government-data.htm> Open Government Data (OGD) is a philosophy – and increasingly a set of policies – that promotes transparency, accountability and value creation by making government **data** available to all.

By making their datasets available, public institutions become **more transparent** and accountable to citizens. By encouraging the use, reuse and free distribution of datasets, governments **promote business creation** and innovative, citizen-centric services.

Korzyści (obywatele vs władza/administracja): **zaangażowanie** (promocja/edukacja/społeczeństwo obywatelskie) | **przejrzystość działania/kontrola** | **konkurencyjność** (zmniejszenie kosztów/wyrównanie szans)

Państwowa Komisja Wyborcza

<https://wybory.gov.pl/sejmsenat2019/> <https://pkw.gov.pl/>

Aktualności, Akty Prawne, Kalendarz wyborczy, Komitety, Kandydaci, Wyszukiwarka obwodów, Organy wyborcze, Dane w arkuszach, Inne Wybory

https://parlament2015.pkw.gov.pl/347_Wyniki

<https://wybory2018.pkw.gov.pl/pl/dane-w-arkuszach>

Wyniki głosowania na listy kandydatów do sejmików województw | Wyniki głosowania i wyniki wyborów wójtów, burmistrzów i prezydentów miast | Wyniki wyborów do rad

Frekwencja i głosy nieważne w obwodach głosowania

Obwody głosowania | Okręgi wyborcze | Kandydaci do rad | Kandydaci na wójta, burmistrza, prezydenta | Komitety wyborcze | Skład terytorialnych i obwodowych komisji wyborczych

Zawartość: rejesty wszystkiego: komitetów wyborczych, kandydatów, komisji wyborczych. Wyniki wyborów i głosowania (frekwencja, głosy nieważne) **Format:** dokument. **Dane zagregowane wg różnych kryteriów (list, kandydatów, jednostek podziału administracyjnego (gmina, województwo) oraz okręgów wyborczych.)**

Dane źródłowe/pierwotne: protokoły z komisji obowodowych. Format: dokument.

Nieco teorii nt wyborów i standardów

PKW organizuje wybory współpracując z innymi organami administracji i korzystając z utrzymywanej przez nie infrastruktury.

PESEL: powszechnie znany numer ewidencyjny, ale także rejestr służy do gromadzenia informacji identyfikujących obywateli polskich oraz cudzoziemców zamieszkujących na terenie Rzeczypospolitej Polskiej.

TERYT: standard identyfikacji terytorialnej Województwa (16), powiaty (314 + 66 miast na prawie powiatu = 380), gminy (2477). Identyfikator TERYT: [wwppggt](#)

OKRĘGI i OBWODY wyborcze: wybory parlamentarne – 41 okręgów wyborczych. Wybory samorządowe do rad sejmików wojewódzkich – 85 okręgów wyborczych.

Około 27 tys obwodowych komisji wyborczych (26983 w wyborach samorządowych 2018)

100000;181405;Szkoła Podstawowa w Kańczudze, Kańczuga, ul. Szkolna 7, 37-220 Kańczuga

100002;181406;Zespół Szkół w Chałupkach, Chałupki 109, 37-200 Przeworsk

albo:

181405;8;Szkoła Podstawowa w Kańczudze, Kańczuga, ul. Szkolna 7, 37-220 Kańczuga

181405;9;Hala Sportowa w Kańczudze, Kańczuga, ul. Parkowa 10, 37-220 Kańczuga

181406;1;Zespół Szkół w Chałupkach, Chałupki 109, 37-200 Przeworsk

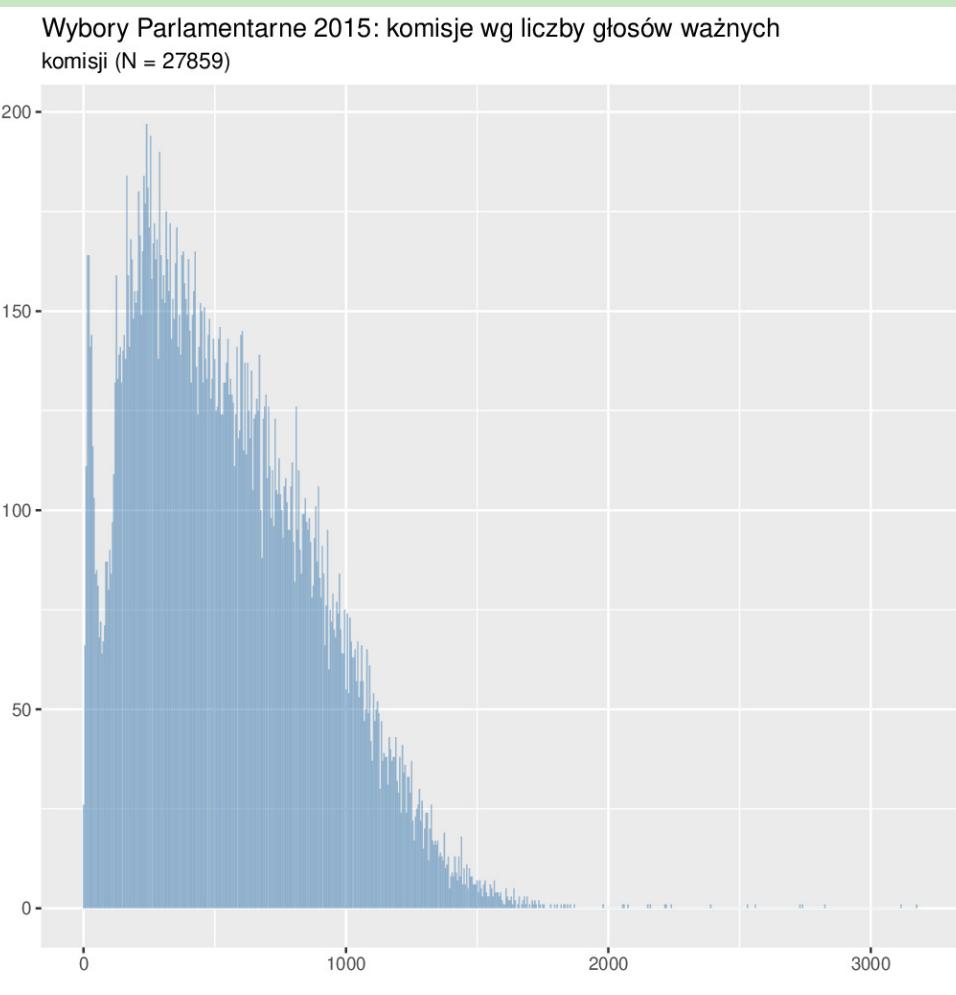
Liczba mandatów w okręgach wyborczych



Herfindahl-Hirschman Index (HH Index): prosta miara koncentracji definiowana jako suma kwadratów udziałów wyrażonych w % (czyli $HH = \sum_{i=1}^n x_i^2$, gdzie $\sum_{i=1}^n = 100$)

Wybory do Sejmu RP 2015

Ze strony http://parlament2015.pkw.gov.pl/349_Wyniki_Sejm za pomocą pobrań *web scrappingu* zostały pobrane 28.10.2015 protokoły 27,859 komisji obwodowych.



Wynikowy plik CSV liczył: 5,730,125 wpisów postaci:

```
idkomisji;nrlisy;numer;kto;komitet;glosy;teryt;woj  
100000;1;1;Marek Tadeusz KUCHCIŃSKI;PiS;71;181405;18  
100000;1;2;Anna Katarzyna SCHMIDT-RODZIEWICZ;PiS;78;181405;18  
100000;1;3;Piotr Marcin URUSKI;PiS;3;181405;18
```

Minimalna liczba głosów – 0, pierwszy kwartyl – 265, mediana – 494, trzeci kwartyl – 782 (czyli w 50% komisji oddano około 500 głosów). Wartość maksymalna 3,177. Średnia 545,3 głosów.

W 577 komisji oddało głosy mniej niż 25 wyborców (2,1%)

Wybory do Sejmu RP 2015

Porównano wyniki uzyskane w ten sposób z wynikami publikowanymi na stronie PKW:

Komitet	L.kandydatów	L.głosów	L.głosów*	Różnica	Okr19	Okr19*	Różnica
KORWIN	899	722921	722999	78	21757	21767	10
KUKIZ	839	1338610	1339094	484	26546	26573	27
Kongres N. Prawicy	116	4852	4852	0	x	x	x
Razem	571	550343	550349	6	9469	9475	6
Samoobrona	119	4266	4266	0	x	x	x
BRAUN	202	13113	13113	0	x	x	x
JOW Bezpartyjni	138	15184	15656	472	x	x	x
Mniejszość Niem.	24	27530	27530	0	x	x	x
OdP	40	1964	1964	0	266	266	0
Ruch Społ. RP	59	3941	3941	0	186	186	0
STONOGA	299	42668	42731	63	x	x	x
Zjed. dla Śląska	42	18668	18668	0	x	x	x
PETRU	858	1155364	1155370	6	15942	15948	6
PiS	918	5711661	5711687	26	58317	58343	26
PO	914	3661455	3661474	19	32240	32259	19
PSL	916	779874	779875	1	796	797	1
ZLEW	905	1146837	1147102	265	7948	7956	8

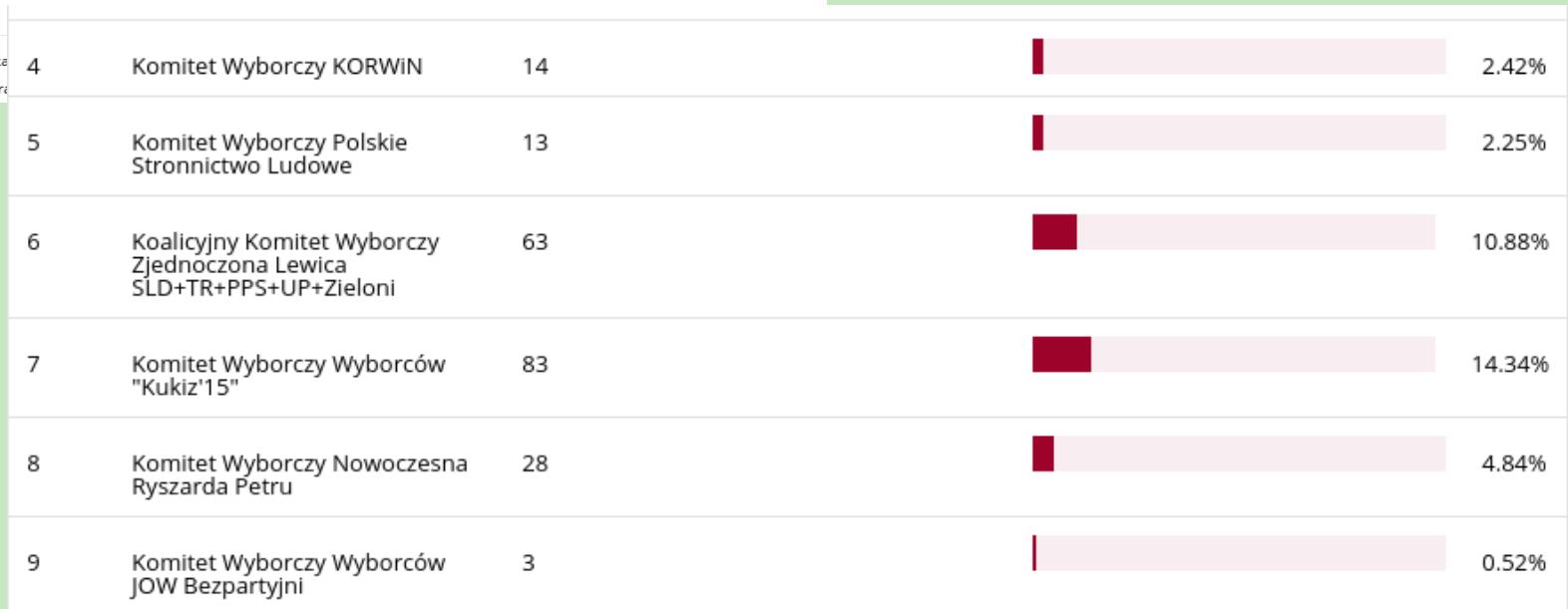
* dane zagregowane ze strony PKW

Jedna komisja jest/była tajna (Bagram/Afganistan /okr. 19)

http://parlament2015.pkw.gov.pl/321_protokol_komisji_obwodowej/97770

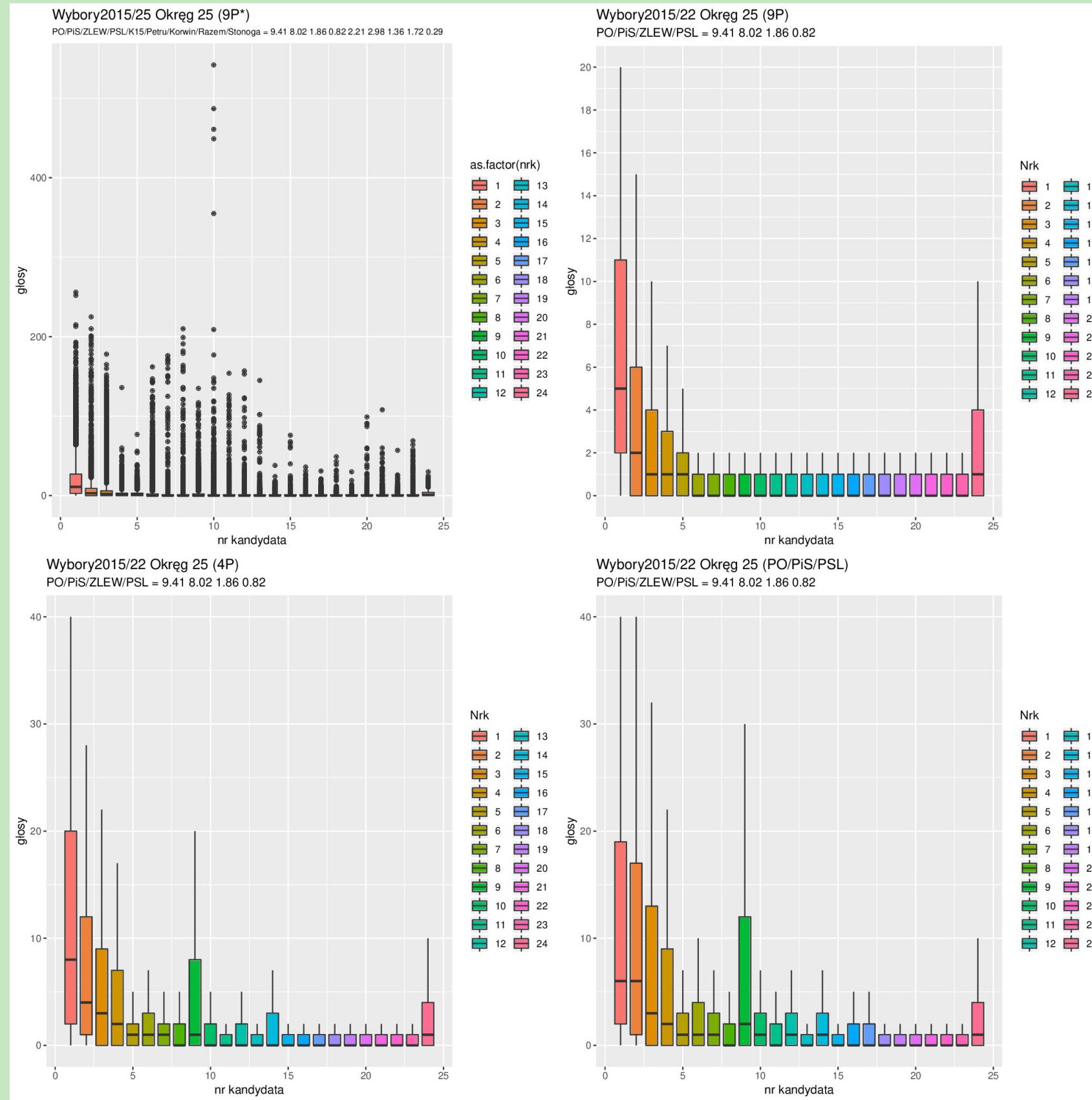
Przypadek komisja obwodowej w Bardzie na ul. Kolejowej 12

parlament2015.pkw.gov.pl/321_protokol_komisji_obwodowej/84873/						
ost Visited Powiadomienia Znalezione obrazy d... gMapToGPX Szukaj znajomych Wybory 2011 - Wyni... bangor - Szukaj w G... Louise Brooks & Rin...						
8	14	Elzbieta BRZĘCZEK	KW Nowoczesna Ryszarda Petru	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
8	15	Tomasz ŁOJEWSKI	KW Nowoczesna Ryszarda Petru	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
8	16	Dariusz Leopold GÓRALEWICZ	KW Nowoczesna Ryszarda Petru	1	<div style="width: 25%; background-color: #e67e22;"></div>	0.17%
9	1	Patryk Edward WILD	KWW JOW Bezpartyjni	XXXXX	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	2	Pawel OZGA	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	3	Mariusz Jarosław GRYŚ	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	4	Dorota Urszula SOZAŃSKA	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	5	Ewa OZIMEK	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	6	Iwona Stanisława STACH-JANYST	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	7	Danuta Zofia PEŁKA	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	8	Robert Wiesław WĄS	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	9	Edmund Zbigniew FRĄCZAK	KWW JOW Bezpartyjni	XXXXX	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	10	Małgorzata JASIAK-GAZDA	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	11	Katarzyna KOCIK	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%
9	12	Zygmunt WORSA	KWW JOW Bezpartyjni	0	<div style="width: 0%; background-color: #e67e22;"></div>	0%

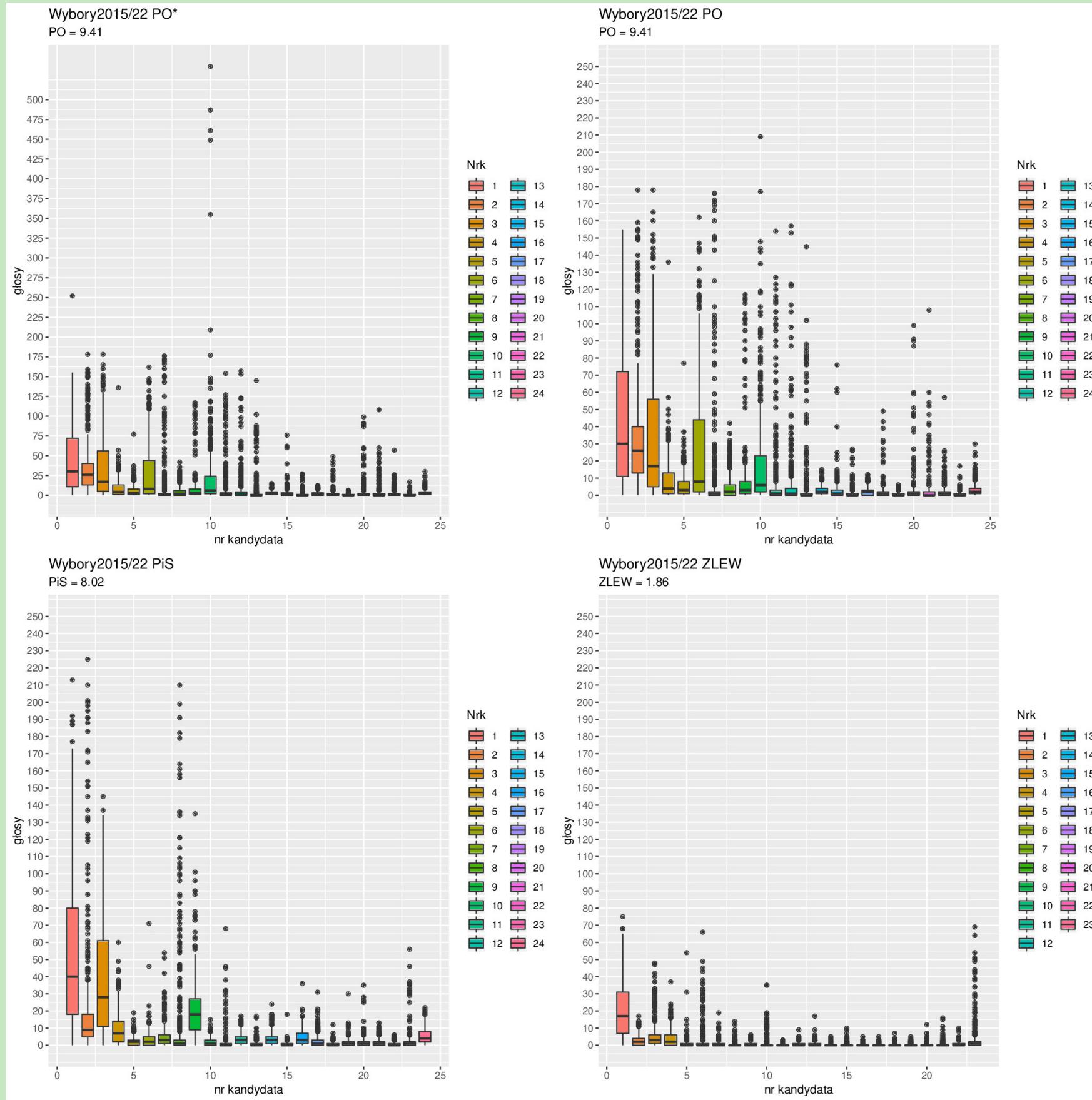


Wybory do Sejmu RP 2015 (Okręg 25)

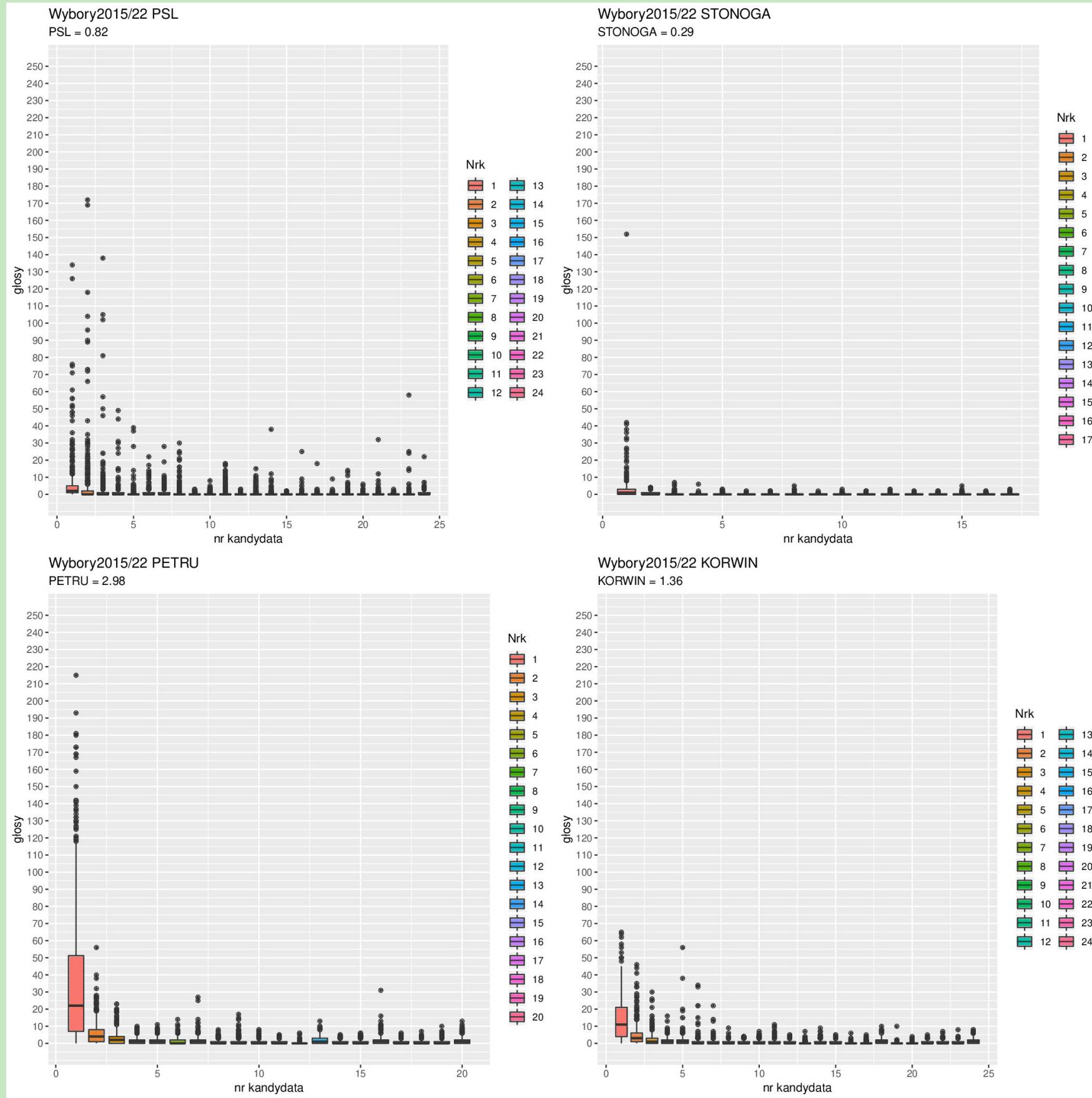
Trzy rodzaje analizy statystycznej: konfirmacyjna, eksploracyjna, numerologiczna...



Okręg 25 (województwo pomorskie)

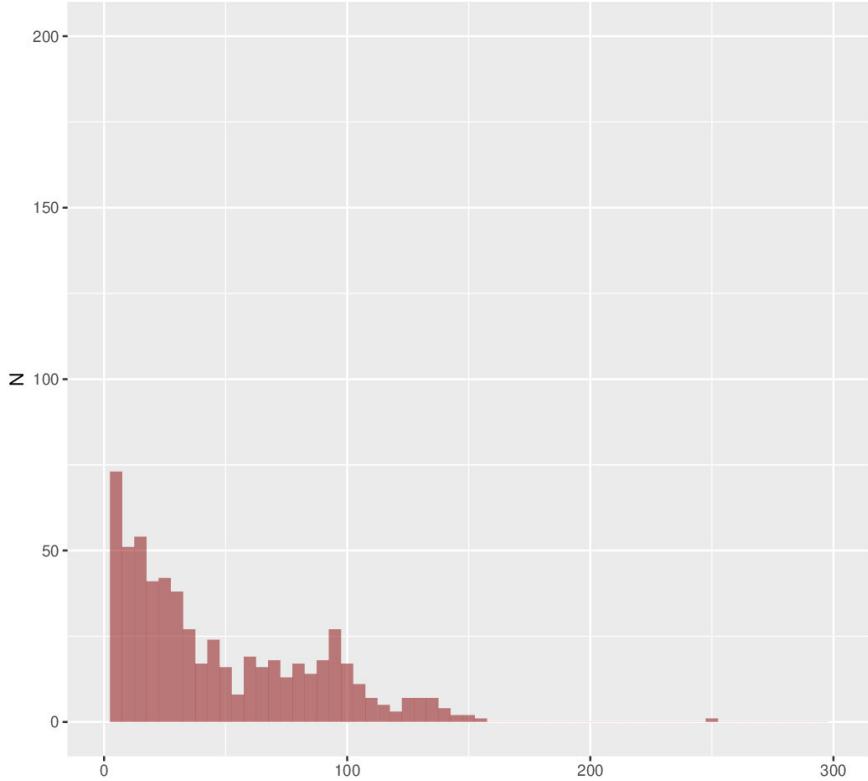


Okręg 25 (województwo pomorskie)

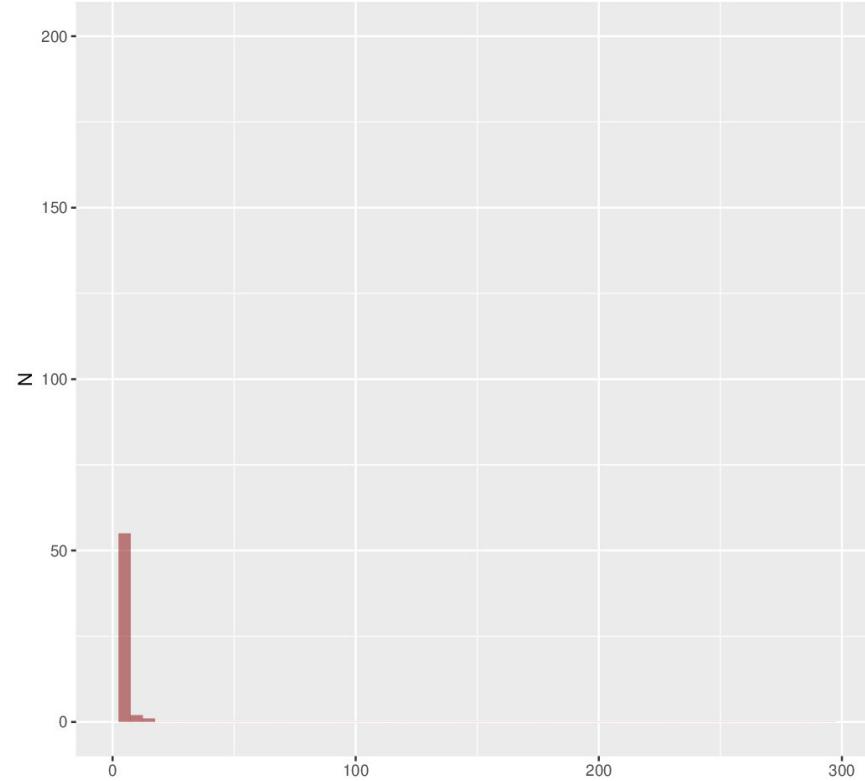


Okręg 25 (województwo pomorskie)

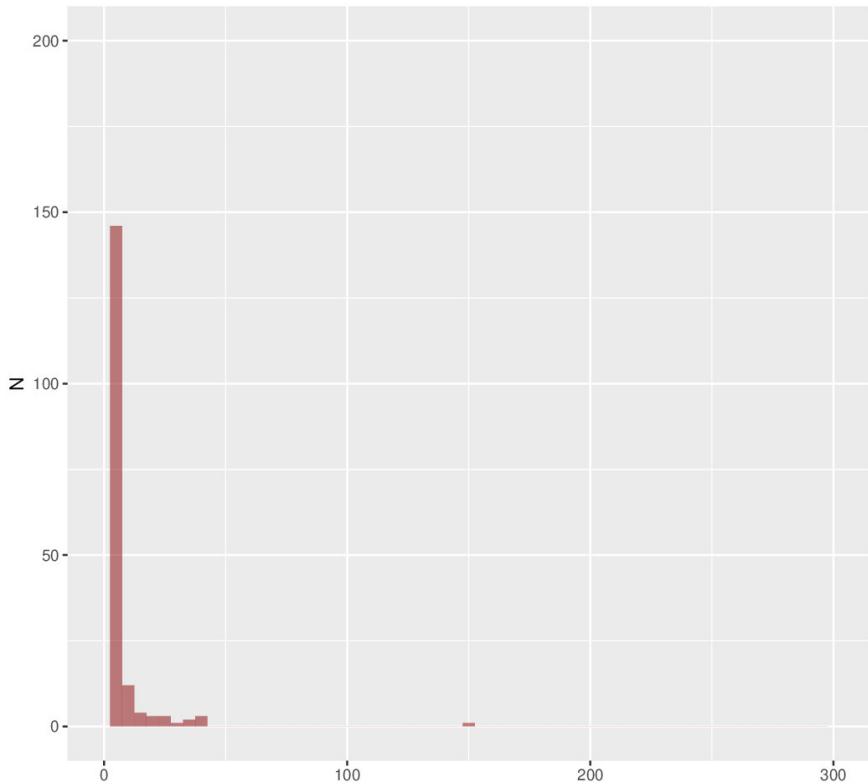
Wybory 2015: Adam Marek KOROL/28727 (1) PO
Komisji: 657 / 0-głosów: 17/2.6 / 2-głosy: 624 10-głosów: 515



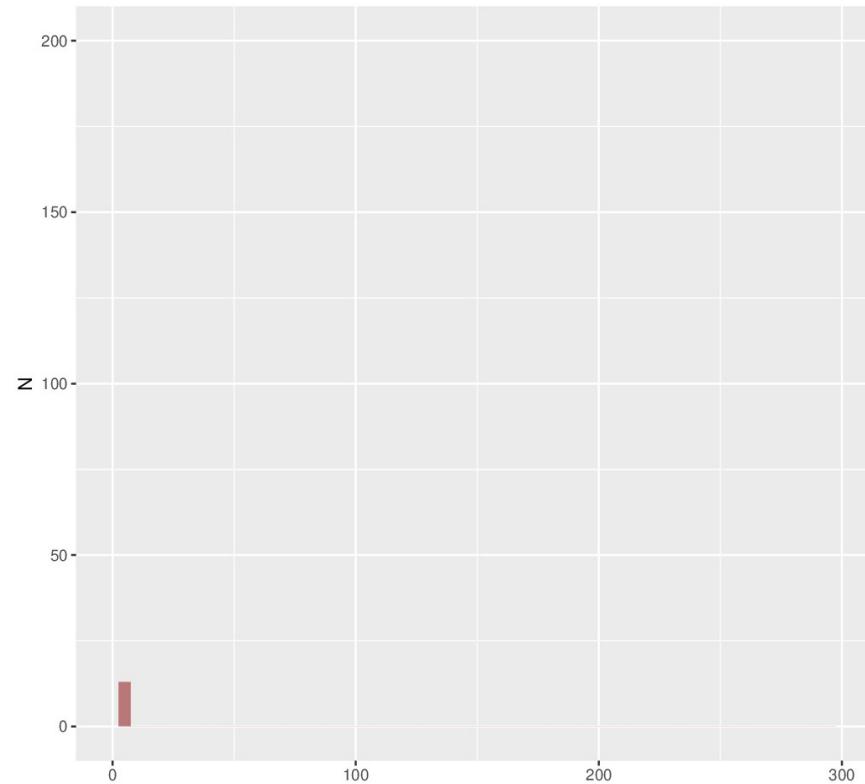
Wybory 2015: Radosław Krzysztof SOBECKI/569 (23) PO
Komisji: 657 / 0-głosów: 347/52.8 / 2-głosy: 140 10-głosów: 3



Wybory 2015: Danuta HOJARSKA/1671 (1) STONOGA
Komisji: 657 / 0-głosów: 191/29.1 / 2-głosy: 300 10-głosów: 22



Wybory 2015: Piotr WASIEK/191 (3) STONOGA
Komisji: 657 / 0-głosów: 523/79.6 / 2-głosy: 35 10-głosów: 0



Wybory do Sejmu RP 2015

Poznanie zjawiska to także sprawdzenie (że wszystko jest OK).

Szukanie anomalii (rozkładu lub nietypowych wartości)

idkomisji;komitet%;lgw;adres

102943;KORWIN;50.00;14;MONAR Ośrodek Leczenia..., Gdańsk, ul. Agrarna 2, 80-298 Gdańsk 16
102947;KORWIN;48.17;328;Dom Studencki Nr 8, Gdańsk, ul. Wyspiańskiego 5A, 80-434 Gdańsk 6
102231;STONOGA;36.93;417;Świetlica Wiejska w Lubieszewie, Lubieszewo, 16, 82-100 Nowy D. Gdańsk
108763;STONOGA;16.67;6;Szpital Powiatowy, Krotoszyn, ul. Boleskiego 4-8, 63-700 Krotoszyn
102562;16.67;6;Dom Pomocy Społecznej, Stanisławie, 31, 83-112 Lubiszewo
102259;33.33;3;Szpital Pucki Sp. z o.o., Puck, 1-go Maja 13A, 84-100 Puck

5% komisji \leq 67 głosów ; 10% komisji \leq 137 głosów; 0 głosów – 2 komisje; 5 głosów i mniej – 48 komisji; 1500 i więcej – 164 komisje.



Geokoder Google

27859 adresów komisji obwodowych do zamiany na współrzędne:

100000;181405;Szkoła Podstawowa w Kańczudze|Kańczuga|ul. Szkolna 7|37-220 Kańczuga

W wersji darmowej limit 2500 wyszukań/dzień.

Wyszukanie po kodach pocztowych:

```
awk -F'|' '{print $NF}' kody_adresy.csv | sort -u | wc -l  
$ 6313
```

Google znajduje 5698 różnych współrzędnych. Błąd po stronie PKW (także?)

```
awk -F'|' '{a[$NF]++; END{ for (x in a) {if (a[x]> 30) { print a[x] " = " x }}}}' kody_adresy.csv  
$ 76      ## a[x]< 10 => 5714
```

69 = 44-200 Rybnik; 80 = 43-300 Bielsko-Biała; 116 = 26-600 Radom; 124 = 87-100 Toruń...

Wyszukanie „z kotwicą” (bounds)

```
https://maps.googleapis.com/maps/api/geocode/json?\n  bounds=53.8710742,16.7827421|54.4431018,17.8002834\n  &address=Borzytuchom, ul. Czterdziestolecia 2&components=country:pl
```

2574 adresów w województwie pomorskim i zachodnio-pomorskim; 2302 adresów unikatowych; 1858 (80,71%) geokodowań z dokładnością ROOFTOP; 210 (9,12%) geokodowań z „dokładnością” Poland.

```
.../api/geocode/json?&components=country:PL&address=Sopot,%20ul.%201-go%20Maja%205  
=>%22formatted_address%22%3A%22Poland%22  
.../api/geocode/json?&components=country:PL&address=Sopot,%20ul.%201%20Maja%205  
=>%22formatted_address%22%3A%221%20Maja%205,%20Sopot,%20Poland%22
```

```
.../api/geocode/json?&components=country:PL|postal_code:81-825  
.../api/geocode/json?&components=country:PL|administrative_area:kartuski&address=Kartuzy
```

Jak często powtarzają się nazwy miejscowości w Polsce

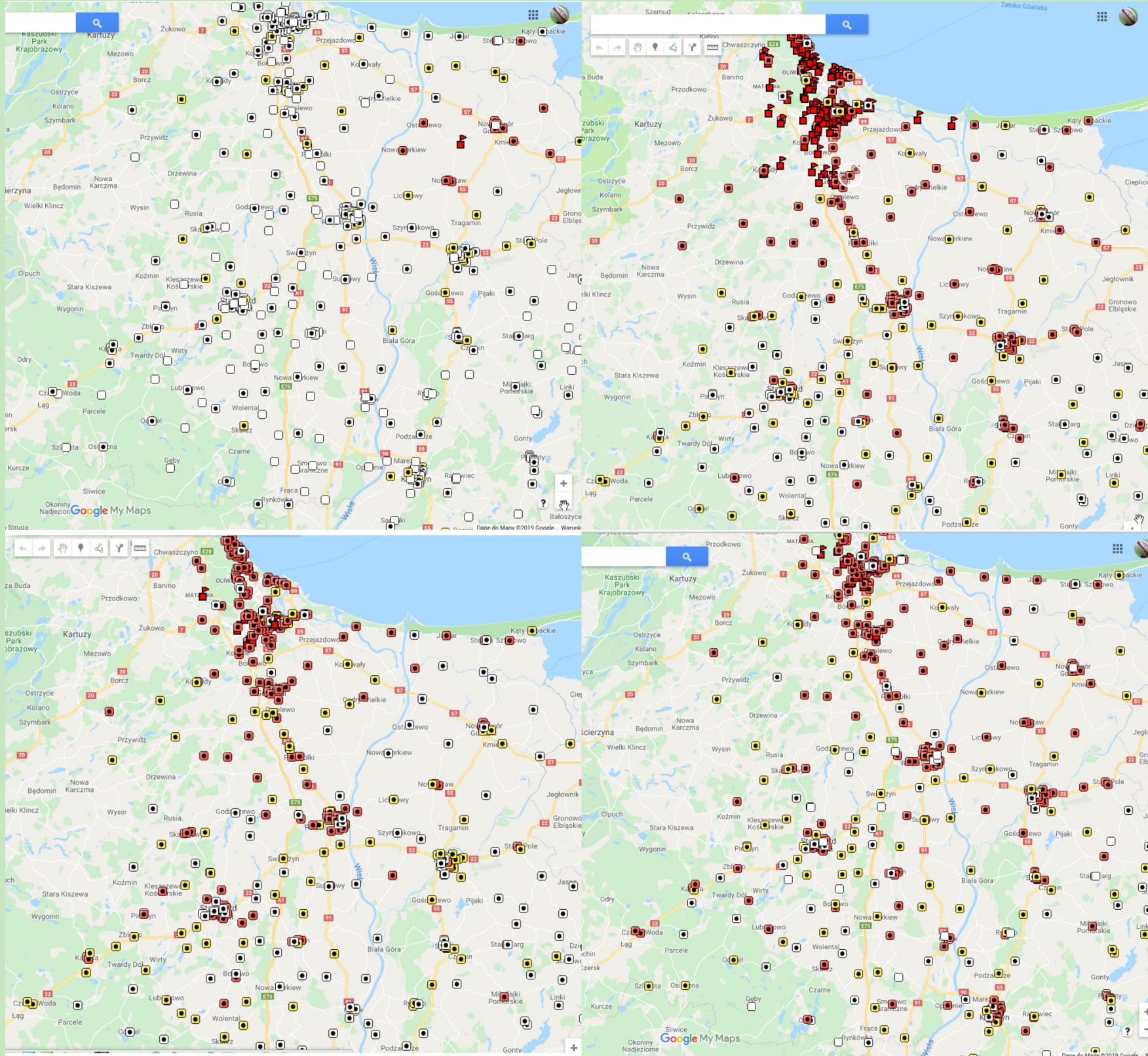
Według obwieszczenia Ministra Administracji i Cyfryzacji z 4/8/2015 r. w wykazie urzędowych nazw miejscowości i ich części (DU 19/10/2015 r., poz. 1636) znajduje się: 103086 nazw, w tym 915 nazw miast, 6710 nazw części miast, 43068 nazw wsi, 36263 nazwy części wsi, 5132 nazwy osad.

Wsi + miast jest $43068 + 915 = 43,983$ nazw.

Województwo	[1,2)	[2,3)	[3,5)	[5,30)	[1,2)	[2,3)	[3,5)	[5,30)
dolnośląskie	2082	117	37	3	92,99	5,23	1,65	0,13
kuj.-pomorskie	2257	183	52	10	90,20	7,30	2,10	0,40
lubelskie	2698	154	65	26	91,68	5,23	2,21	0,88
lubuskie	965	57	5	0	93,96	5,55	0,49	0,00
łódzkie	3276	275	99	45	88,70	7,40	2,70	1,20
małopolskie	1594	105	21	4	92,46	6,09	1,22	0,23
mazowieckie	5780	466	177	79	88,90	7,20	2,70	1,20
opolskie	941	39	9	1	95,05	3,94	0,91	0,10
podkarpackie	1429	49	17	2	95,46	3,27	1,14	0,13
podlaskie	2852	136	42	9	93,80	4,50	1,40	0,30
pomorskie	1579	58	13	2	95,58	3,51	0,79	0,12
śląskie	1033	48	7	2	94,77	4,40	0,64	0,18
świętokrzyskie	1844	111	51	12	91,38	5,50	2,53	0,59
war.-mazurskie	2060	128	34	3	92,58	5,75	1,53	0,13
wielkopolskie	3378	310	99	19	88,80	8,10	2,60	0,50
zach-pomorskie	1504	117	19	1	91,65	7,13	1,15	0,06

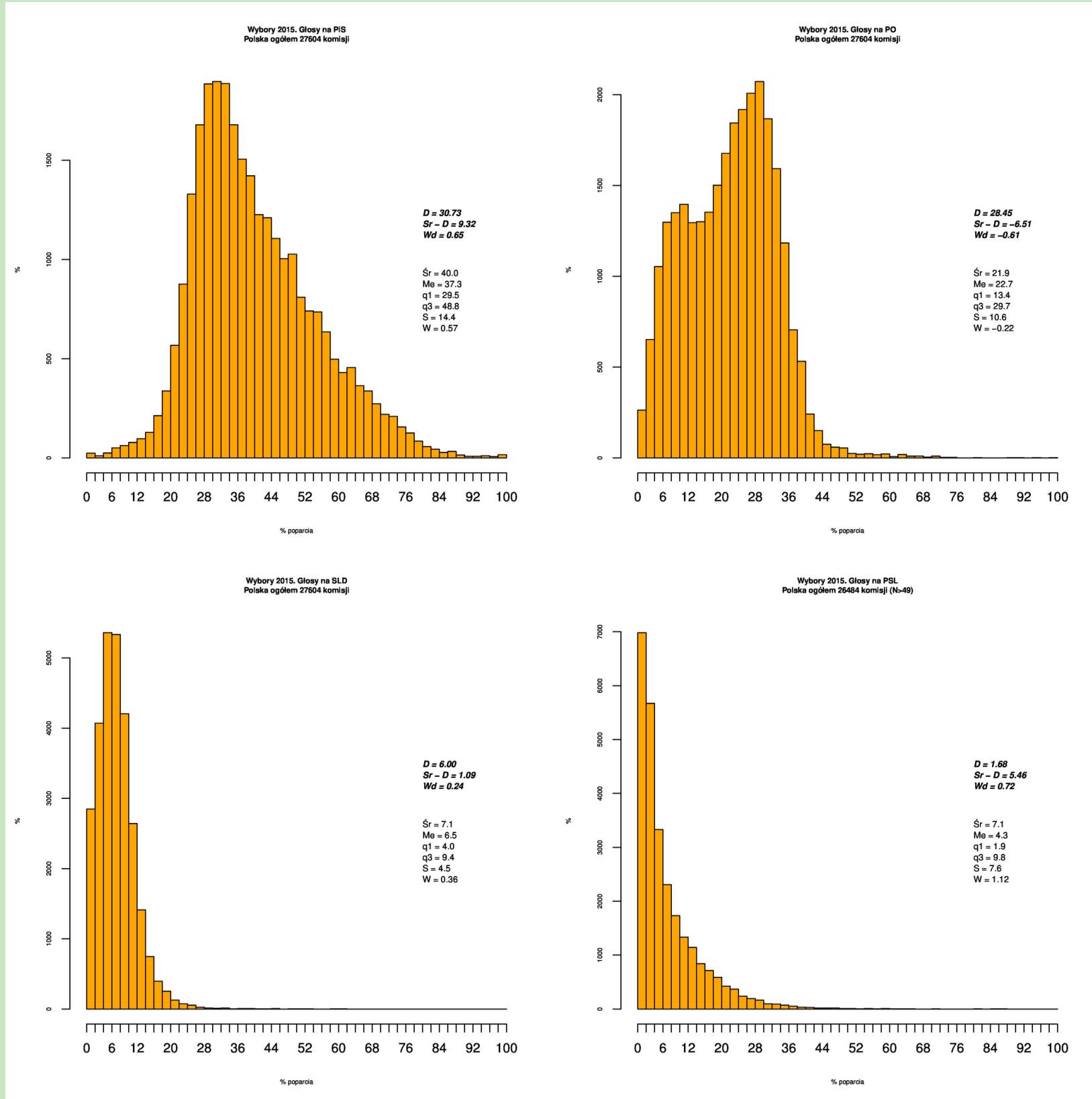
W skali całego kraju jest 5080 nazw nieunikatowych;, w przypadku 868 miejscowości nazwy powtarzają się 5 i więcej razy. Pierwsza dziesiątka: **Nowa Wieś** (113 powtórzeń), Dąbrowa (92), Zalesie (86), Dąbrówka (65), Józefów (49), Kamionka (41), Ruda (39), Janów (39), Zawady (39) Stanisławów (38).

Okręg 25 (Hojarska/Pomaska/Rabenda/Senyszyn)



Wybory do Sejmu RP 2015

W poszukiwaniu anomalii rozkładu (pytanie do przyjaciela: kto wygrał?)



Nihil novi sub sole

The Wrong Man is President! Overvotes in the 2000 Presidential Election

Walter R. Mebane Jr.

Using ballot-level data from the NORC Florida ballots project and ballot-image files, I argue that overvoted ballots in the 2000 presidential election in Florida included more than 50,000 votes that were intended to go to either Bush or Gore but instead were discarded. This was primarily due to defective election administration in the state, especially the failure to use a system to warn each voter when too many marks were on a ballot and allow the voter to make corrections. If the best type of vote tabulation system used in Florida in 2000—precinct-tabulated optical scan ballots—had been used everywhere in the state, Gore would have won by more than 30,000 votes. Florida's election experience points to the need to gather ballot-level data to evaluate the success of election reform efforts now underway in much of the United States.

Skandal wyborczy w USA (2000) znany jako google: Bush vs Gore

W stanie Floryda Bush wygrał 537 głosami ale 50 tysięcy głosów uznano za nieważne albo z powodu niewskazania żadnego kandydata (pusty głos) albo wskazania więcej niż jednego

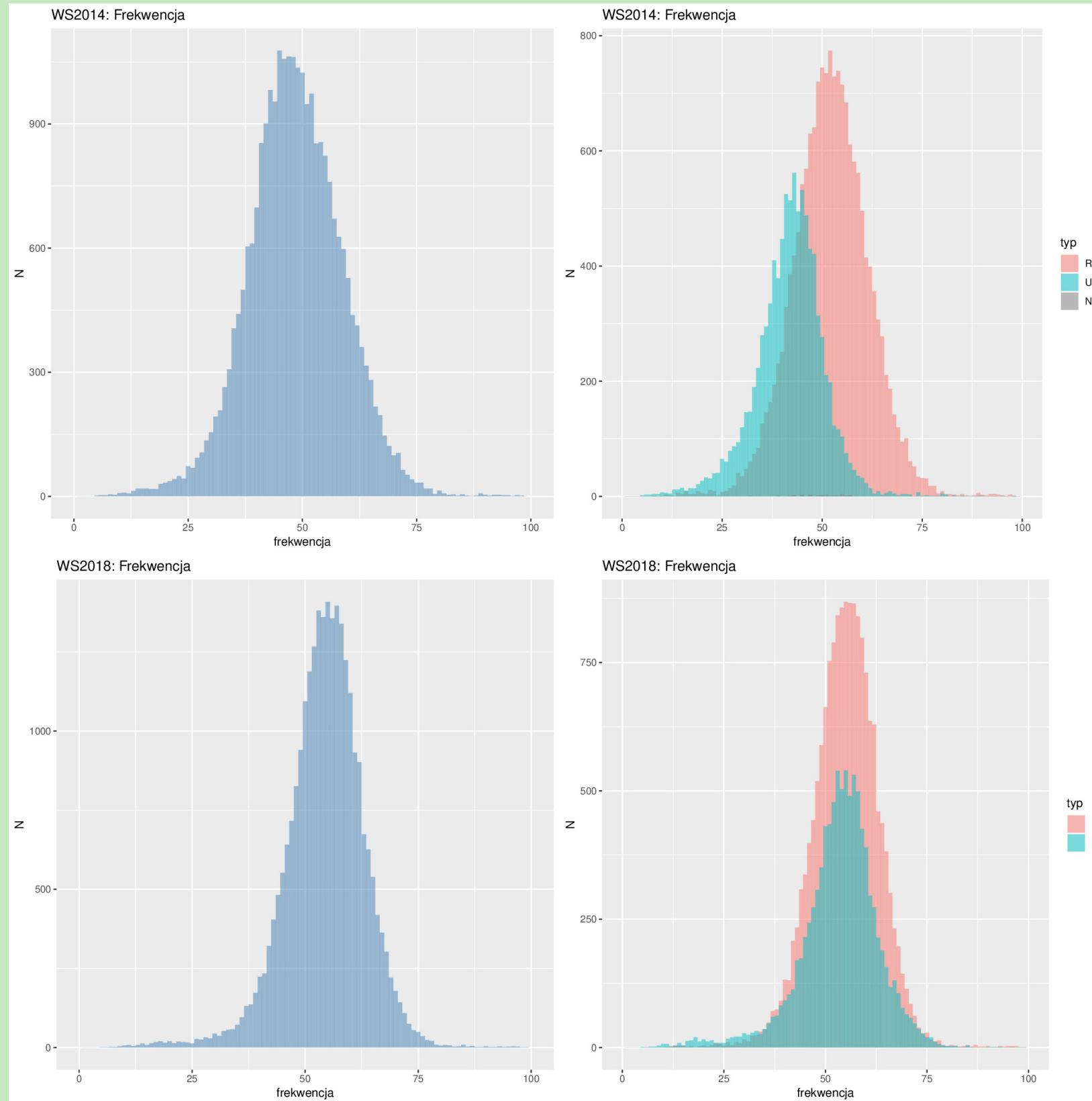
Organizacja wyborów w USA jest mocno zdecentralizowana. Co do zasady głos jest oddawany z wykorzystaniem maszyn. Maszyny mogą wspomagać wyborcę lub nie.

Porównując wyniki z obwodów korzystających z lepszych maszyn z tymi używającymi gorszych niektórzy oceniają że Gore wygrałby 30 tys głosów gdyby używano tylko lepszych.

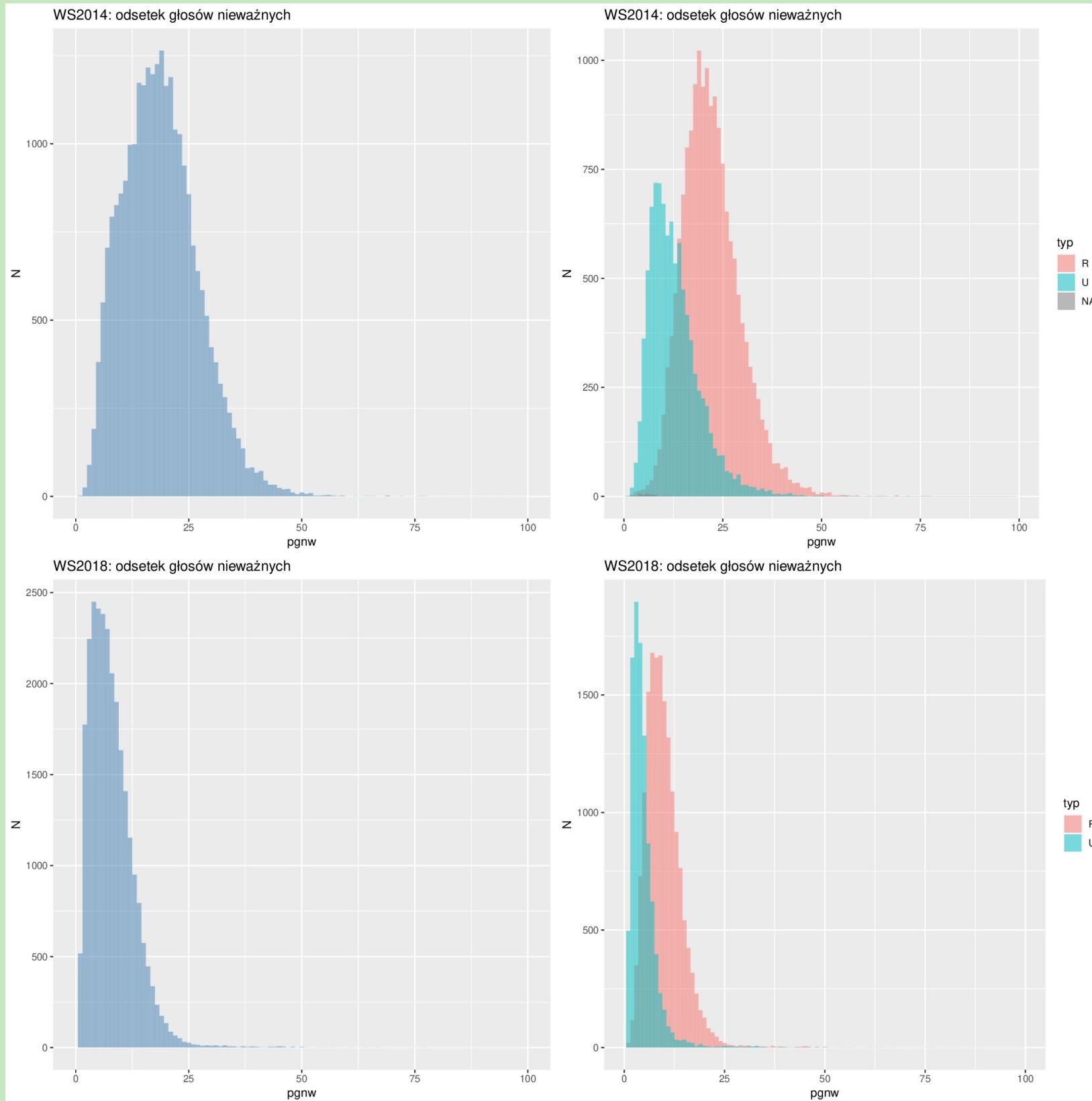
Podstawą analiz tego typu są dane na poziomie karty wyborczej (ballot-level).

W prawdzie Gore nie został prezydentem, ale usprawniono organizację wyborów.

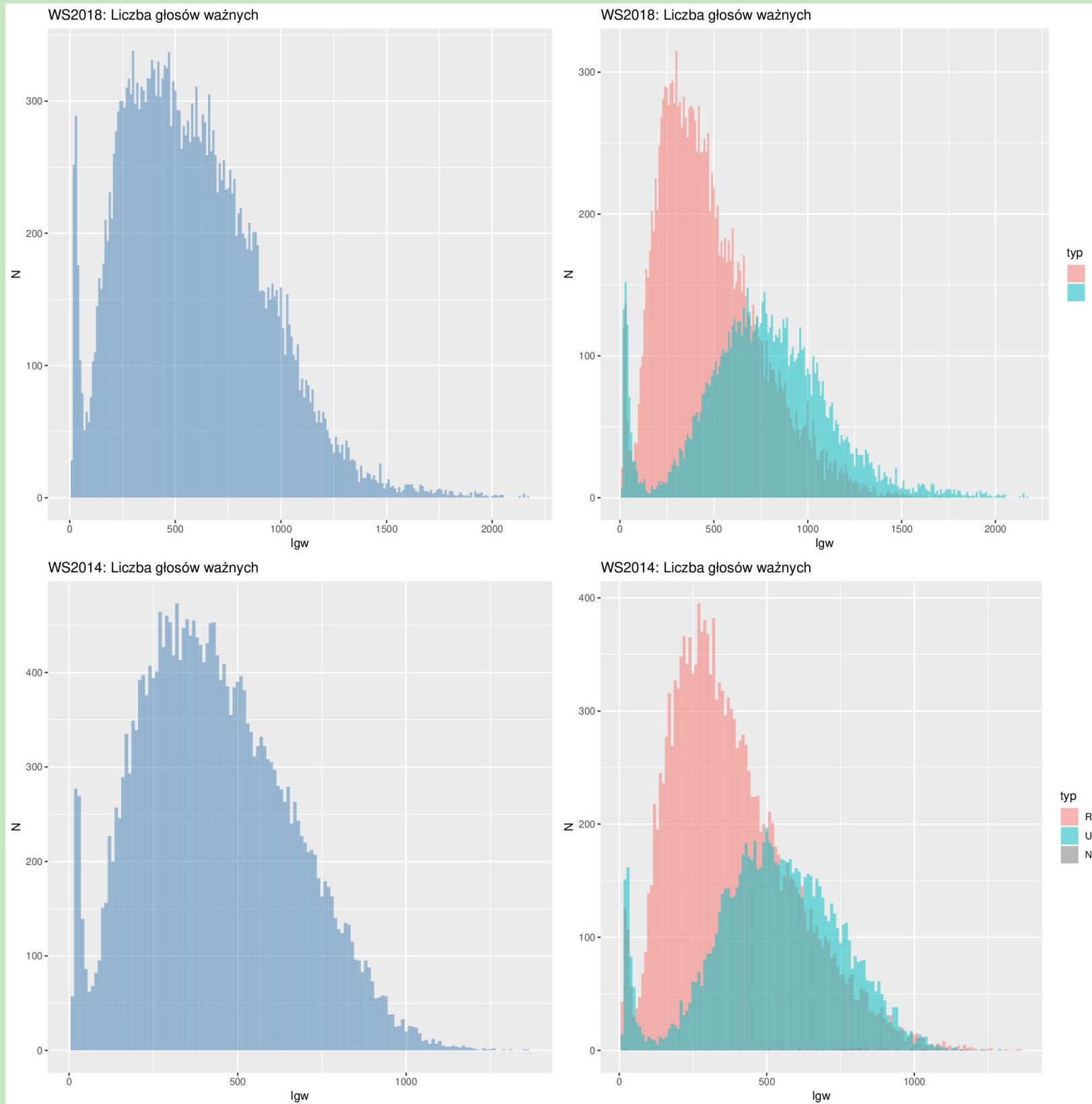
Wybory samorządowe 2014/2018: frekwencja



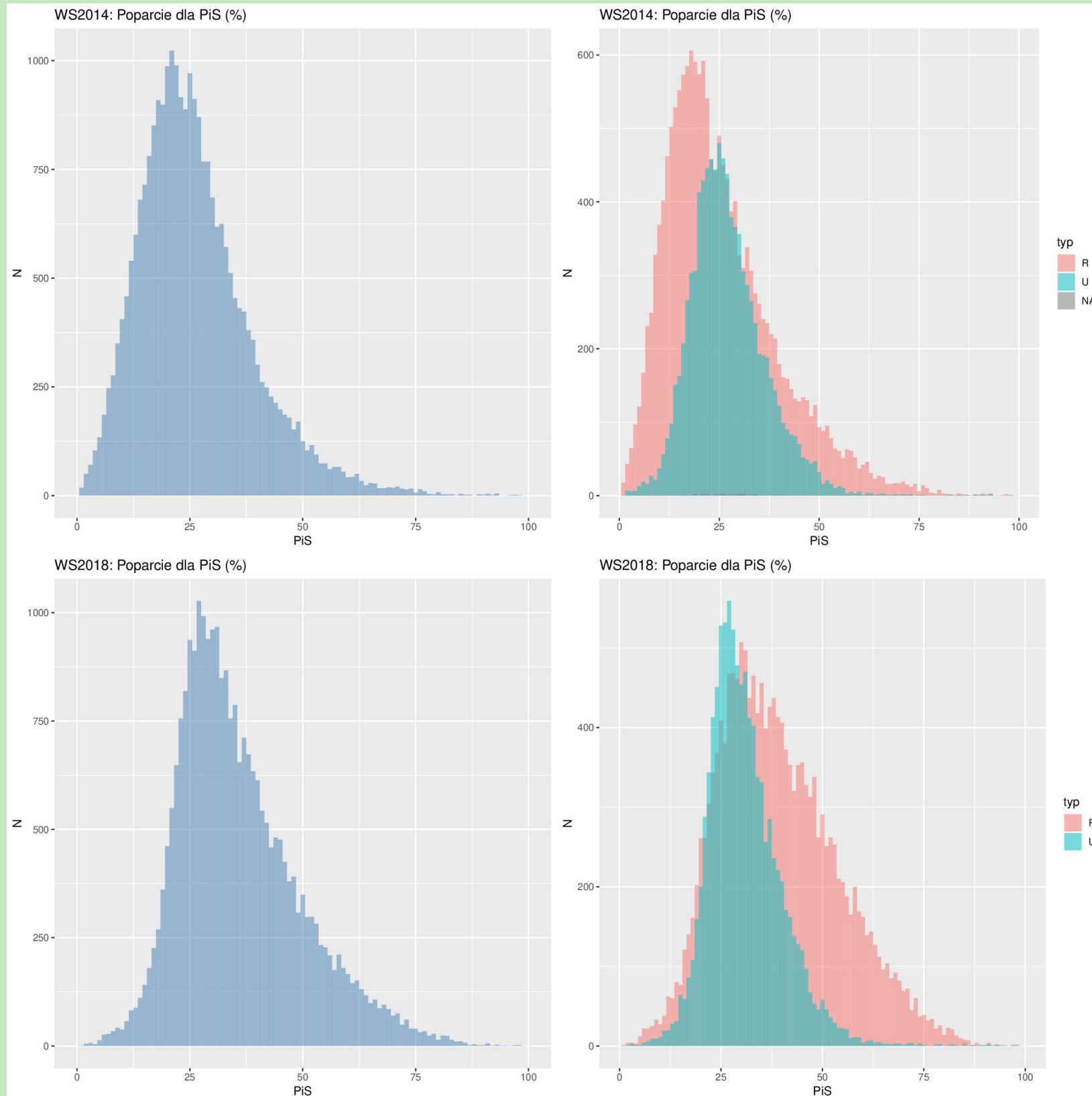
WS 2014/2018: odsetek głosów nieważnych



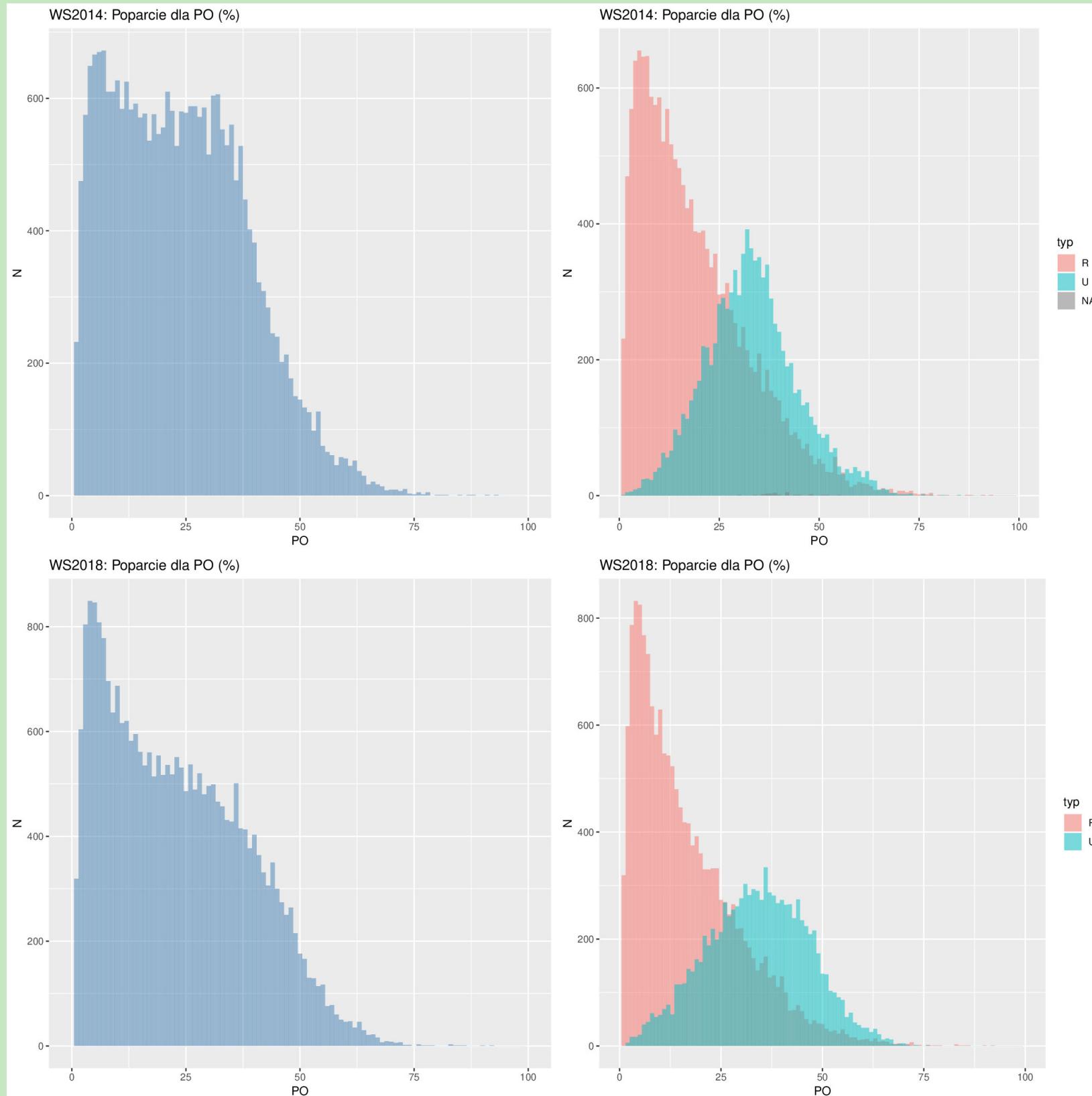
WS 2014/2018: liczba głosów ważnych



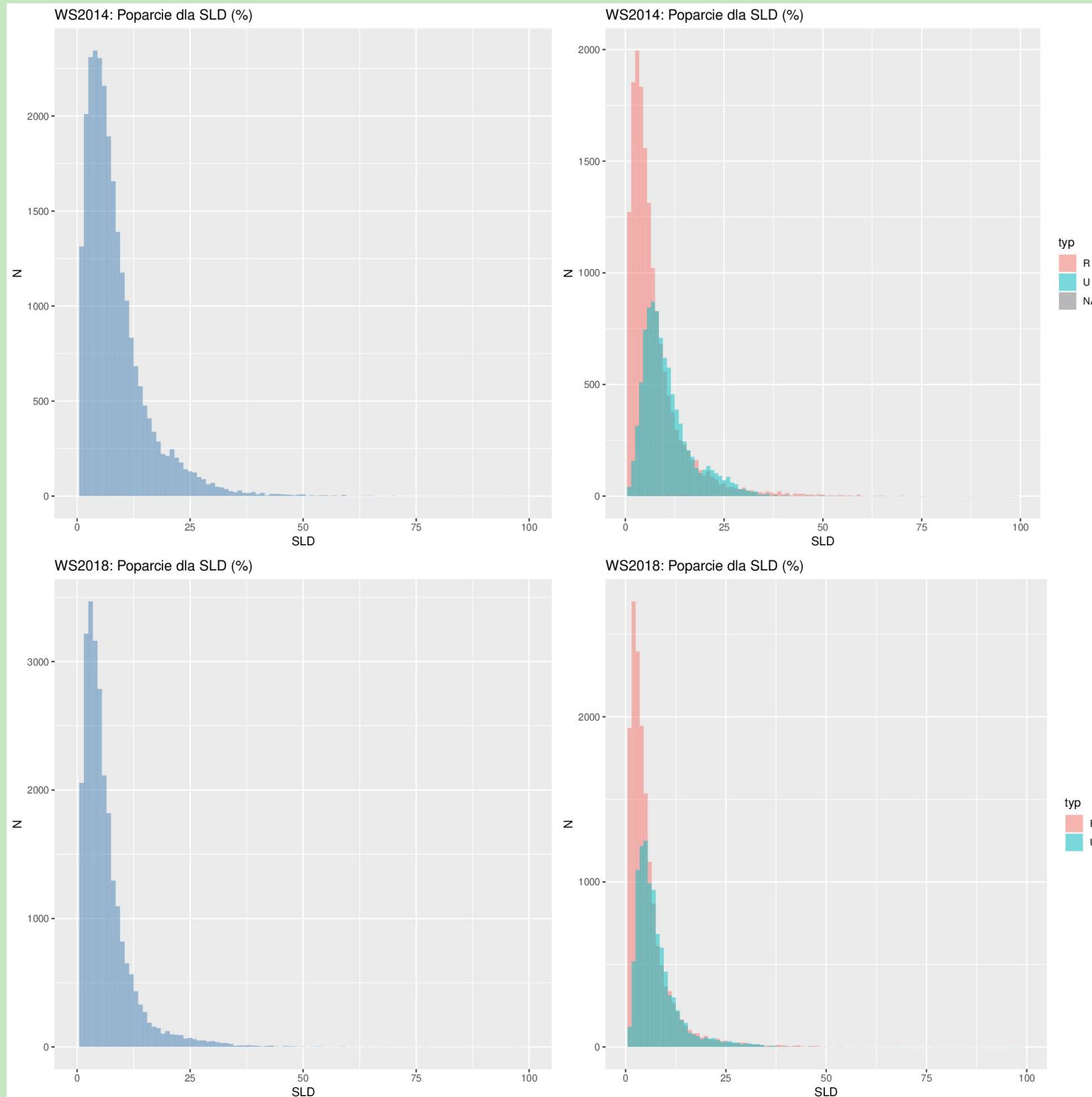
Wybory samorządowe 2014/2018: PiS



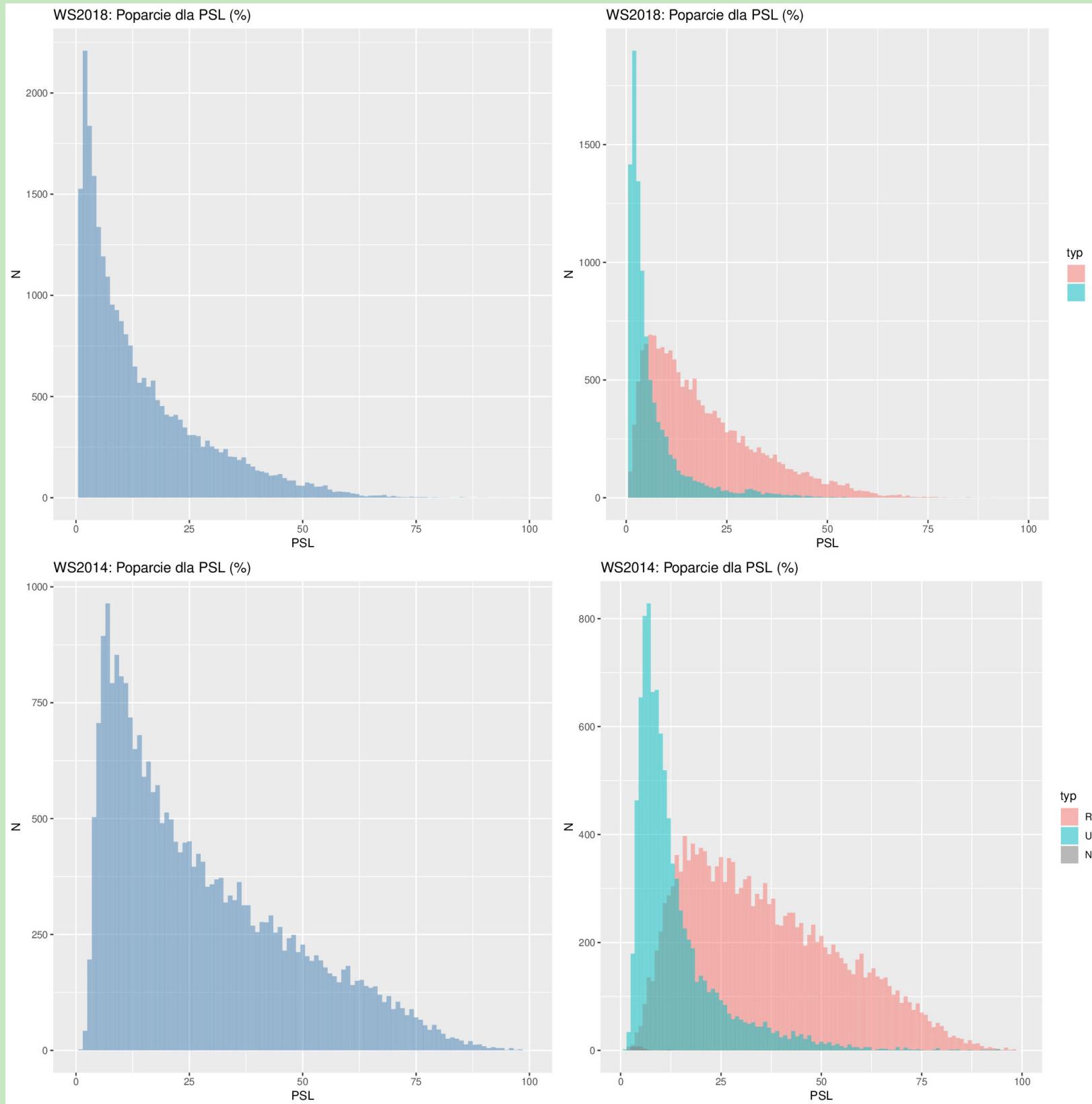
Wybory samorządowe 2014/2018: PO



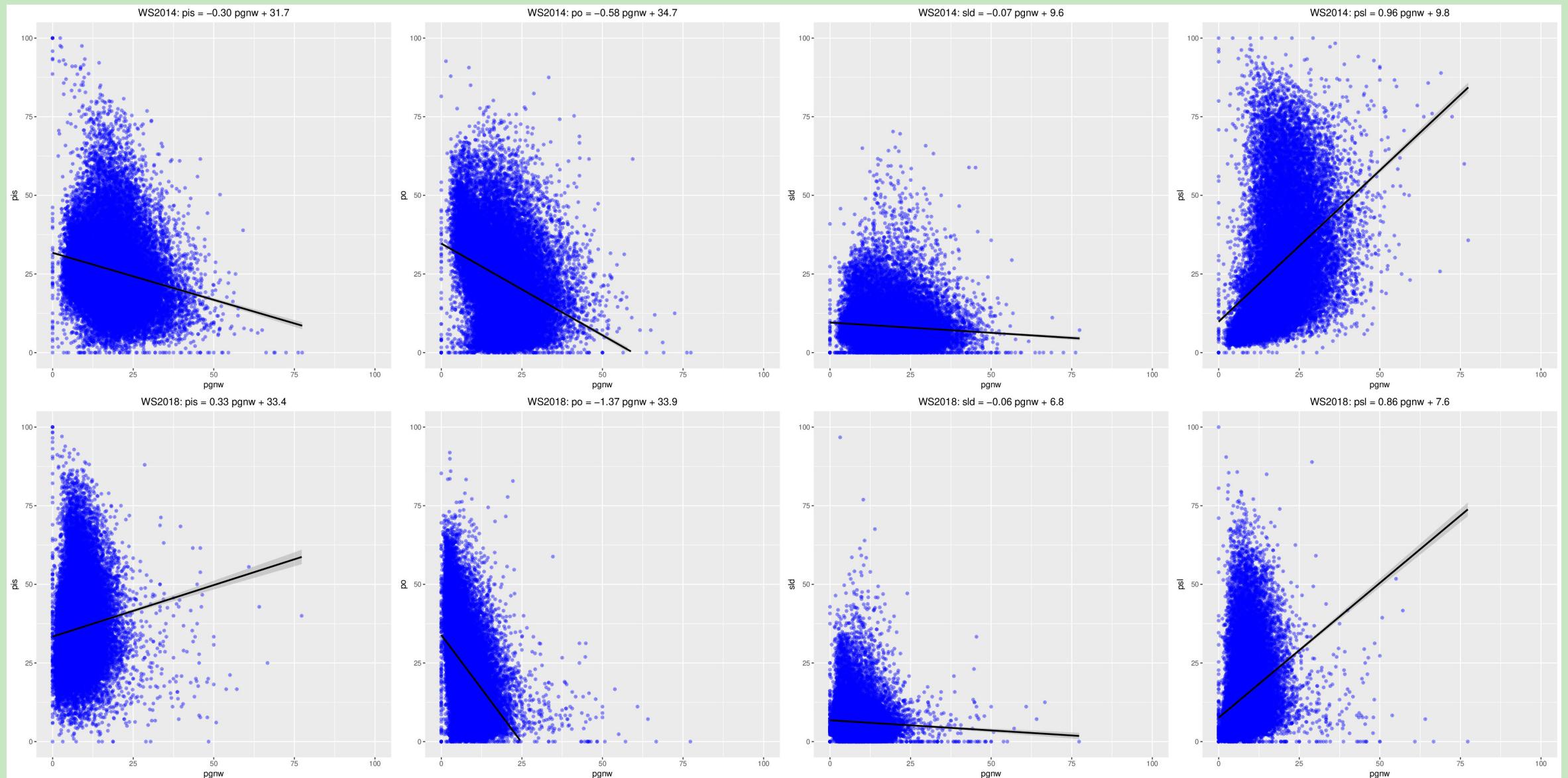
Wybory samorządowe 2014/2018: SLD



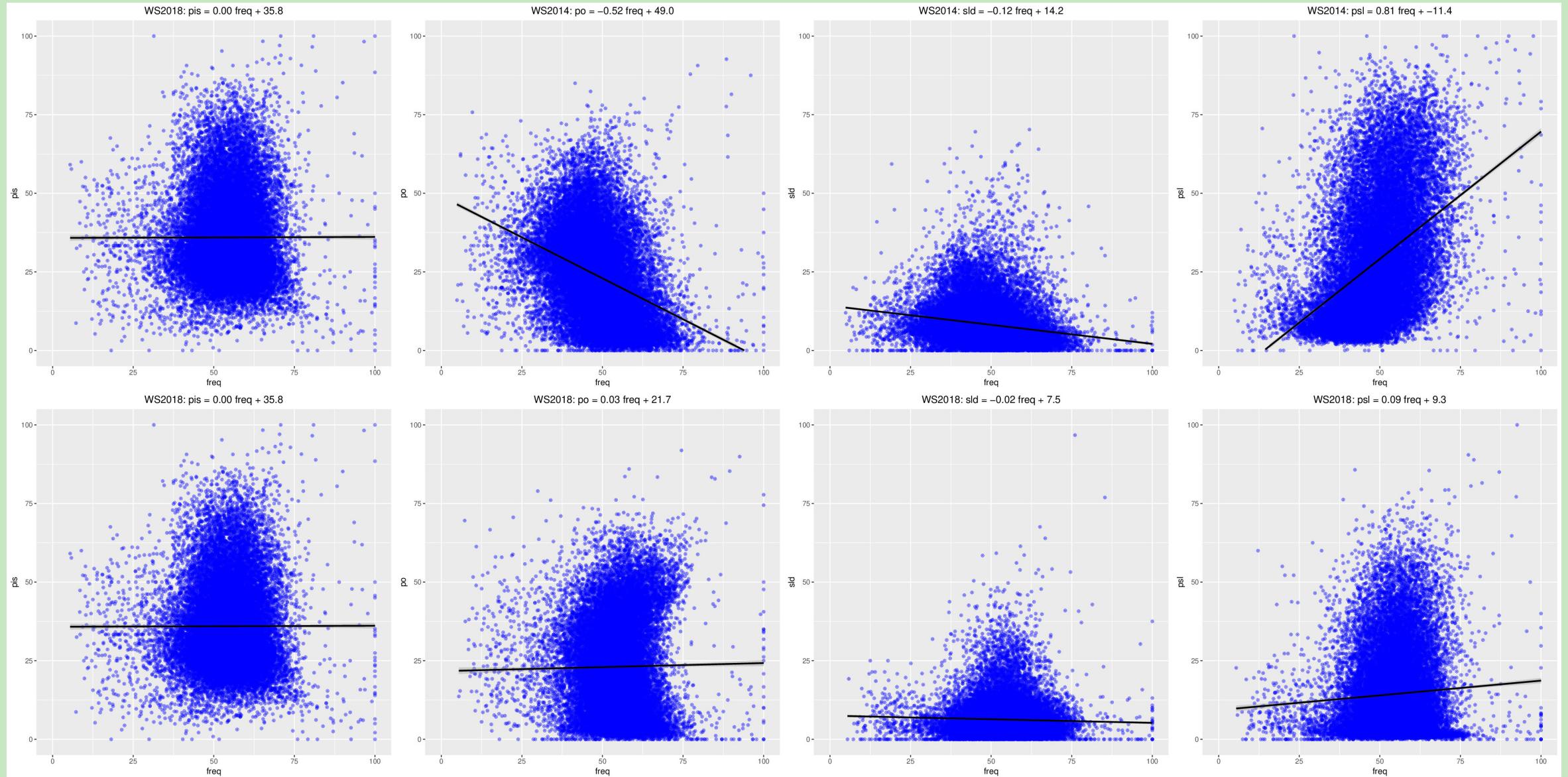
Wybory samorządowe 2014/2018: PSL



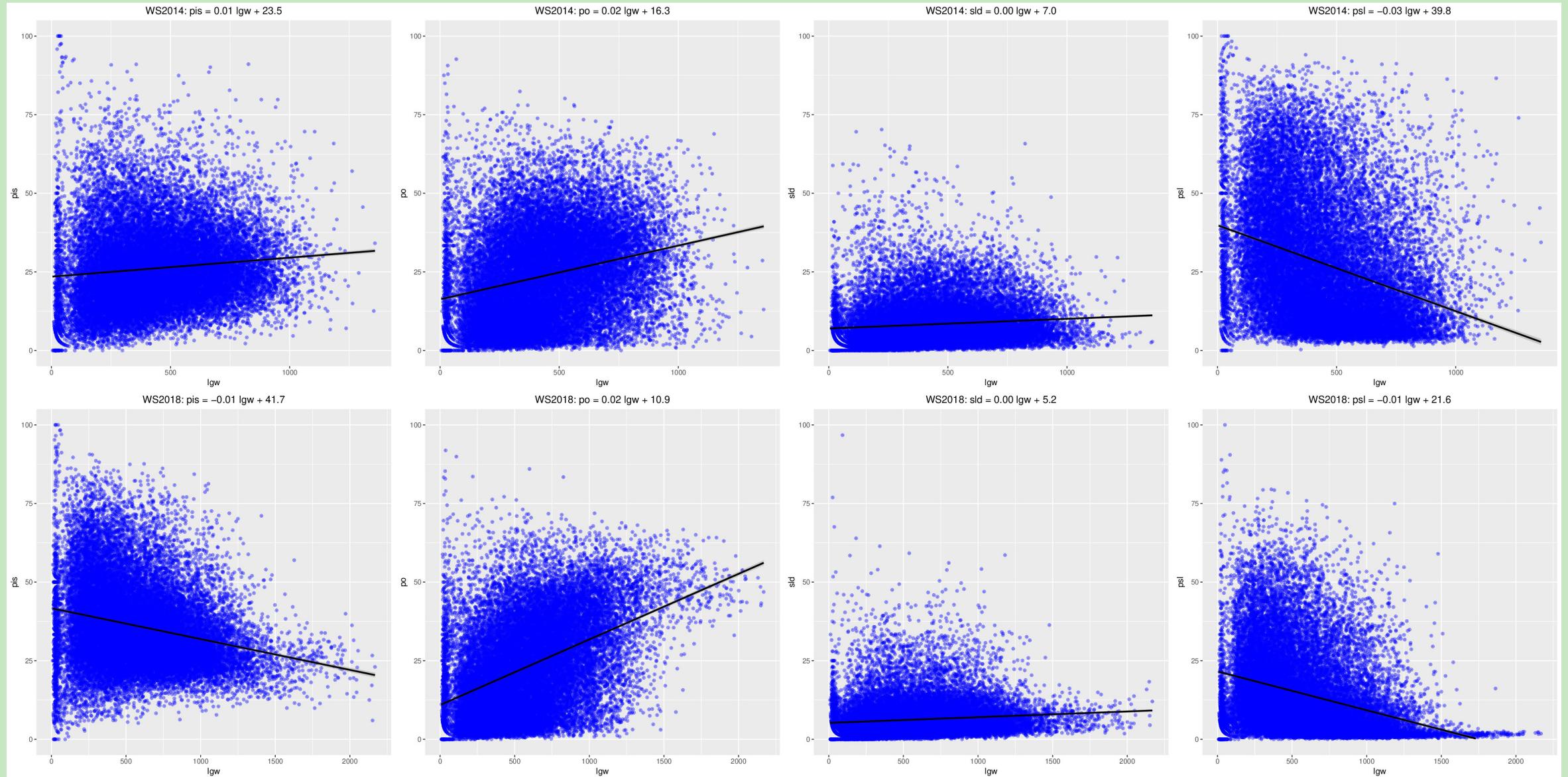
Odsetek głosów nieważnych a poparcie dla partii (Polska)



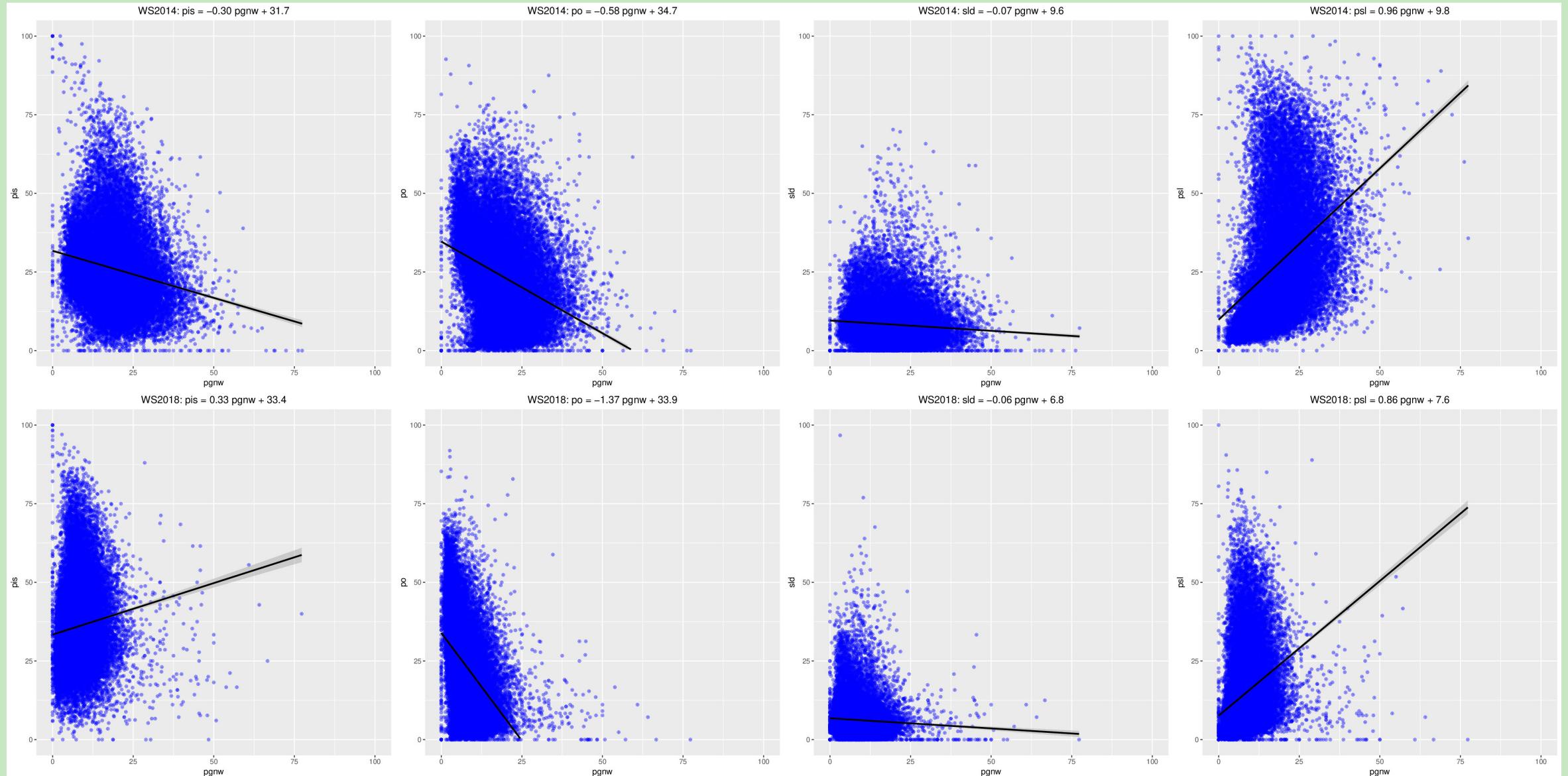
Frekwencja a poparcie dla partii (Polska)



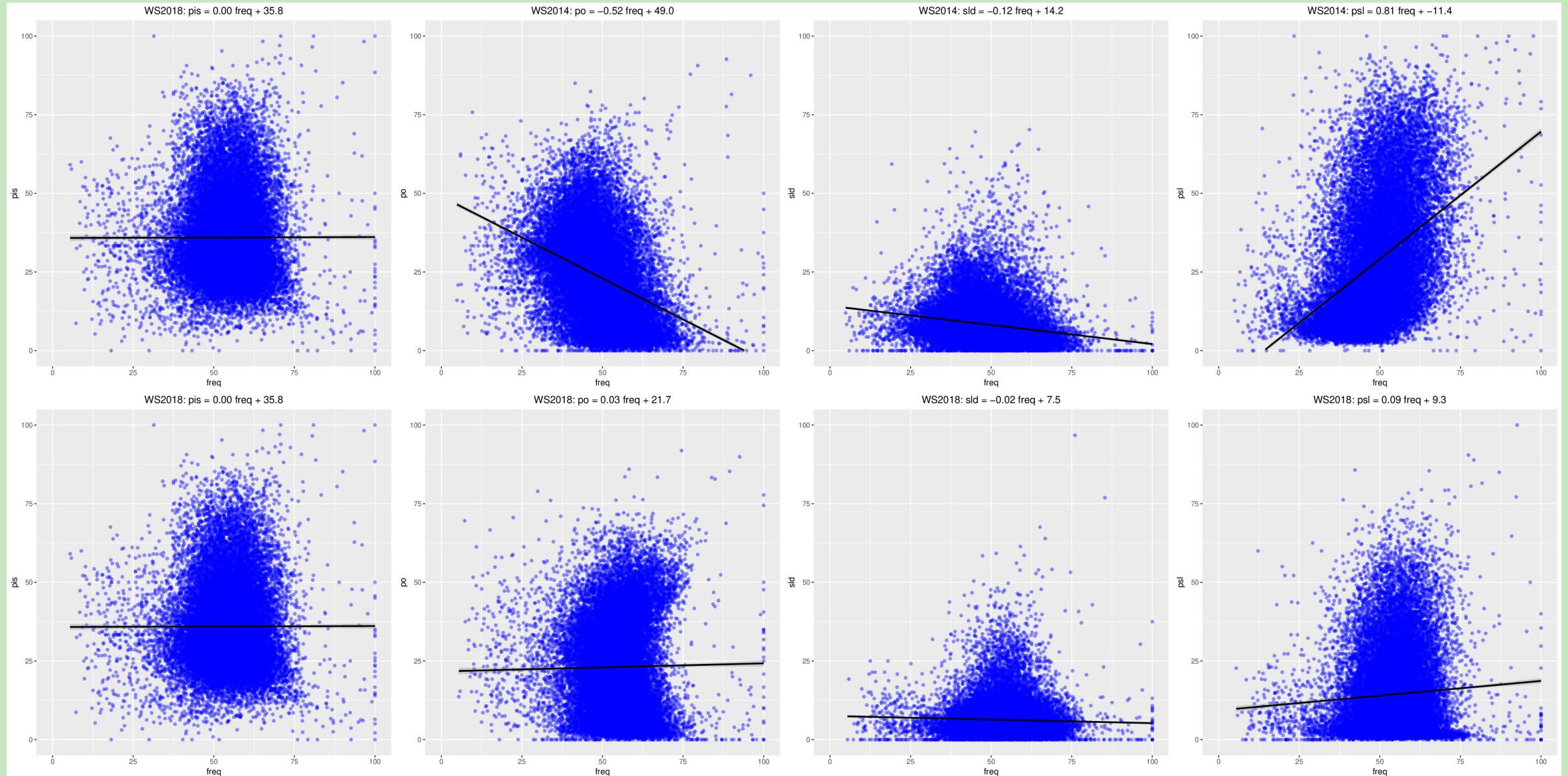
Liczba głosów ważnych a poparcie dla partii (Polska)



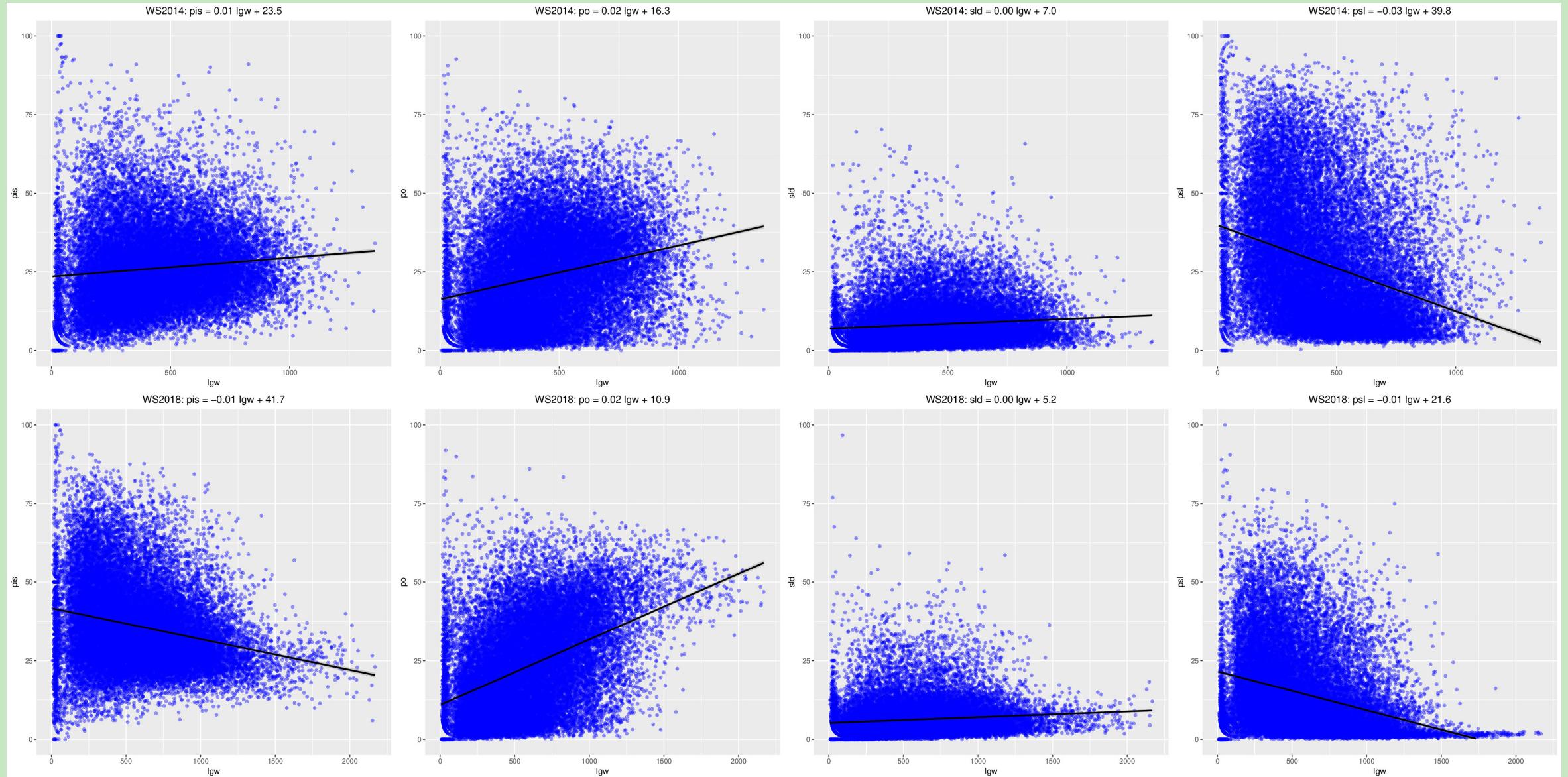
Pomorskie % głosów nieważnych a poparcie dla partii



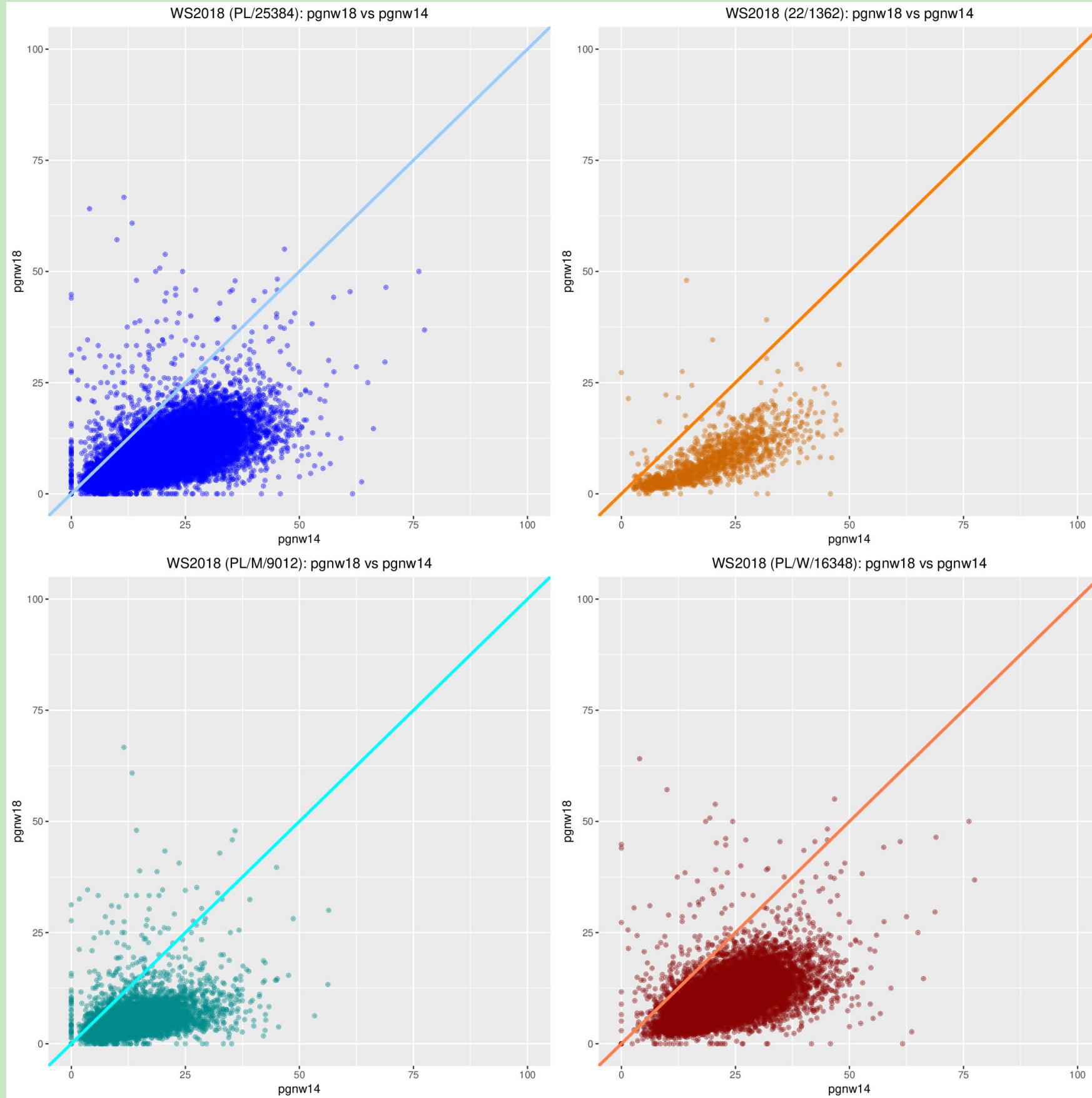
Pomorskie frekwencja a poparcie dla partii



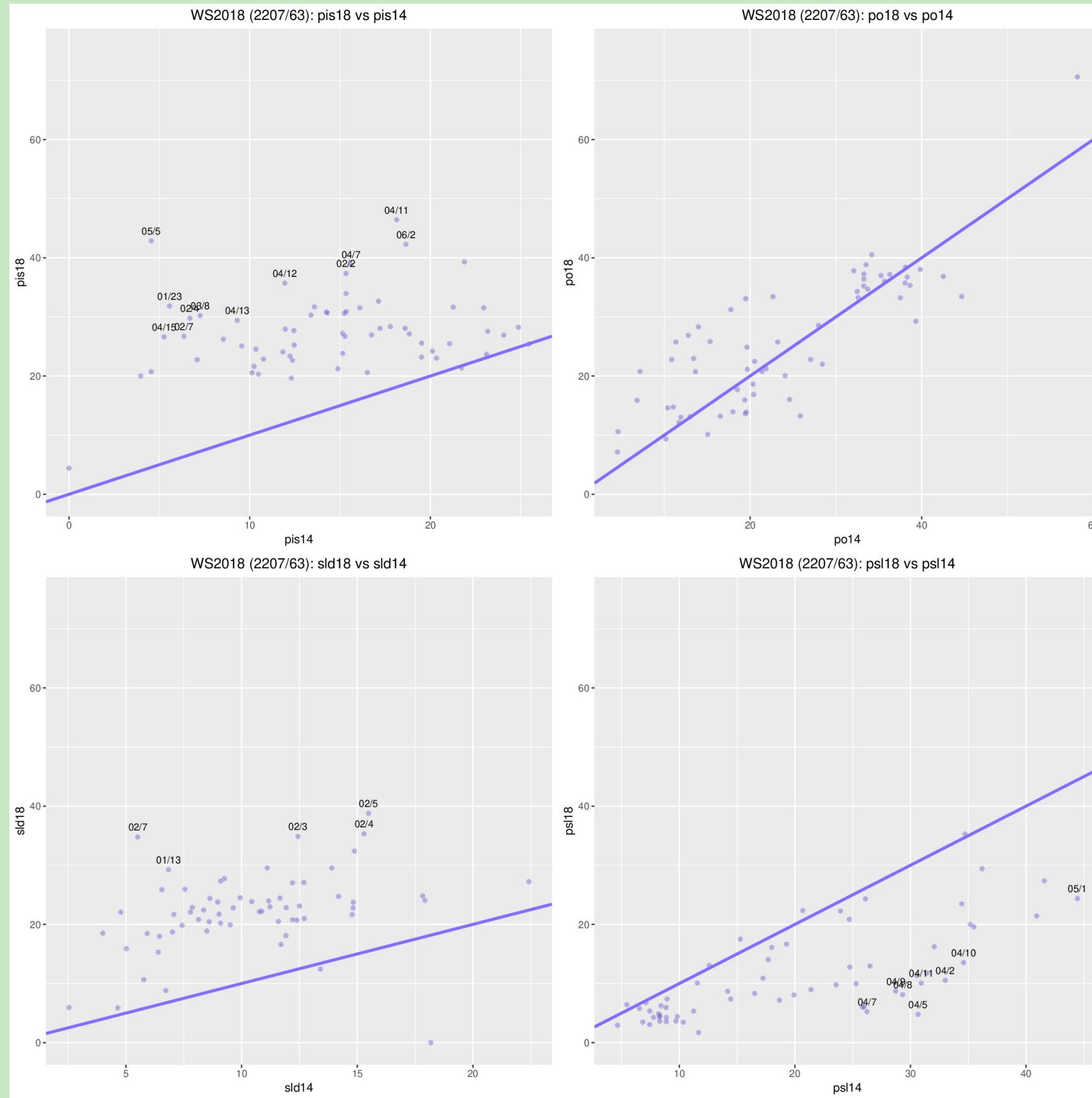
Pomorskie liczba głosów ważnych a poparcie dla partii



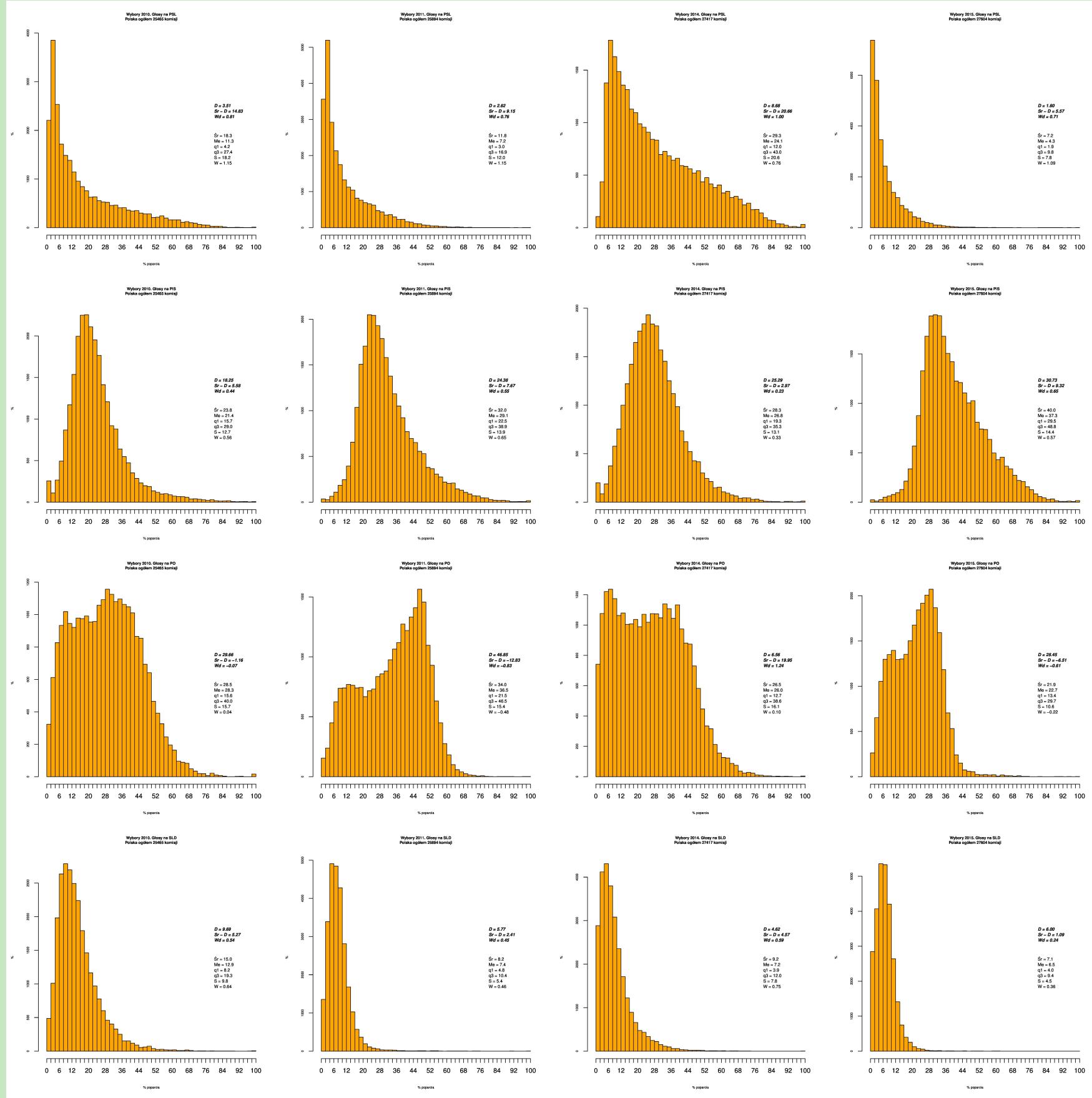
Odsetek głosów nieważnych 2014–2018



Odsetek poparcia dla partii w powiecie 2207 (2014–2018)



W poszukiwaniu anomalii rozkładu (cd)



Podsumowanie

Dane udostępnione przez PKW mogłyby być elementem edukacji obywatelskiej oraz systemu kontroli ale ich stan na dzień dzisiejszy pozwala na to w ograniczonym zakresie.

Można zauważać postęp w jakości oraz ilości udostępnianych danych.

W latach 2014–2019 pozyskałem głównie metodą webscrappingu dane dotyczące wyborów z lat 2005, 2007, 2011, 2015 (parlamentarne), 2010, 2014, 2018 (samorządowe), 2014, 2018 (do parlamentu europejskiego) oraz 2015 (prezydenckie). Zbiory danych są dostępne tutaj:

<https://github.com/hrpunio/Data>



Dziękuję za uwagę