factorMerger: hierarchiczna klasteryzacja i wizualizacja factorów

Agnieszka Sitko

MIMUW, Grupa MI^2

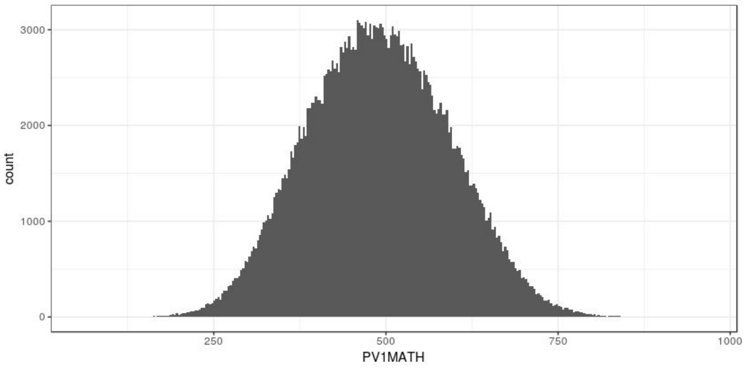
SER XXVI | 25-05-2017

Agenda

1. Dane PISA - jak uczniowie Europy radzą sobie z matematyką?
2. ANOVA i tradycyjne testy post hoc
3. factorMerger - nowe podejście do testów post hoc

PISA 2012

- 0.5 mln studentów,
- 65 krajów,
- 3 kategorie: matematyka, czytanie, wiedza



Pleasible values:

- ocena osiągnięć studentów,
- dane normalizowane,
- średnia i odchylenie standardowe OECD: 500, 100.

Więcej tutaj.

PISA 2012 - trzy klastry

	Mathematics				Reading		Science	
	Mean score in PISA 2012	Share of low achievers in mathematics (Below Level 2)	Share of top performers in mathematics (Level 5 or 6)	Annualised change in score points	Mean score in PISA 2012	Annualised change in score points	Mean score in PISA 2012	Annualised change in score points
OECD average	494	23.0	12.6	**-0.3**	496	**0.3**	501	**0.5**
Shanghai-China	613	3.8	55.4	**4.2**	570	**4.6**	580	1.8
Singapore	573	8.3	40.0	**3.8**	542	**5.4**	551	**3.3**
Hong Kong-China	561	8.5	33.7	**1.3**	545	**2.3**	555	**2.1**
Chinese Taipei	560	12.8	37.2	1.7	523	**4.5**	523	-1.5
Korea	554	9.1	30.9	1.1	536	**0.9**	538	**2.6**
Macao-China	538	10.8	24.3	**1.0**	509	**0.8**	521	**1.6**
Japan	536	11.1	23.7	0.4	538	**1.5**	547	**2.6**
Liechtenstein	535	14.1	24.8	0.3	516	**1.3**	525	0.4
Switzerland	531	12.4	21.4	0.6	509	**1.0**	515	0.6
Netherlands	523	14.8	19.3	**-1.6**	511	-0.1	522	-0.5
Estonia	521	10.5	14.6	0.9	516	**2.4**	541	1.5
Finland	519	12.3	15.3	**-2.8**	524	**-1.7**	545	**-3.0**
Canada	518	13.8	16.4	**-1.4**	523	**-0.9**	525	**-1.5**
Poland	518	14.4	16.7	**2.6**	518	**2.8**	526	**4.6**
Belgium	515	19.0	19.5	**-1.6**	509	0.1	505	-0.9
Germany	514	17.7	17.5	**1.4**	508	**1.8**	524	1.4
Viet Nam	511	14.2	13.3	m	508	m	528	m
Austria	506	18.7	14.3	0.0	490	-0.2	506	-0.8
Australia	504	19.7	14.8	**-2.2**	512	**-1.4**	521	-0.9
Ireland	501	16.9	10.7	-0.6	523	**-0.9**	522	**2.3**
Slovenia	501	20.1	13.7	-0.6	481	**-2.2**	514	-0.8
Denmark	500	16.8	10.0	**-1.8**	496	0.1	498	0.4
New Zealand	500	22.6	15.0	**-2.5**	512	**-1.1**	516	**-2.5**
Czech Republic	499	21.0	12.9	**-2.5**	493	-0.5	508	-1.0
France	495	22.4	12.9	**-1.5**	505	0.0	499	0.6
United Kingdom	494	21.8	11.8	-0.3	499	0.7	514	-0.1
Iceland	493	21.5	11.2	**-2.2**	483	**-1.3**	478	**-2.0**
Latvia	491	19.9	8.0	0.5	489	**1.9**	502	**2.0**
Luxembourg	490	24.3	11.2	-0.3	488	**0.7**	491	0.9
Norway	489	22.3	9.4	-0.3	504	0.1	495	1.3
Portugal	487	24.9	10.6	**2.8**	488	**1.6**	489	**2.5**
Italy	485	24.7	9.9	**2.7**	490	0.5	494	**3.0**
Spain	484	23.6	8.0	0.1	488	-0.3	496	1.3
Russian Federation	482	24.0	7.8	1.1	475	**1.1**	486	1.0
Slovak Republic	482	27.5	11.0	**-1.4**	463	-0.1	471	**-2.7**
United States	481	25.8	8.8	0.3	498	-0.3	497	1.4
Lithuania	479	26.0	8.1	-1.4	477	1.1	496	1.3
Sweden	478	27.1	8.0	**-3.3**	483	**-2.8**	485	**-3.1**
Hungary	477	28.1	9.3	**-1.3**	488	**1.0**	494	-1.6
Croatia	471	29.9	7.0	0.6	485	1.2	491	-0.3
Israel	466	33.5	9.4	**4.2**	486	**3.7**	470	**2.8**
Greece	453	35.7	3.9	**1.1**	477	0.5	467	-1.1
Serbia	449	38.9	4.6	**2.2**	446	**7.6**	445	1.5
Turkey	448	42.0	5.9	**3.2**	475	**4.1**	463	**6.4**
Romania	445	40.8	3.2	**4.9**	438	1.1	439	**3.4**
Cyprus[1, 2]	440	42.0	3.7	m	449	m	438	m
Bulgaria	439	43.8	4.1	**4.2**	436	0.4	446	2.0
United Arab Emirates	434	46.3	3.5	m	442	m	448	m
Kazakhstan	432	45.2	0.9	**9.0**	393	0.8	425	**8.1**
Thailand	427	49.7	2.6	1.0	441	**1.1**	444	**3.9**
Chile	423	51.5	1.6	**1.9**	441	**3.1**	445	1.1
Malaysia	421	51.8	1.3	**8.1**	398	**-7.8**	420	-1.4
Mexico	413	54.7	0.6	**3.1**	424	**1.1**	415	0.9
Montenegro	410	56.6	1.0	**1.7**	422	**5.0**	410	-0.3
Uruguay	409	55.8	1.4	**-1.4**	411	**-1.8**	416	**-2.1**
Costa Rica	407	59.9	0.6	-1.2	441	-1.0	429	-0.6
Albania	394	60.7	0.8	**5.6**	394	**4.1**	397	2.2
Brazil	391	67.1	0.8	**4.1**	410	**1.2**	405	**2.3**
Argentina	388	66.5	0.3	1.2	396	-1.6	406	2.4
Tunisia	388	67.7	0.8	**3.1**	404	**3.8**	398	**2.2**
Jordan	386	68.6	0.6	0.2	399	-0.3	409	**-2.1**
Colombia	376	73.8	0.3	1.1	403	**3.0**	399	1.8
Qatar	376	69.6	2.0	**9.2**	388	**12.0**	384	**5.4**
Indonesia	375	75.7	0.3	0.7	396	**2.3**	382	-1.9
Peru	368	74.6	0.6	1.0	384	**5.2**	373	1.3

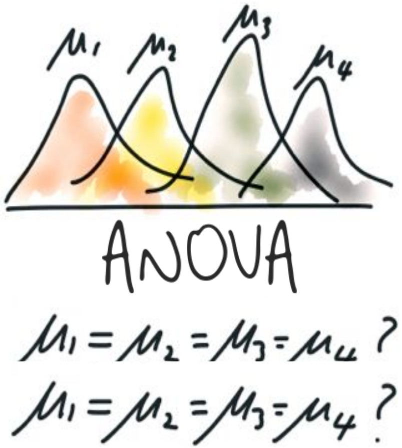
Kraje o wynikach statystycznie:

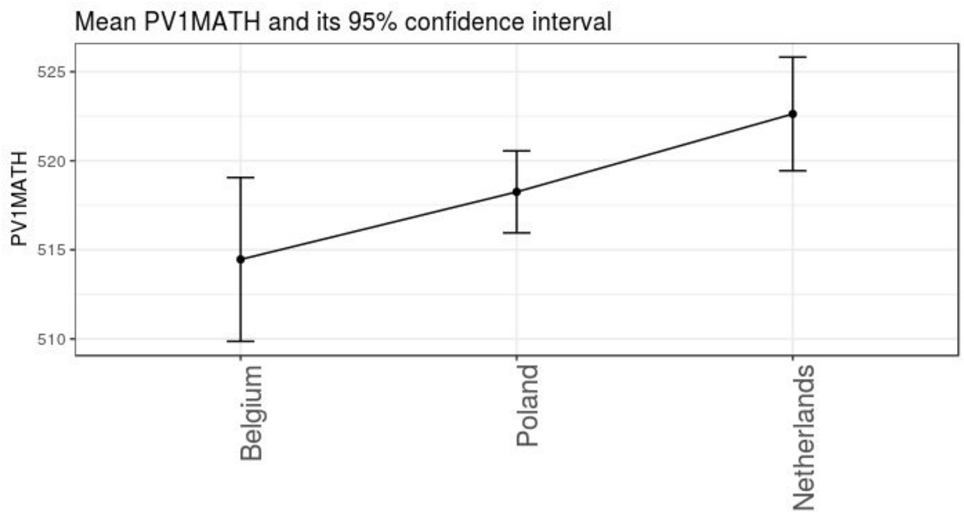
lepszych niż średnia OECD

równych średniej OECD

niższych niż średnia OECD

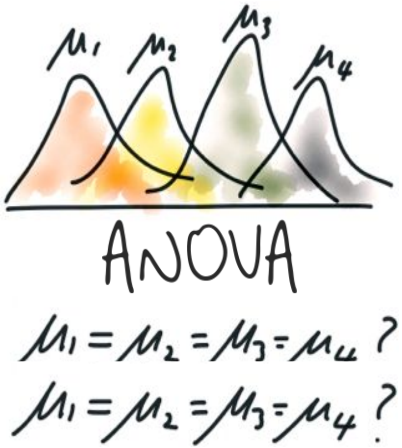
http://www.oecd.org/pisa/keyfindings/PISA-2012-results-snapshot-Volume-I-ENG.pdf

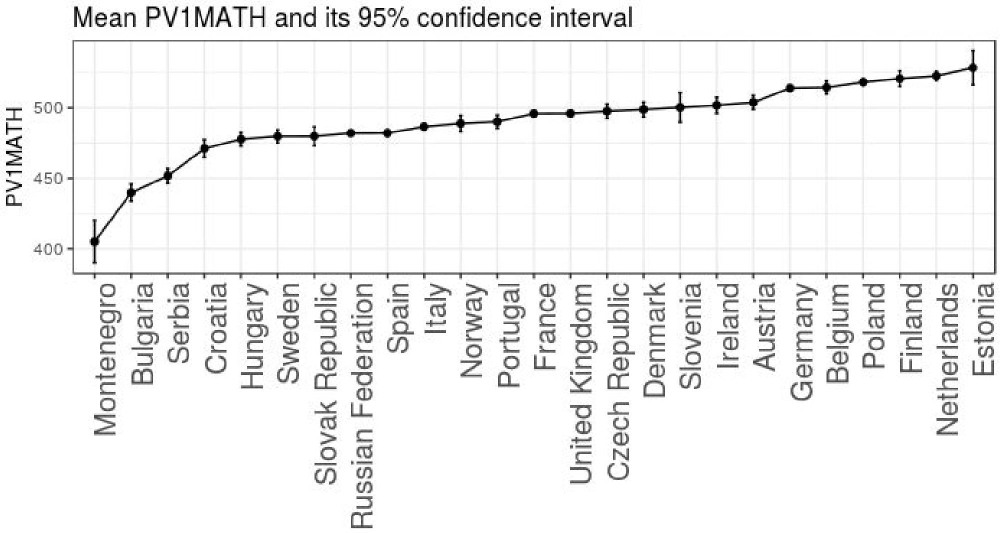





```
anova(lm(PV1MATH ~ CNT, data = filter(pisaEuropean, CNT %in% c("Belgium", "Poland", "Netherlands"))))
```

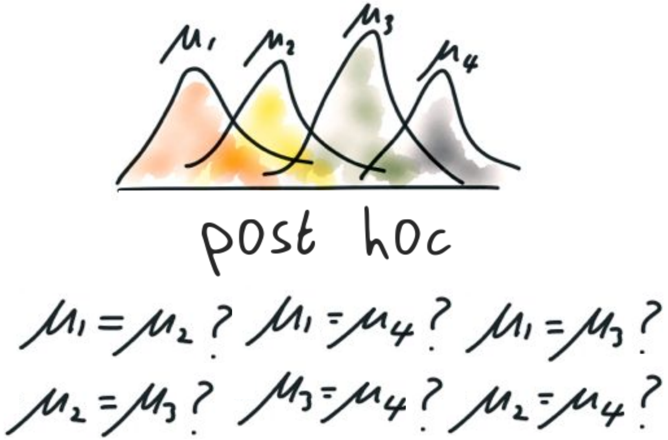
```
## Analysis of Variance Table
##
## Response: PV1MATH
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## CNT           2    84272   42136  4.8359 0.007956 **
## Residuals 11113 96829278    8713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```






```r
anova(lm(PV1MATH ~ CNT, data = pisaEuropean))
```

```
## Analysis of Variance Table
##
## Response: PV1MATH
##              Df    Sum Sq Mean Sq F value    Pr(>F)
## CNT          24  23790251  991260  115.87 < 2.2e-16 ***
## Residuals 92411 790577926    8555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

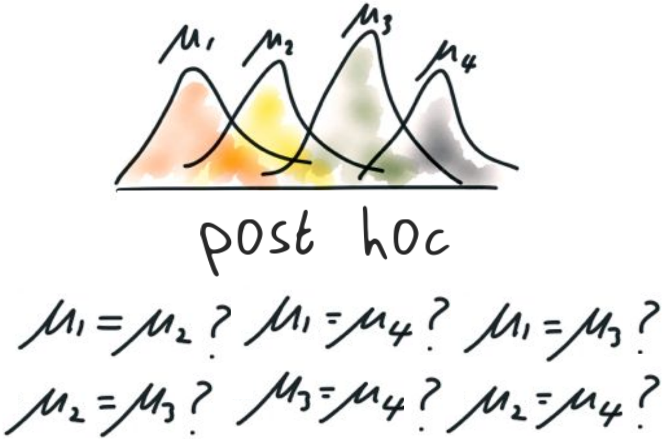


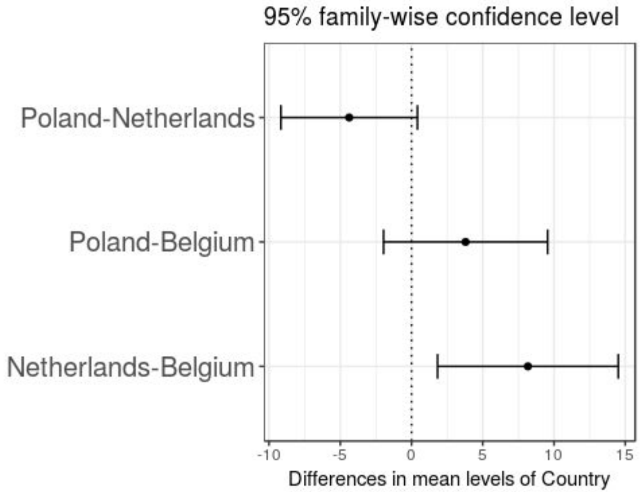
Testy post hoc

Post hoc - po fakcie

Tukey HSD	`TukeyHSD{stats}, glht{multcomp}, HSD.test{agricolae}`
LSD Fishera	`LSD.test{agricolae}`
Student-Newman-Keuls	`SNK.test {agricolae}`
Scheffe	`scheffe.test {agricolae}`

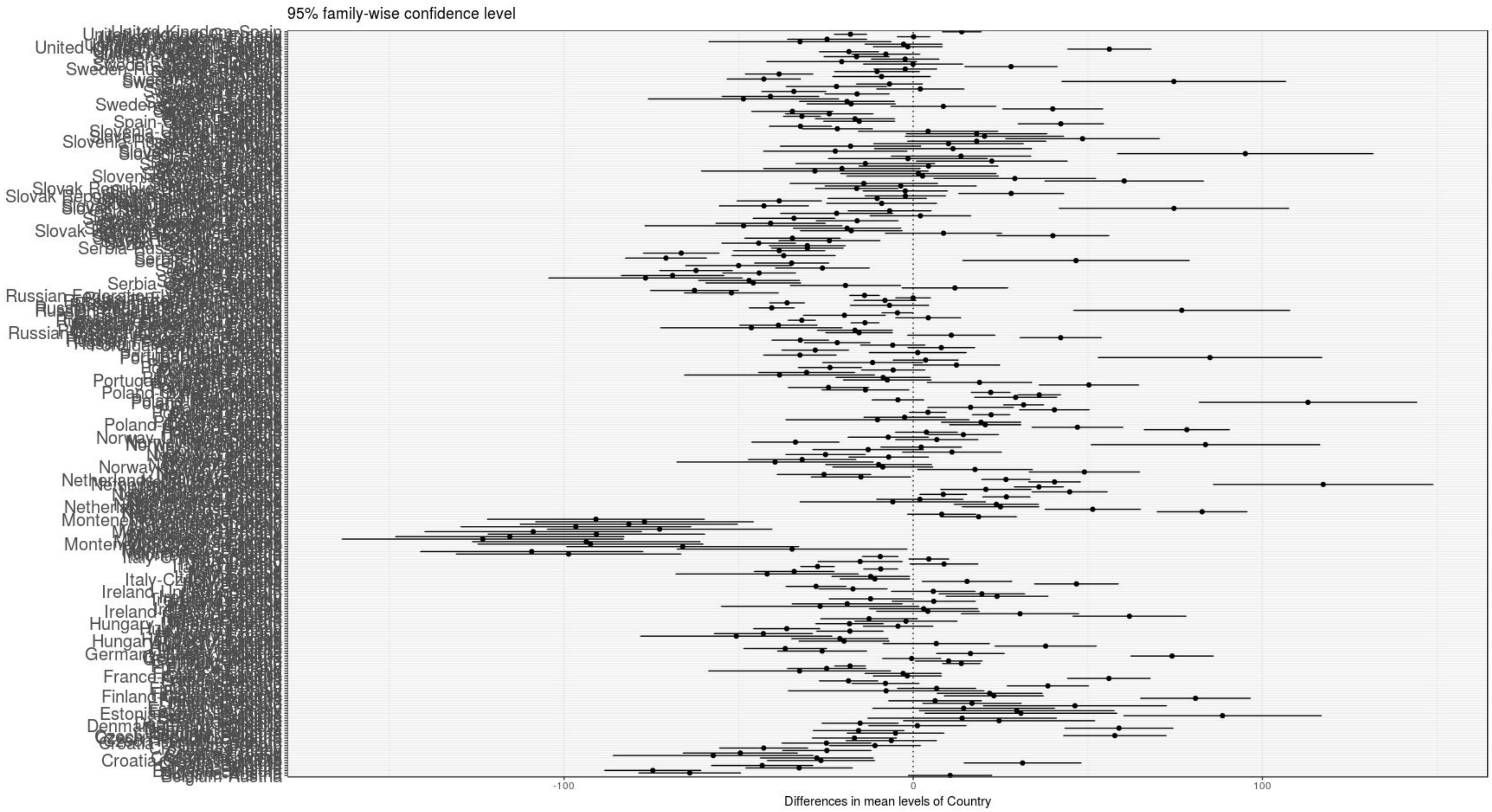
Więcej o testach post hoc: *Biecek, Przemysław. Analiza danych z programem R: modele liniowe z efektami stałymi, losowymi i mieszanymi. [2.2.4. Zagadnienie: testy post hoc]*



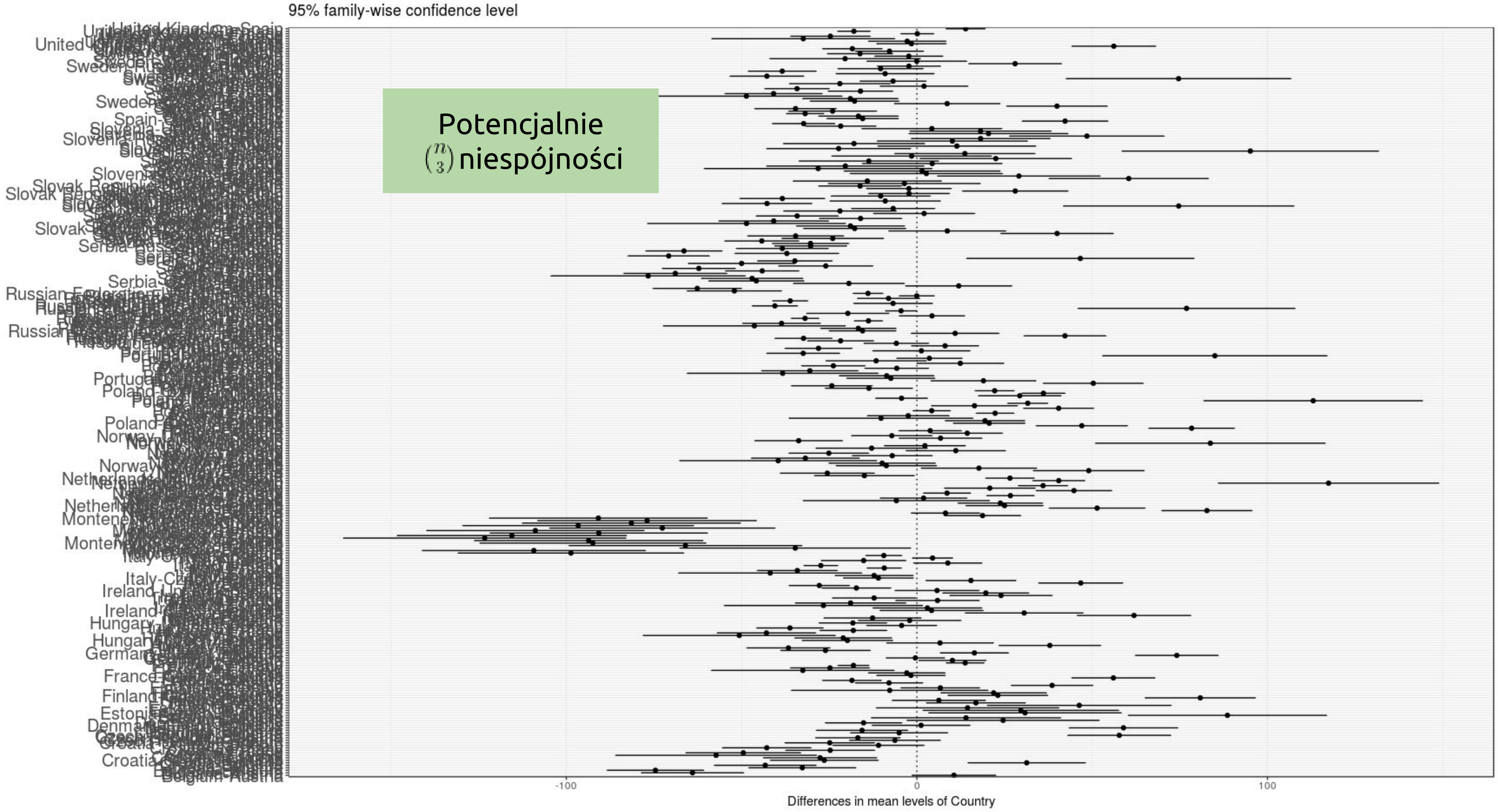


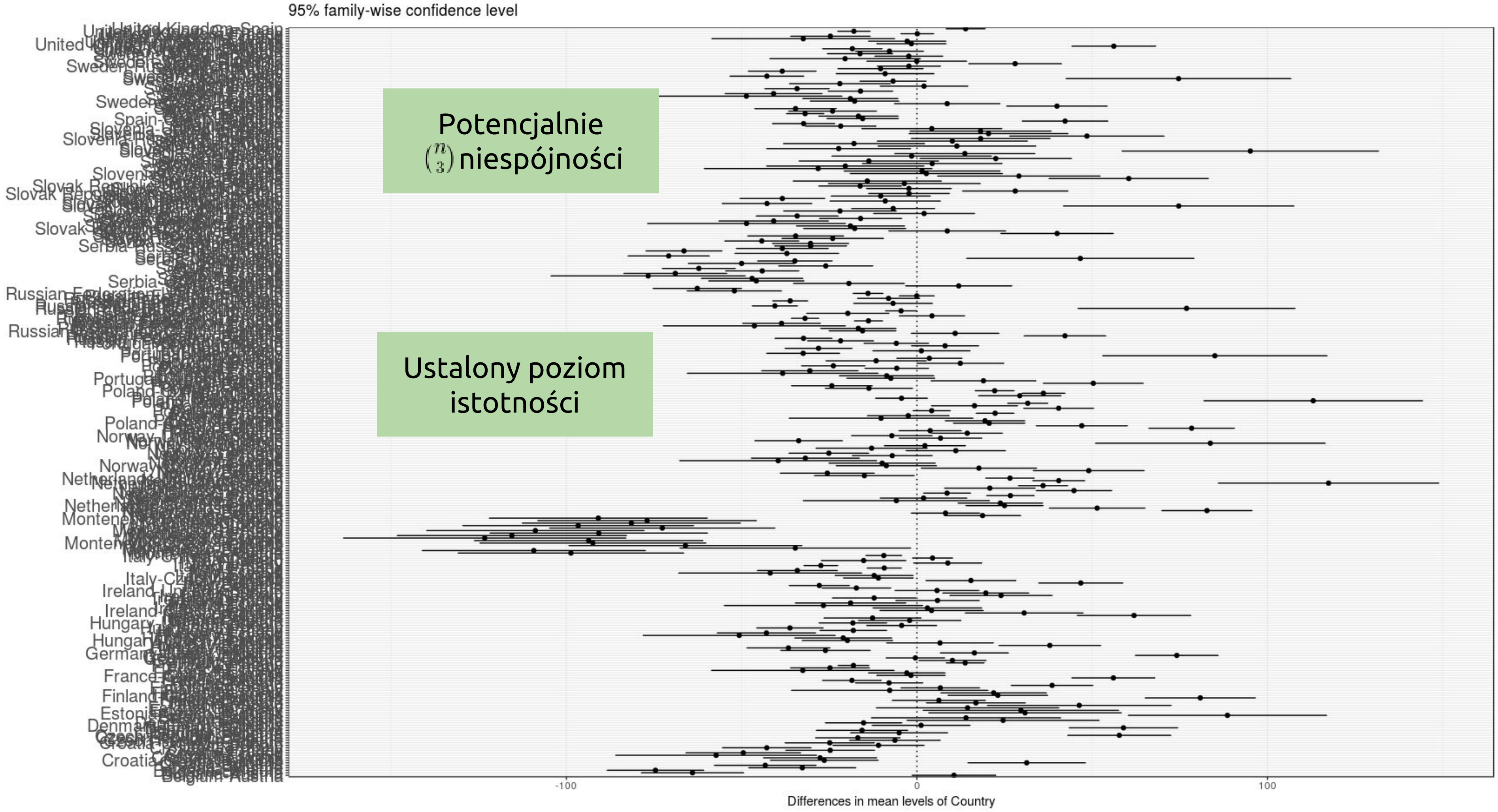

```r
tk <- TukeyHSD(aovPISA, "CNT")
tk
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = PV1MATH ~ CNT, data = filter(pisaEuropean, CNT %in% c("Poland", "Belgium", "Netherlands")))
##
## $CNT
##                          diff       lwr        upr     p adj
## Netherlands-Belgium  8.168042  1.827307 14.5087766 0.0071606
## Poland-Belgium       3.793268 -1.962187  9.5487229 0.2700258
## Poland-Netherlands  -4.374774 -9.166789  0.4172409 0.0819931
```



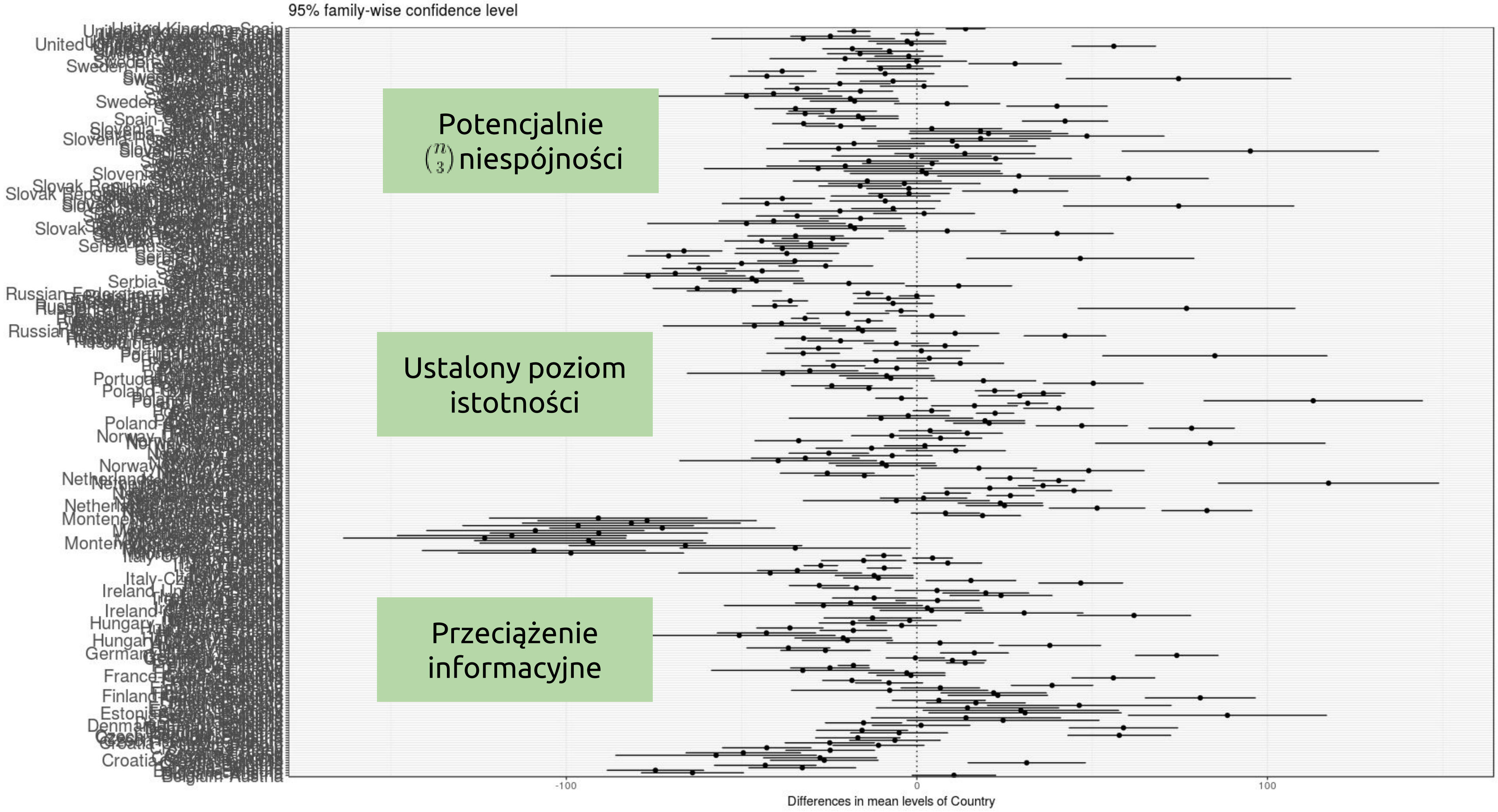
95% family-wise confidence level
-100
0
100
Differences in mean levels of Country


95% family-wise confidence level
Potencjalnie
$\binom{n}{3}$ niespójności
-100
0
100
Differences in mean levels of Country

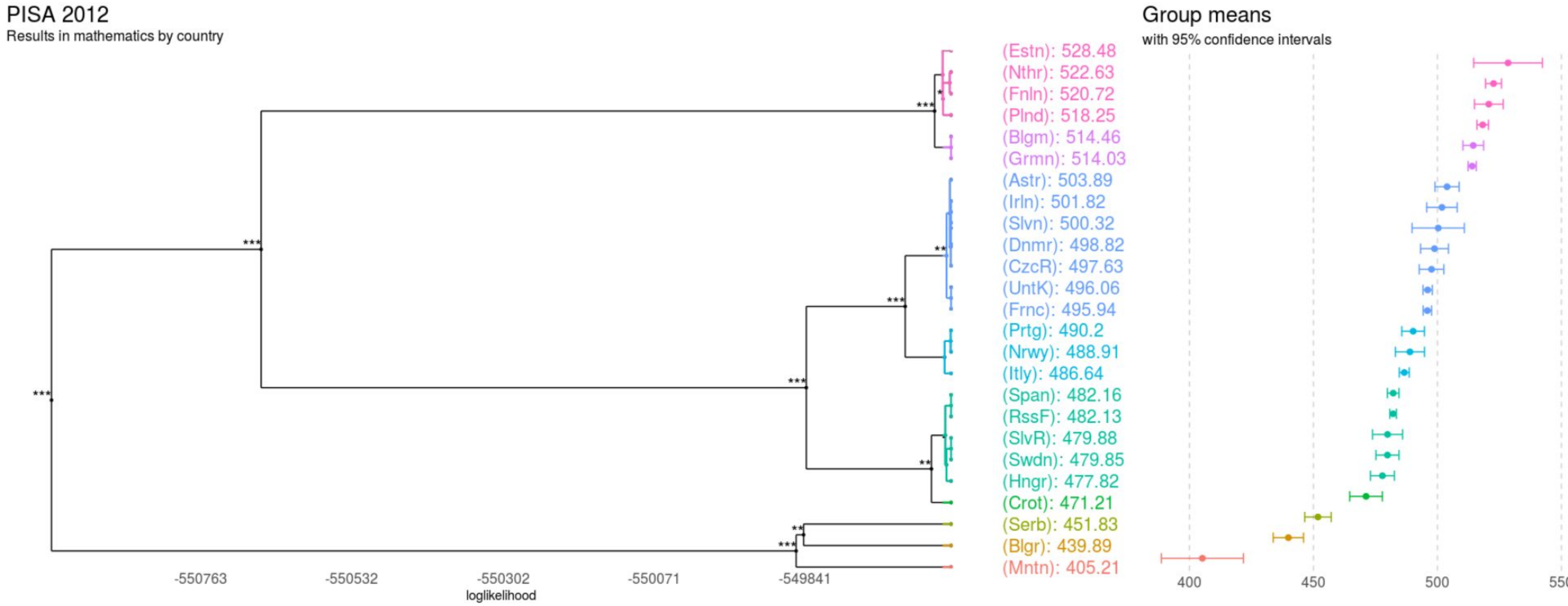


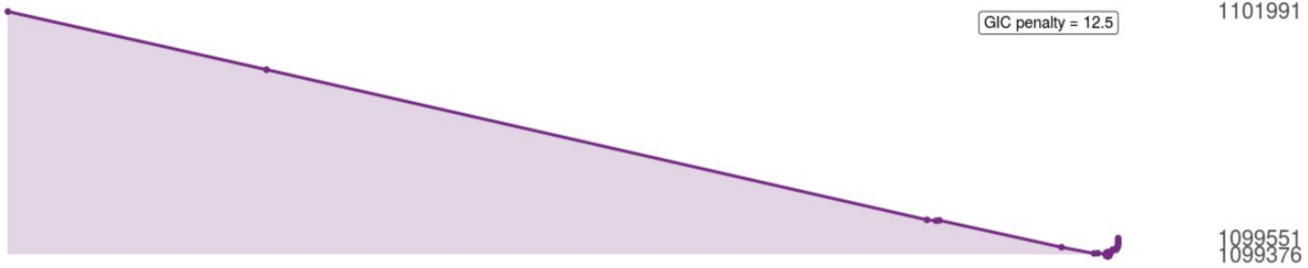
Potencjalnie $\binom{n}{3}$ niespójności

Ustalony poziom istotności


95% family-wise confidence level
Potencjalnie $\binom{n}{3}$ niespójności
Ustalony poziom istotności
Przeciążenie informacyjne
-100
0
100
Differences in mean levels of Country

Miło mi przedstawić factorMerger





ANOVA table

	Df	F	p-value
factor	25	105939.6	< 2.2e-16
Res	92411		

Merge

1. Testy ilorazu wiarygodności
2. Delete or Merge Regressors

```
factorMerger::mergeFactors(response = myResponse,
                           factor = myFactor,
                           method = "LRT")
```

```
factorMerger::mergeFactors(response = myResponse,
                           factor = myFactor,
                           method = "hclust",
                           successive = TRUE)
```

Merge

1. Testy ilorazu wiarygodności
2. Delete or Merge Regressors

Algorithm 1 Merging with *LRT*

function MERGEFACTORS($response, factor, successive$)
  $pairsSet := generatePairs(response, factor, successive)$
  $M_0 :=$ full model
  while $levels(factor) > 1$ **do**
    $toBeMerged := \text{argmax}_{pair \in pairsSet} l(updateModel(M_0, pair))$
    $M_0 := updateModel(M_0, toBeMerged)$
    $factor := mergeLevels(factor, pair)$
    $pairsSet := pairsSet \setminus pair$
  end while
end function

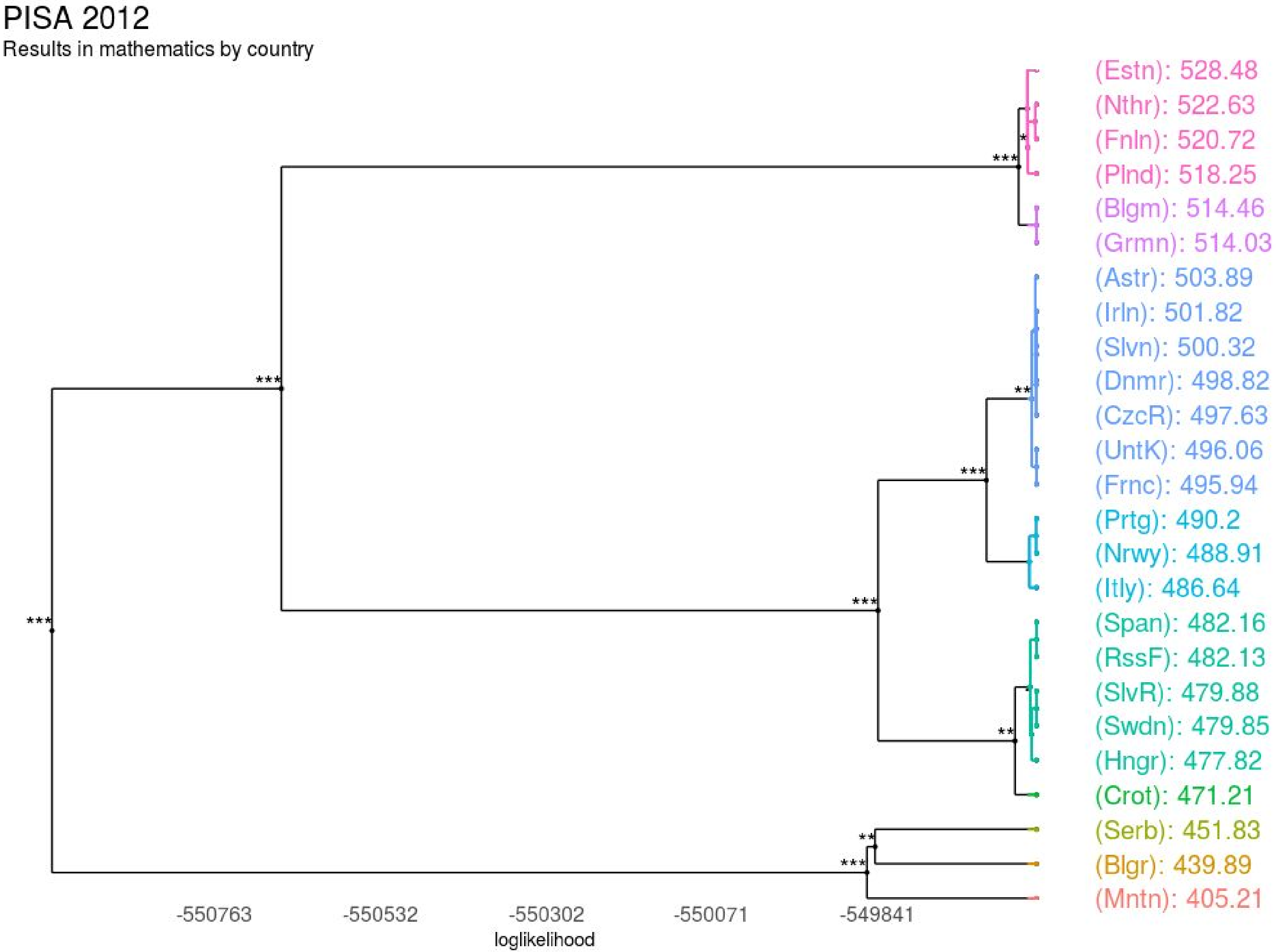
Merge

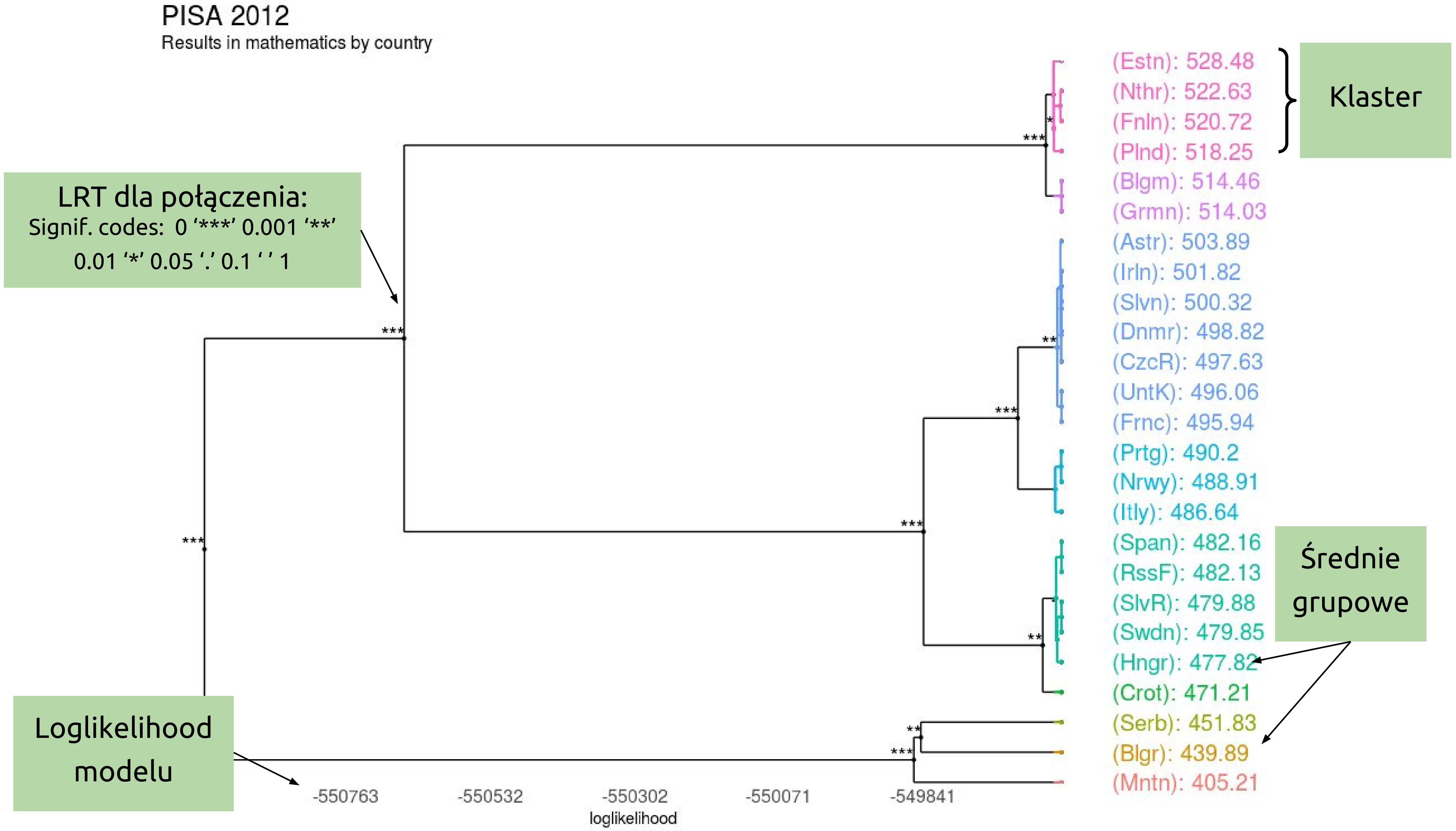
1. Testy ilorazu wiarygodności
2. Delete or Merge Regressors

Algorithm 2 Merging with agglomerative clustering

```
function MergeFactors(response, factor, successive)
    pairsSet := generatePairs(response, factor, successive)
    dist := set of distances
    for all pair ∈ pairsSet do
        h := {μ_pair1 = μ_pair2}                ▷ hypothesis under which pair is merged
        dist[pair] = LRT(M_h|M_0)
    end for
    if successive then
        hClust(dist, method = "single")
    else
        hClust(dist, method = "complete")
    end if
end function
```

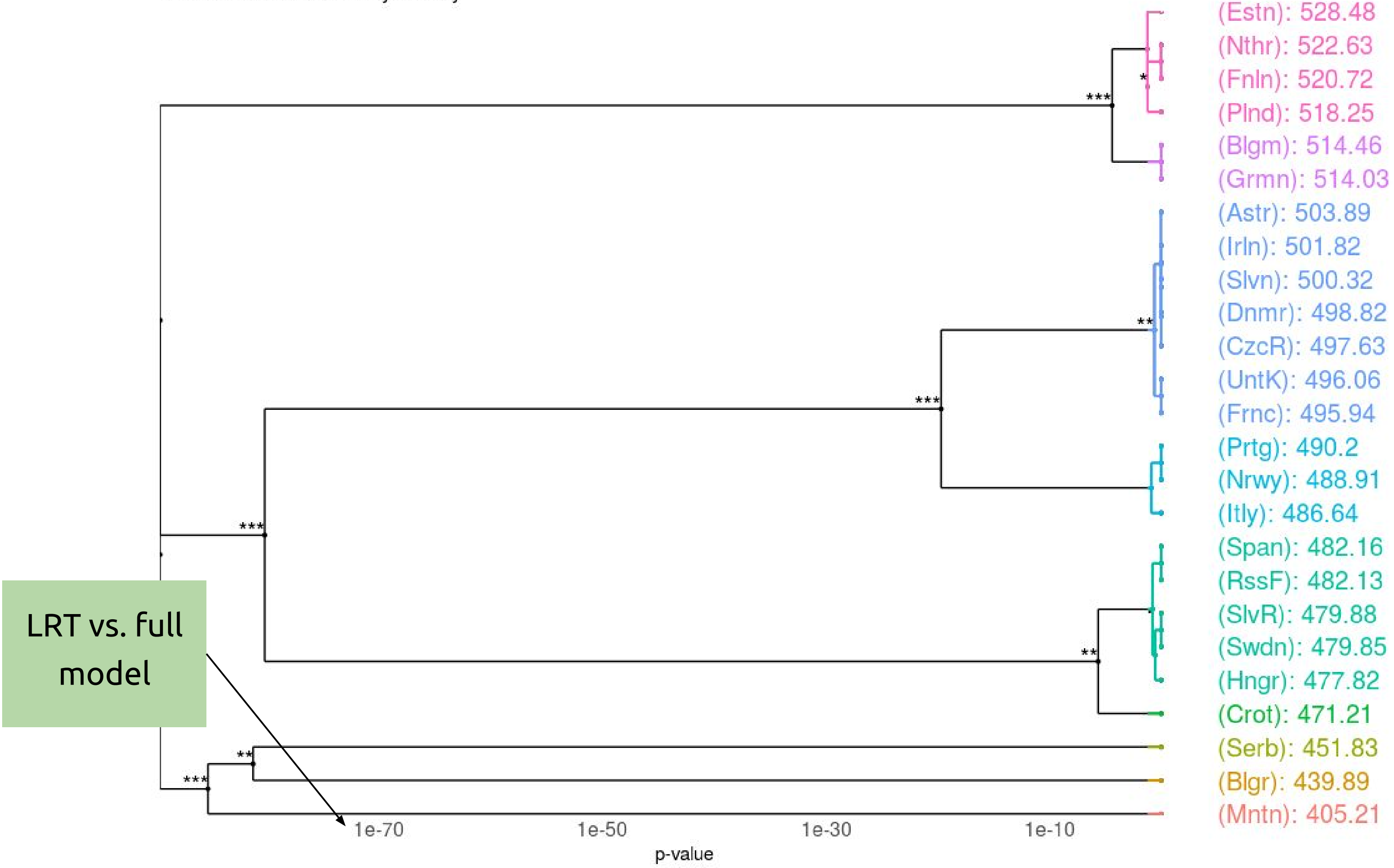
Więcej o algorytmie: https://arxiv.org/abs/1505.04008

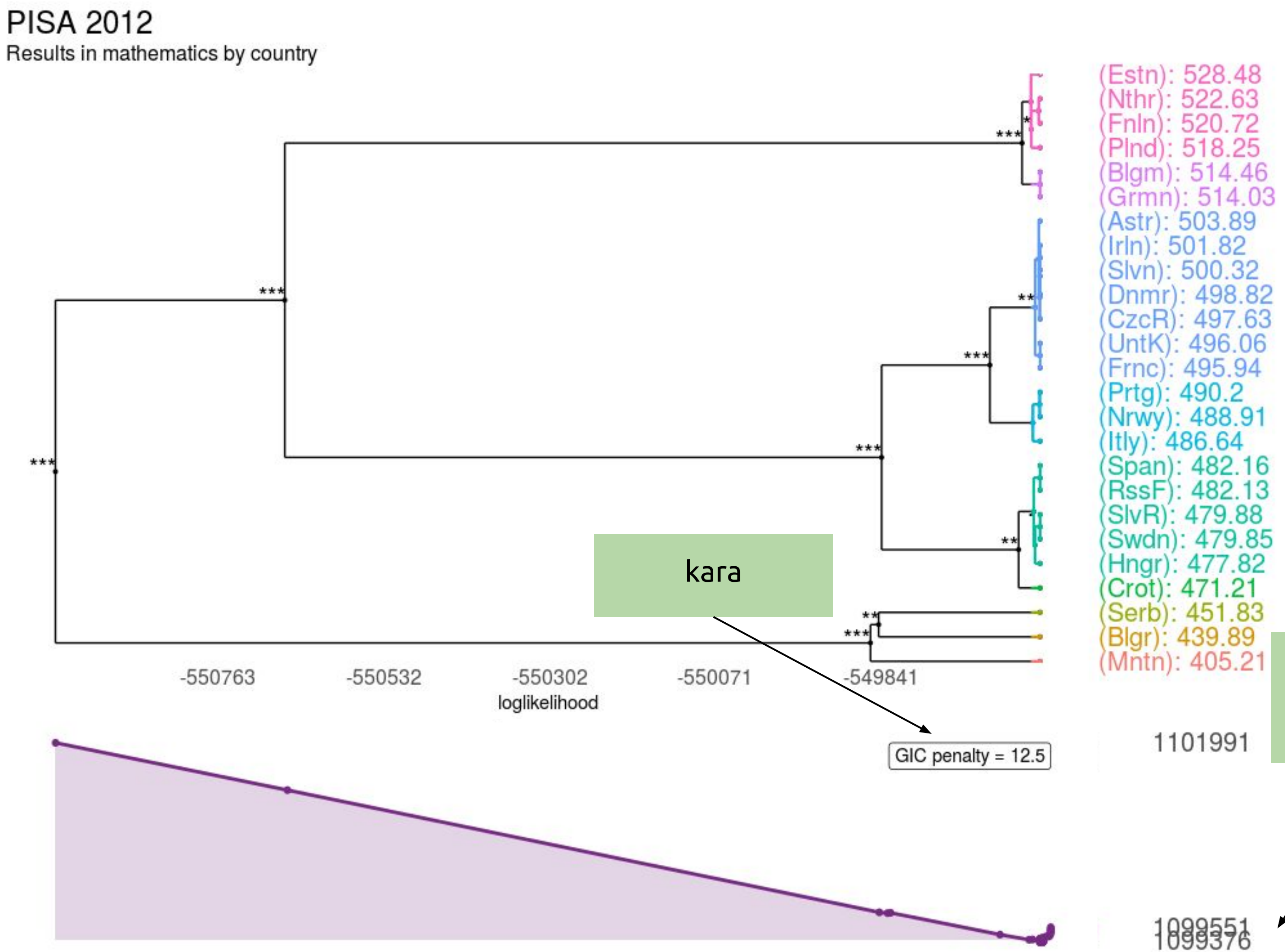

PISA 2012
Results in mathematics by country
(Estn): 528.48
(Nthr): 522.63
(Fnln): 520.72
(Plnd): 518.25
(Blgm): 514.46
(Grmn): 514.03
(Astr): 503.89
(Irln): 501.82
(Slvn): 500.32
(Dnmr): 498.82
(CzcR): 497.63
(UntK): 496.06
(Frnc): 495.94
(Prtg): 490.2
(Nrwy): 488.91
(Itly): 486.64
(Span): 482.16
(RssF): 482.13
(SlvR): 479.88
(Swdn): 479.85
(Hngr): 477.82
(Crot): 471.21
(Serb): 451.83
(Blgr): 439.89
(Mntn): 405.21
-550763
-550532
-550302
-550071
-549841
loglikelihood


PISA 2012
Results in mathematics by country
LRT dla połączenia:
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Klaster
Średnie grupowe
Loglikelihood modelu
(Estn): 528.48
(Nthr): 522.63
(Fnln): 520.72
(Plnd): 518.25
(Blgm): 514.46
(Grmn): 514.03
(Astr): 503.89
(Irln): 501.82
(Slvn): 500.32
(Dnmr): 498.82
(CzcR): 497.63
(UntK): 496.06
(Frnc): 495.94
(Prtg): 490.2
(Nrwy): 488.91
(Itly): 486.64
(Span): 482.16
(RssF): 482.13
(SlvR): 479.88
(Swdn): 479.85
(Hngr): 477.82
(Crot): 471.21
(Serb): 451.83
(Blgr): 439.89
(Mntn): 405.21
-550763
-550532
-550302
-550071
-549841
loglikelihood

PISA 2012

Results in mathematics by country



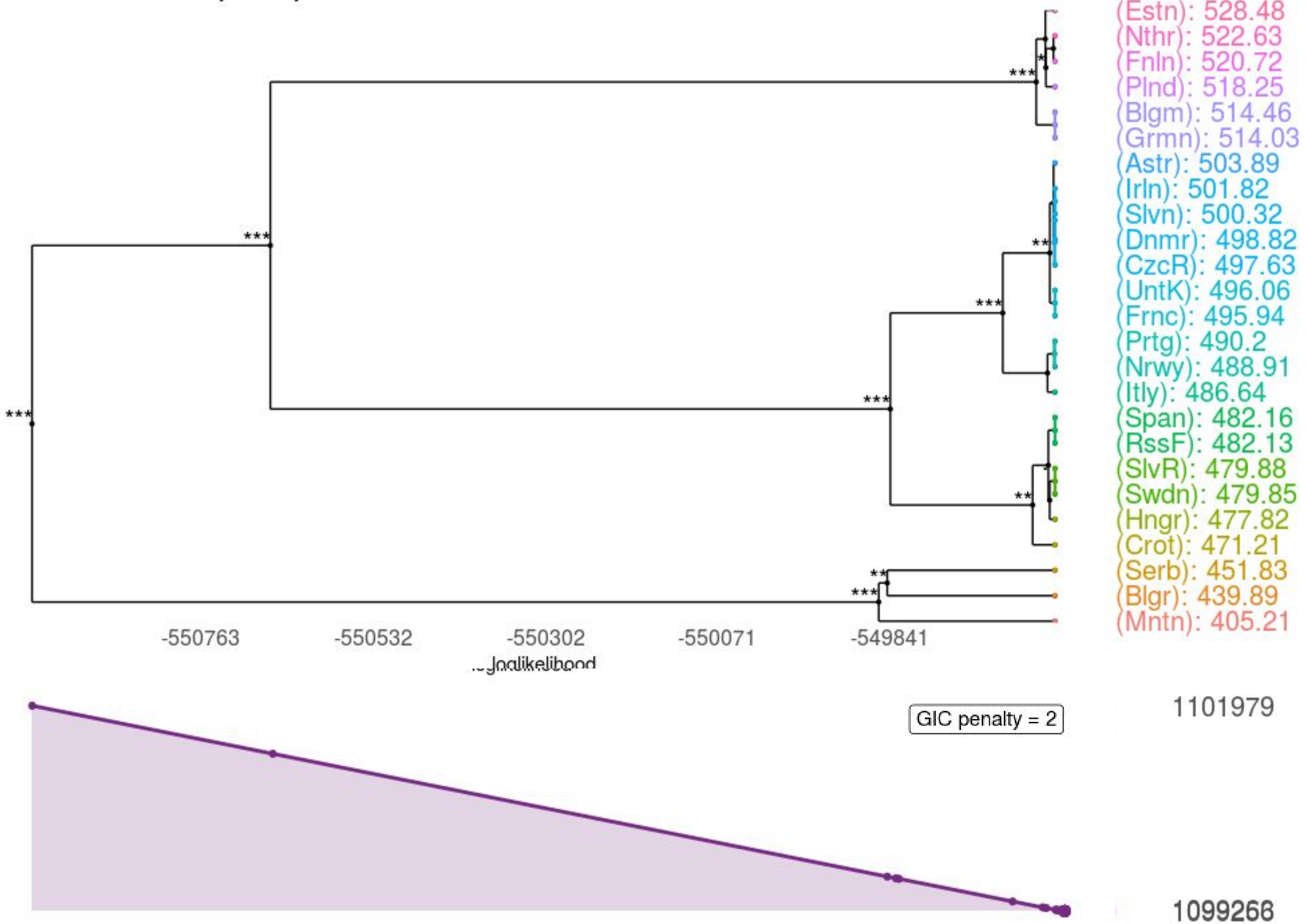


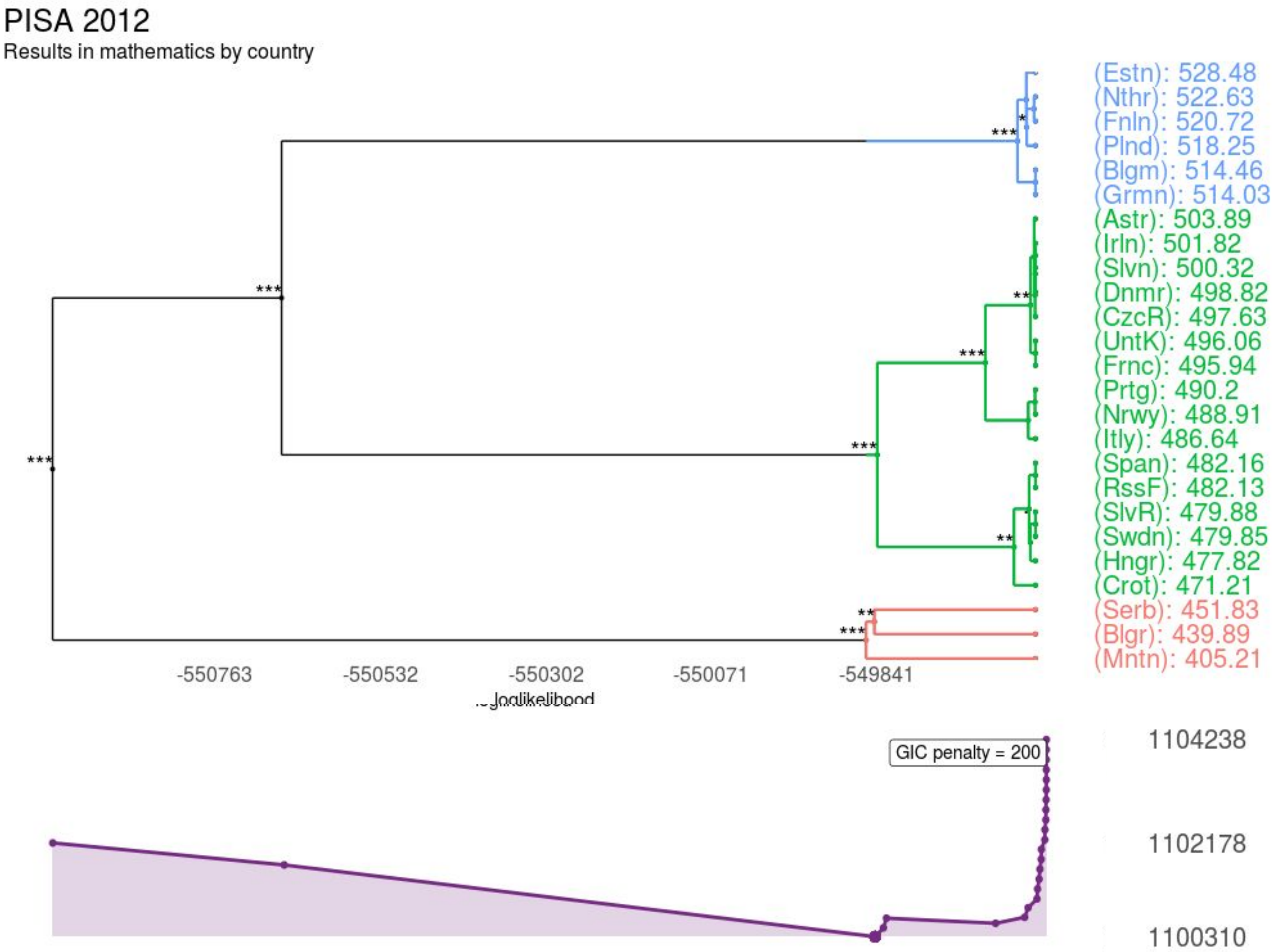
kara

Modele:
stały, pełny,
najlepszy

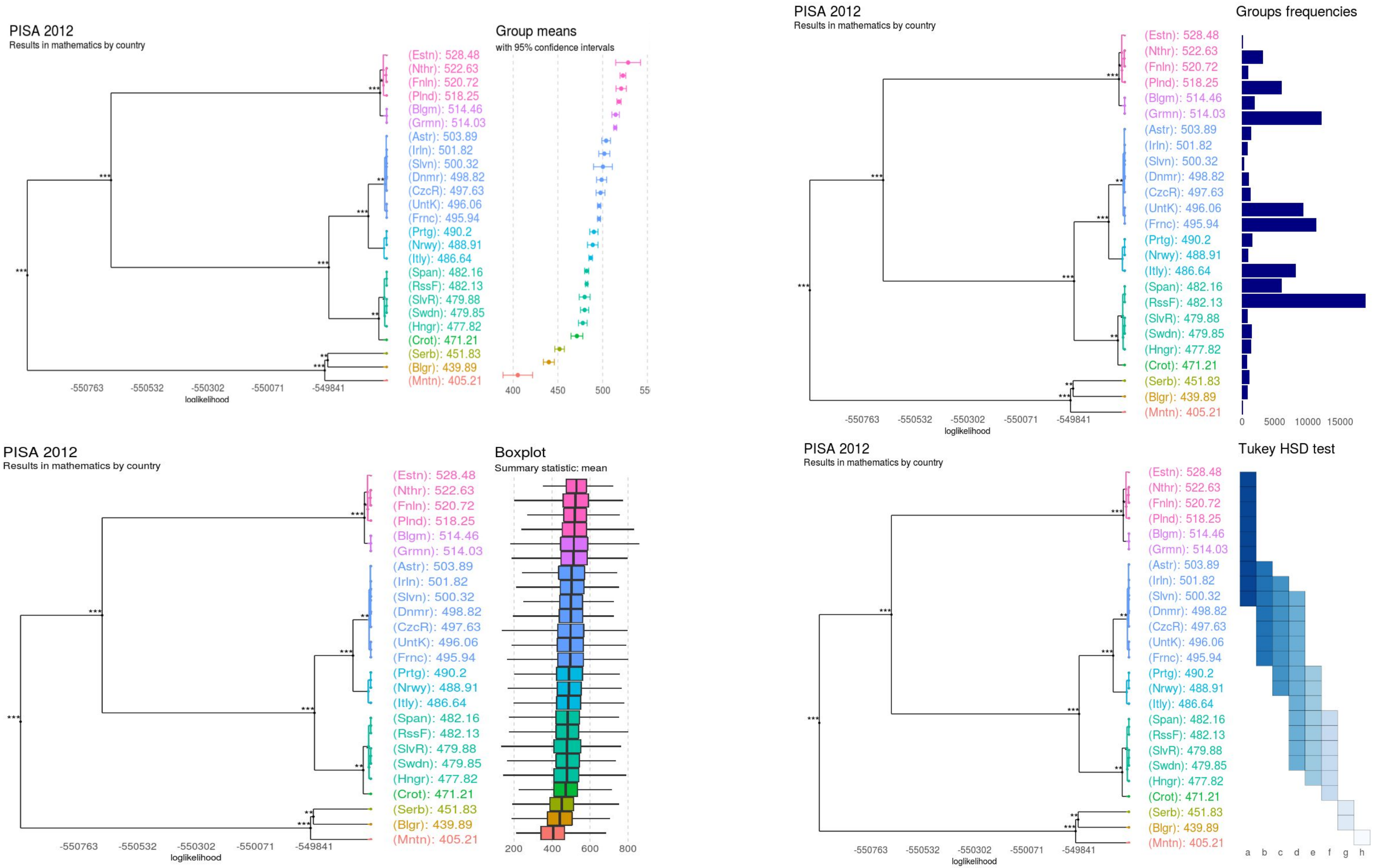
PISA 2012

Results in mathematics by country



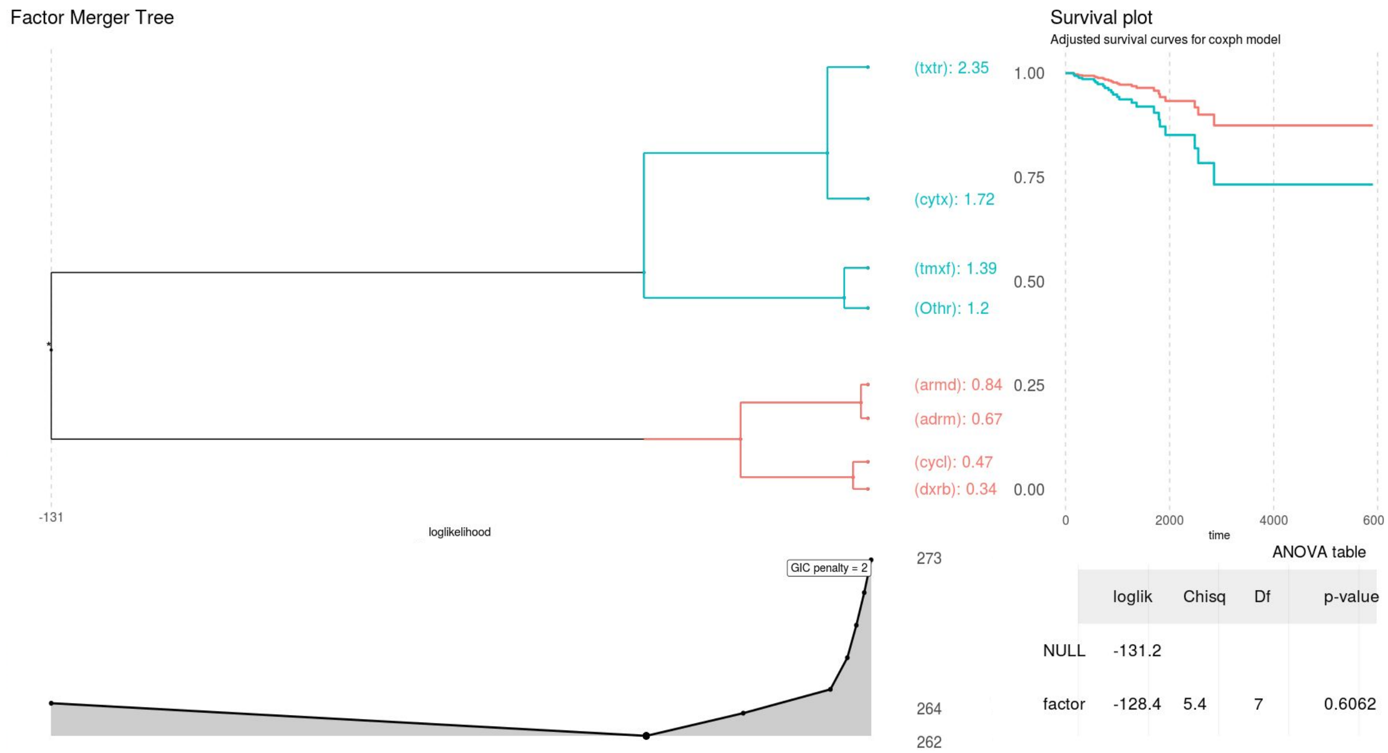


PISA 2012
Results in mathematics by country
(Estn): 528.48
(Nthr): 522.63
(Fnln): 520.72
(Plnd): 518.25
(Blgm): 514.46
(Grmn): 514.03
(Astr): 503.89
(Irln): 501.82
(Slvn): 500.32
(Dnmr): 498.82
(CzcR): 497.63
(UntK): 496.06
(Frnc): 495.94
(Prtg): 490.2
(Nrwy): 488.91
(Itly): 486.64
(Span): 482.16
(RssF): 482.13
(SlvR): 479.88
(Swdn): 479.85
(Hngr): 477.82
(Crot): 471.21
(Serb): 451.83
(Blgr): 439.89
(Mntn): 405.21
-550763
-550532
-550302
-550071
-549841
loglikelihood
GIC penalty = 200
1104238
1102178
1100310


PISA 2012
Results in mathematics by country
Group means
with 95% confidence intervals
Groups frequencies
Boxplot
Summary statistic: mean
Tukey HSD test
(Estn): 528.48
(Nthr): 522.63
(Fnln): 520.72
(Plnd): 518.25
(Blgm): 514.46
(Grmn): 514.03
(Astr): 503.89
(Irln): 501.82
(Slvn): 500.32
(Dnmr): 498.82
(CzcR): 497.63
(UntK): 496.06
(Frnc): 495.94
(Prtg): 490.2
(Nrwy): 488.91
(Itly): 486.64
(Span): 482.16
(RssF): 482.13
(SlvR): 479.88
(Swdn): 479.85
(Hngr): 477.82
(Crot): 471.21
(Serb): 451.83
(Blgr): 439.89
(Mntn): 405.21
-550763 -550532 -550302 -550071 -549841
loglikelihood
400 450 500 550
0 5000 10000 15000
200 400 600 800
a b c d e f g h

Nie tylko jednowymiarowy Gauss

1. Wielowymiarowy Gauss
2. Regresja logistyczna
3. Analiza przeżycia



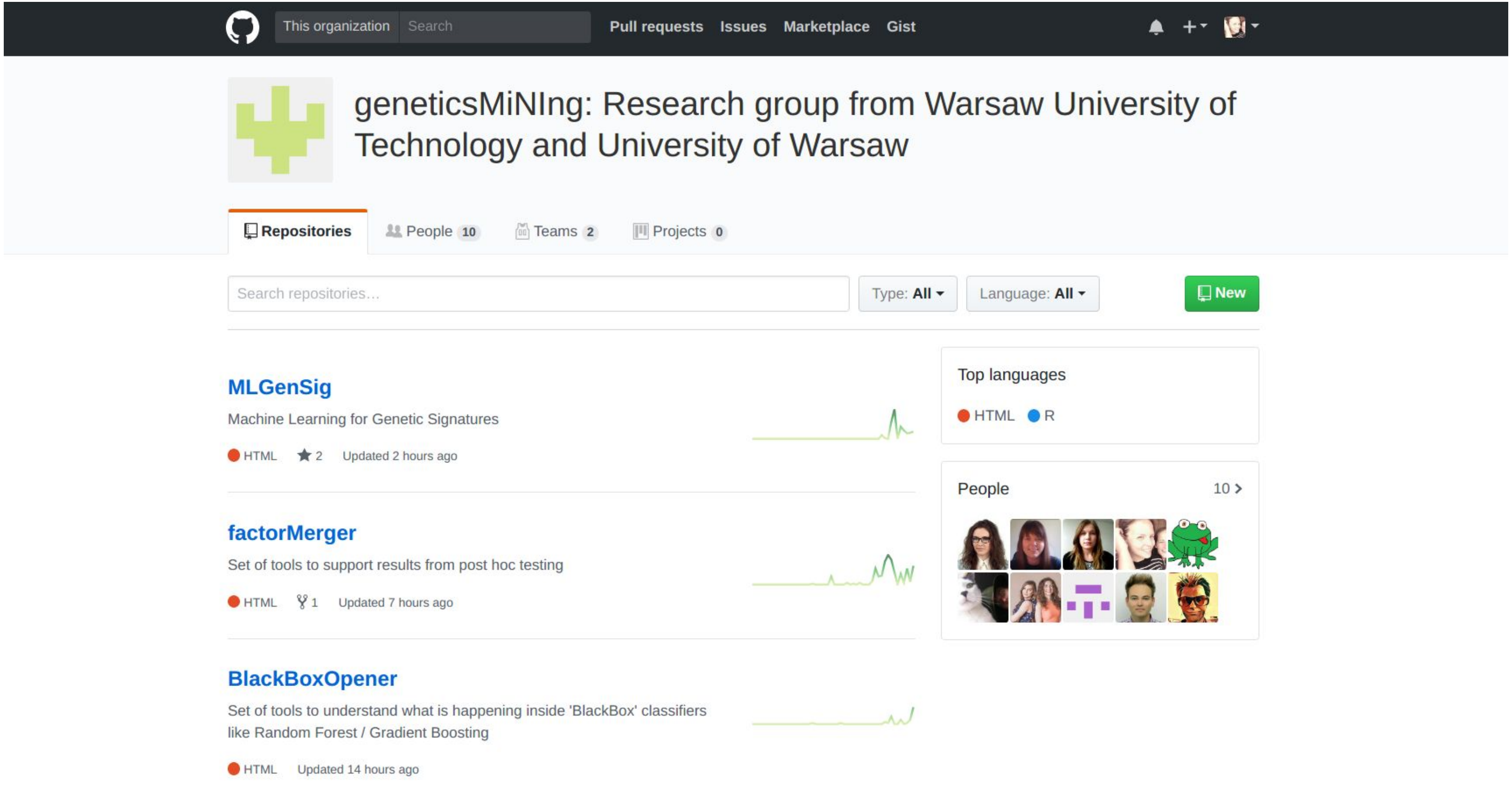
Posumowując

```r
devtools::install_github("geneticsMiNIng/factorMerger")
library(factorMerger)

fm <- mergeFactors(response = myResponse,
                   factor = myFactor,
                   successive = TRUE,
                   method = "hclust",
                   family = "binomial")
plot(fm)
```

Więcej: https://github.com/geneticsMiNIng/factorMerger

geneticsMiNIng



Dziękuję za uwagę

Agnieszka Sitko

ag.agnieszka.sitko@gmail.com

25.05.2017