MINISTERSTVO VNÚTRA SLOVENSKEJ REPUBLIKY
Operačný program Efektívna verejná správa
Európska únia
Európsky sociálny fond

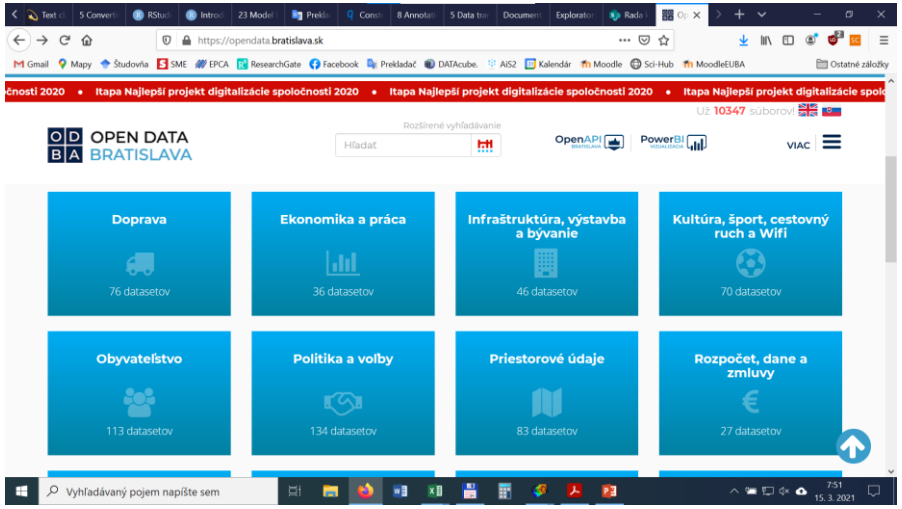
ESDA - Exploratívna analýza priestorových dát v GeoDa

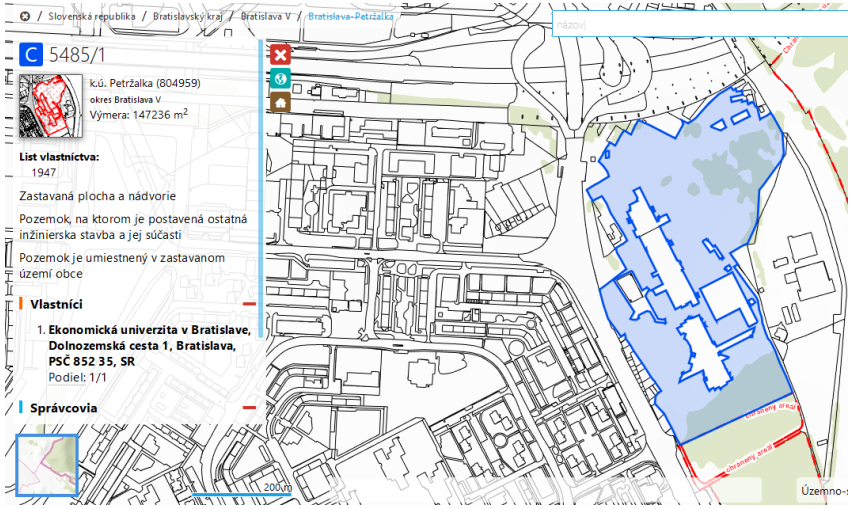
Štefan Rehák

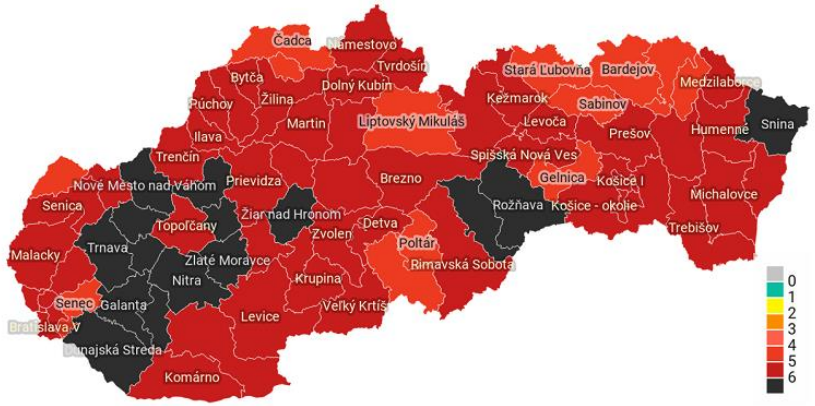
Katedra verejnej správy a regionálneho rozvoja

Národohospodárska fakulta EU v Bratislave

Priestorové údaje všade okolo nás





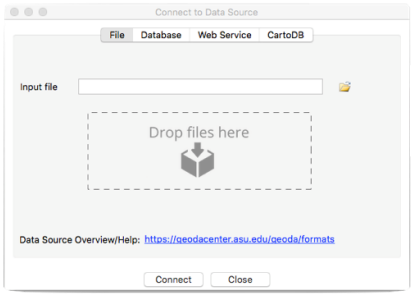


GeoDa

- **GeoDa** je otvorený softvér na exploratívnu analýzu priestorových údajov
  - **E**xploratory **S**patial **D**ata **A**nalysis **ESDA**
- ESDA je súčasť Exploratory Data Analysis (**EDA**) čo je jedna z častí dátovej vedy pričom dáva dôraz na priestorový aspekt spoločenských, ekonomických a environmentálnych procesov
  - Kľúčovým atribútom je preto umiestnenie, vzdialenosť a interakcia
- Je vyvíjaný od roku 2003 v ***Center for Spatial Data Science*** na *University of Chicago* pod vedením Luc Anselin a súčasne ho používa asi **360 tisíc užívateľov** po celom svete, tak praktikov ako aj výskumných tímov (Harvard, MIT, a Cornell)

Platformy a dátové formáty

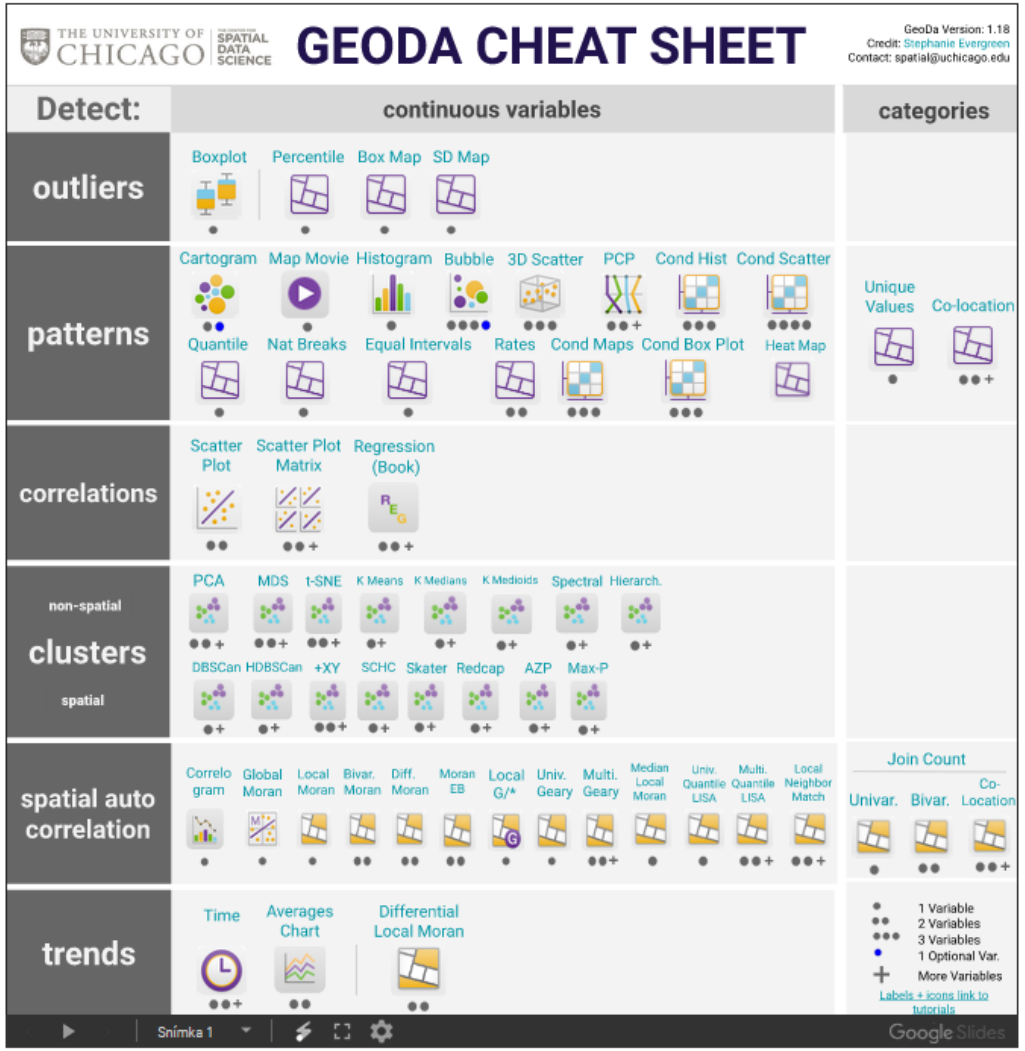
- **Point and click** formát (ale aj R rozhranie)
- Program beží na viacerých **platformách**: Windows, MacOSX a Linux (Ubuntu)
- Dátové **formáty**:
  - shapefile, geodatabase, GeoJSON, MapInfo, GML, KML
  - Konvertovanie koordinátov z tabuľkových súborov (.csv, .dbf, .xls, .ods)
  - Konvertovanie medzi rôznymi typmi súborov (csv do .dbf alebo shapefile do GeoJSON)



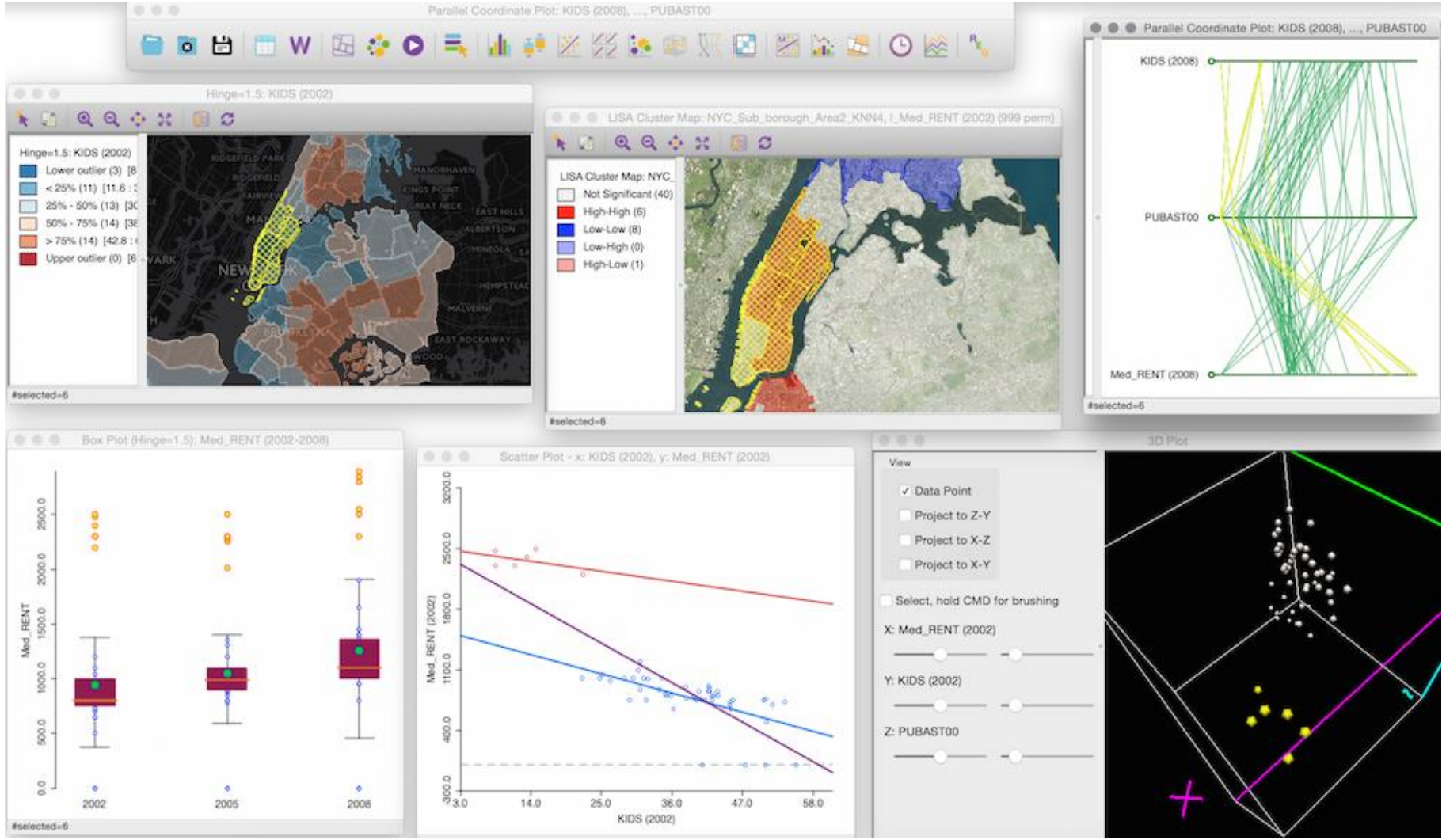
Prehľad funkcionalít

- Detekcia extrémnych hodnôt
- Analýza vzorov
- Korelačná analýza
- Priestorové a nepriestorové klastre
- Priestorová autokorelácia
- Časové trendy

- Modelovanie (regresie, priestorové regresie)



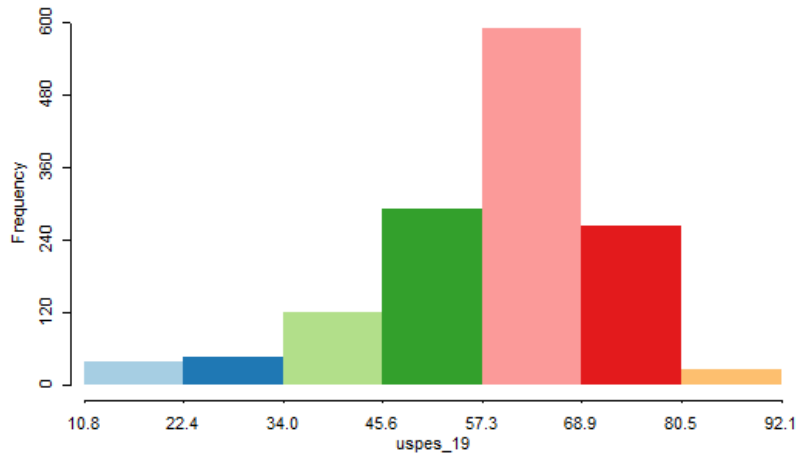
Kľúčový koncept - linking and brushing



Ilustratívny príklad

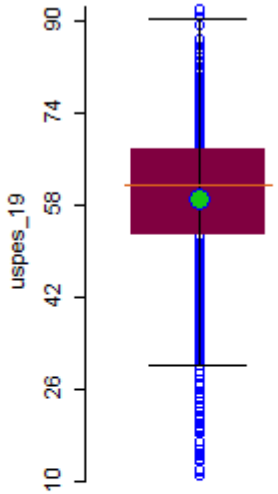
- Exploratívna analýza kvality v školstve na základe testovania deviatakov (T9) v matematike
- Dáta 2014 – 2019:
  - **NUCEM**: veľkosť triedy, počet zdravotne znevýhodnených, počet zo znevýhodneného sociálneho prostredia, známka, úspešnosť MAT, percentil
  - **Register školských zariadení**: vlastníctvo (štátna, cirkevná, súkromná), jazyk (slovenský, iný), typ (ZŠ, iná)
- Dáta sú v skutočnosti o kvalite deviatakov, nie o kvalite samotných škôl
- **Geokódovanie** dát:
  - Na základe adresy školy – QGIS
- Finálna vzorka **1448 škôl** (to je náhoda!)

Analýza distribúcie (histogram, boxplot a mapa)

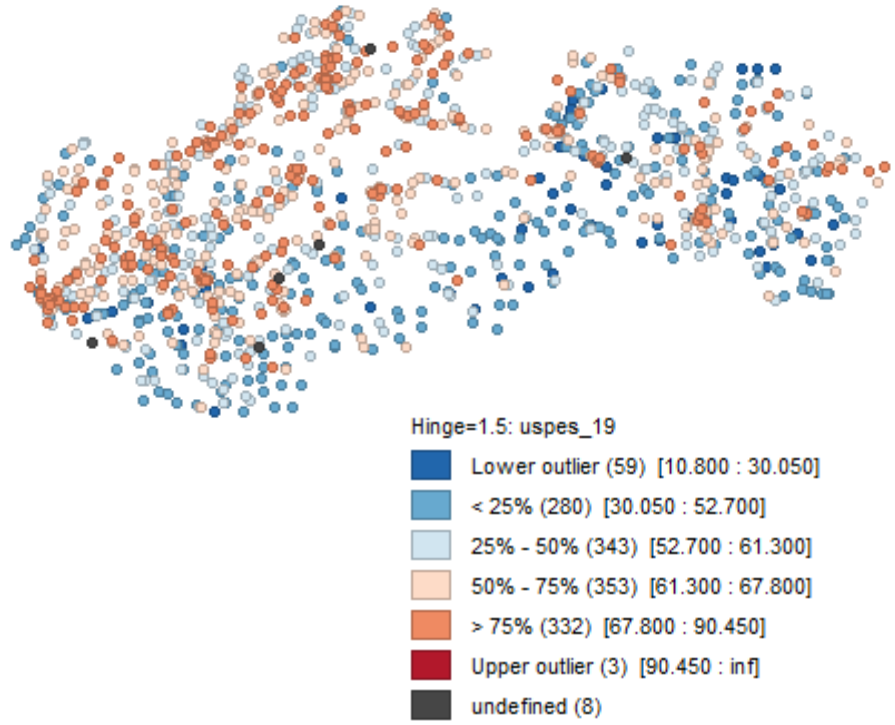


from	10.800	22.414	34.029	45.643	57.257	68.871	80.486
to	22.414	34.029	45.643	57.257	68.871	80.486	92.100
#obs	37	45	118	290	592	263	25
% of total	2.685	3.266	8.563	21.045	42.961	19.086	1.814
sd from mean	-2.714	-1.854	-0.994	-0.134	0	0.726	1.586

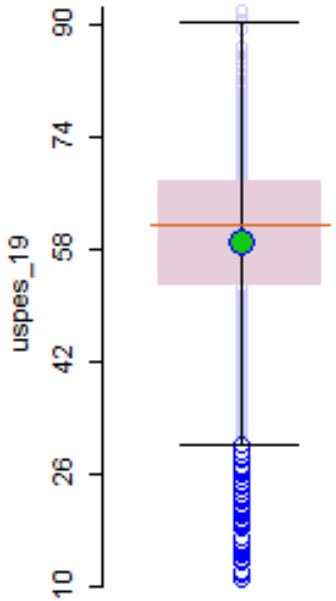
min: 10.8, max: 92.1, median: 61.3, mean: 59.0624, s.d.: 13.5049, #obs: 1378



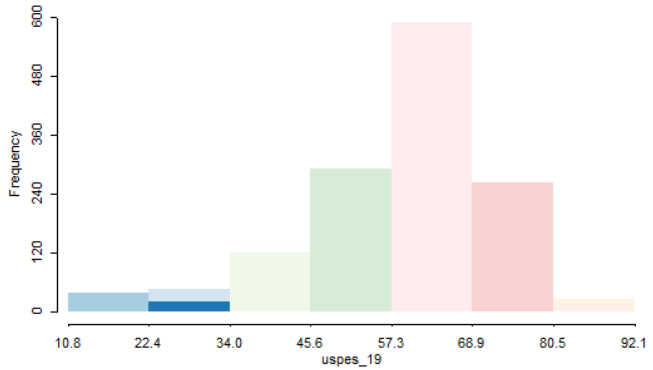
	uspes_19
min	10.8000
max	92.1000
Q1	52.7000
median	61.3000
Q3	67.8000
IQR	15.1000
mean	59.0624
s.d.	13.5049



Analýza distribúcie – linking



	uspes_19
min	10.8000
max	92.1000
Q1	52.7000
median	61.3000
Q3	67.8000
IQR	15.1000
mean	59.0624
s.d.	13.5049



from	10.800	22.414	34.029	45.643	57.257	68.871	80.486
to	22.414	34.029	45.643	57.257	68.871	80.486	92.100
#obs	37	45	118	290	592	263	25
% of total	2.685	3.266	8.563	21.045	42.961	19.086	1.814
sd from mean	-2.714	-1.854	-0.994	-0.134	0	0.726	1.586

min: 10.8, max: 92.1, median: 61.3, mean: 59.0624, s.d.: 13.5049, #obs: 1378

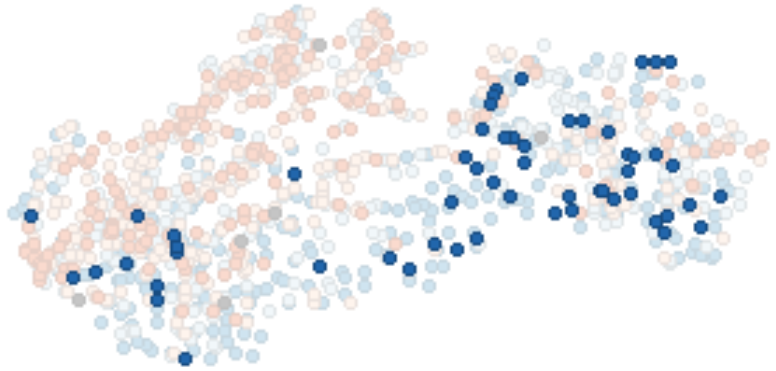
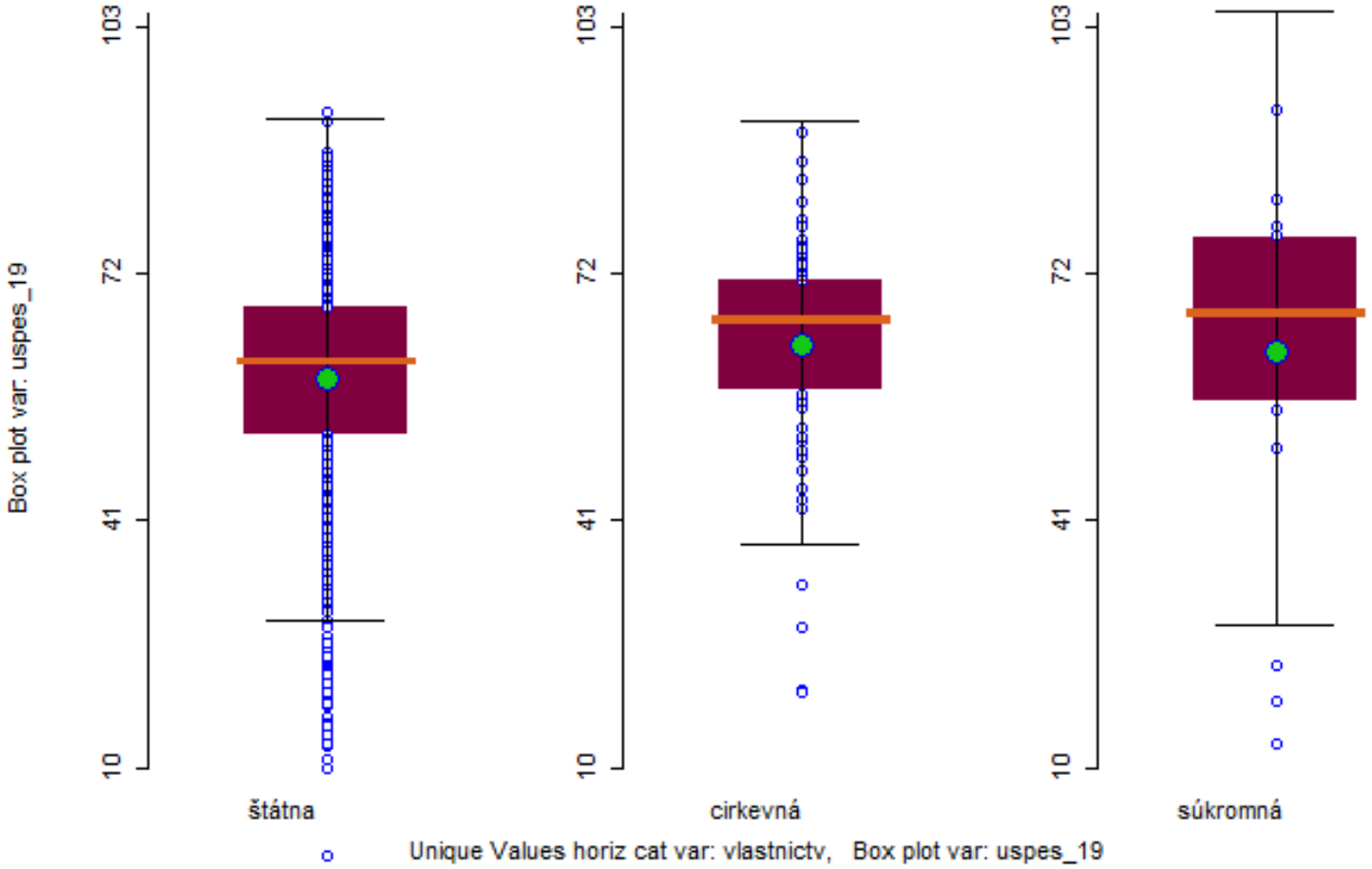


Table - MAT9_2014_19_FINAL

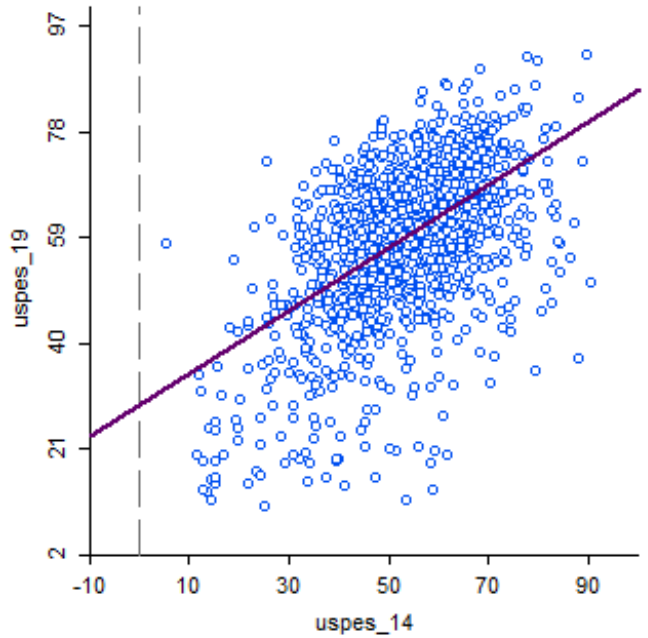
	skola	nazov	ulica_cisl	obec
1251	580608	ZŠ	Ďurkov 192	Ďurkov
1253	580610	ZŠ, Školská 3, Jasov	Školská 3	Jasov
1254	580611	ZŠ	Kecerovce 79	Kecerovce
1259	580617	ZŠ a gymnázium s VJM	Československej armády 1!	Moldava nad
1269	580627	ZŠ s VJM - Alapiskola	Školská ulica 301/12	Turňa nad Bo
1297	580725	CZŠ s VJM VOJANY	Elektrárenská 50	Vojany
1299	580727	ZŠ s MŠ	Žbince 145	Žbince
1305	580809	ZŠ s MŠ	Letná 14	Nižná Slaná
1308	580812	ZŠ	Rejdová 43	Rejdová
1310	580815	Spojená škola, ZS s VJM a SOS s VJM	Komenského 5	Rožňava
1318	580902	ZŠ	Blatné Remety 98	Blatné Remet
1330	581006	ZŠ s MŠ	Školská 16	Markušovce
1348	581037	ZŠ sv. Michala	Školská 1	Spišské Tomá
1361	581113	ZŠ	Školská 58	Nižný Žipov
1371	581125	ZŠ	Ivana Krasku 342/1	Trebišov
1377	581137	ZŠ s MŠ	Hlavná 75	Hrčeľ
1	510101	ZŠ s MŠ M.R.Štefánika	Grösslingová 48	Bratislava
2	510102	ZŠ	Mudroňova 83	Bratislava
3	510105	ZŠ	Hlboká cesta 4	Bratislava
4	510106	ZŠ	Jelenia 16	Bratislava

#row=1378 #selected=59

Analýza distribúcie – conditional plot

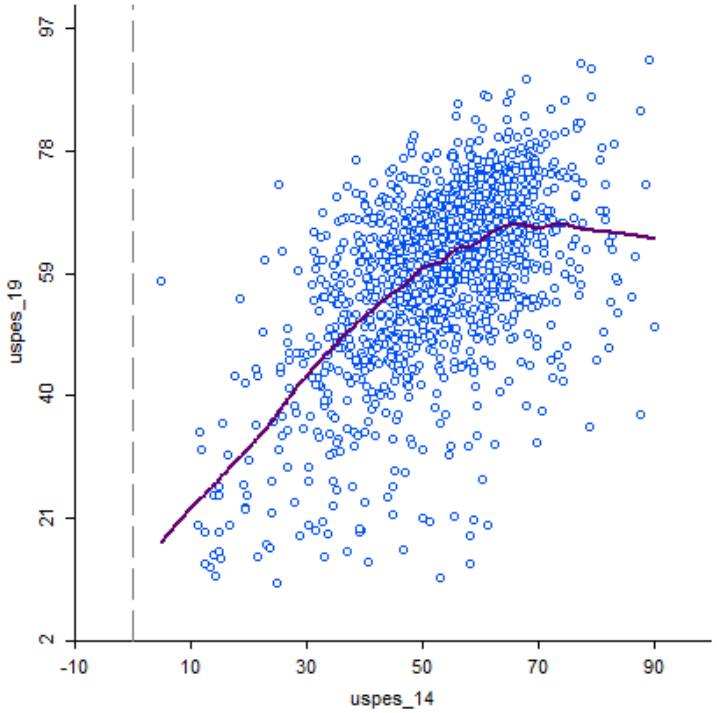


Analýza závislostí – scatter plot (+ LOWESS)

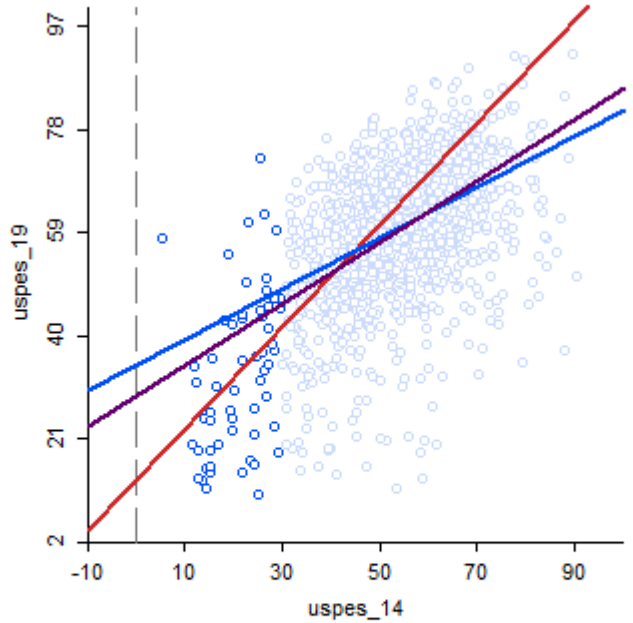


#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
1370	0.296	28.939	1.293	22.379	0	0.568	0.024	23.978	0
0	0	0	0	0	0	0	0	0	0
1370	0.296	28.939	1.293	22.379	0	0.568	0.024	23.978	0

Chow test for sel/unsel regression subsets: need two valid regressions

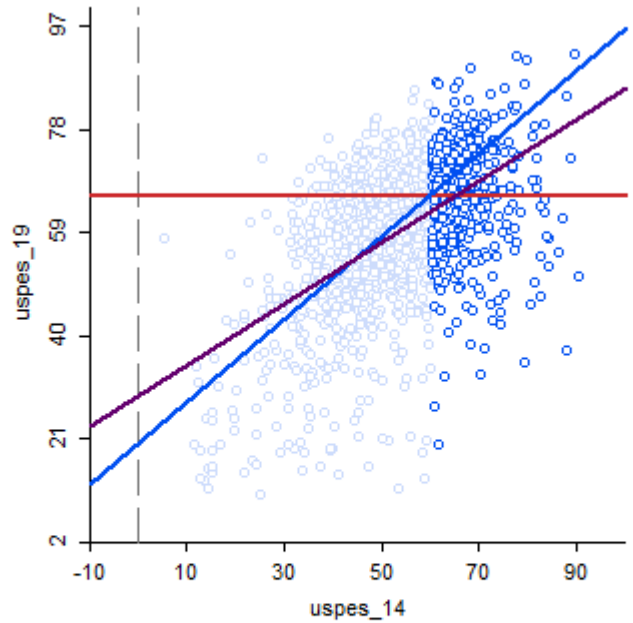


Analýza závislostí – scatter plot (brushing)



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
1370	0.296	28.939	1.293	22.379	0	0.568	0.024	23.978	0
60	0.144	13.354	6.667	2.003	0.050	0.947	0.303	3.126	0.003
1310	0.185	34.566	1.520	22.745	0	0.471	0.027	17.243	0

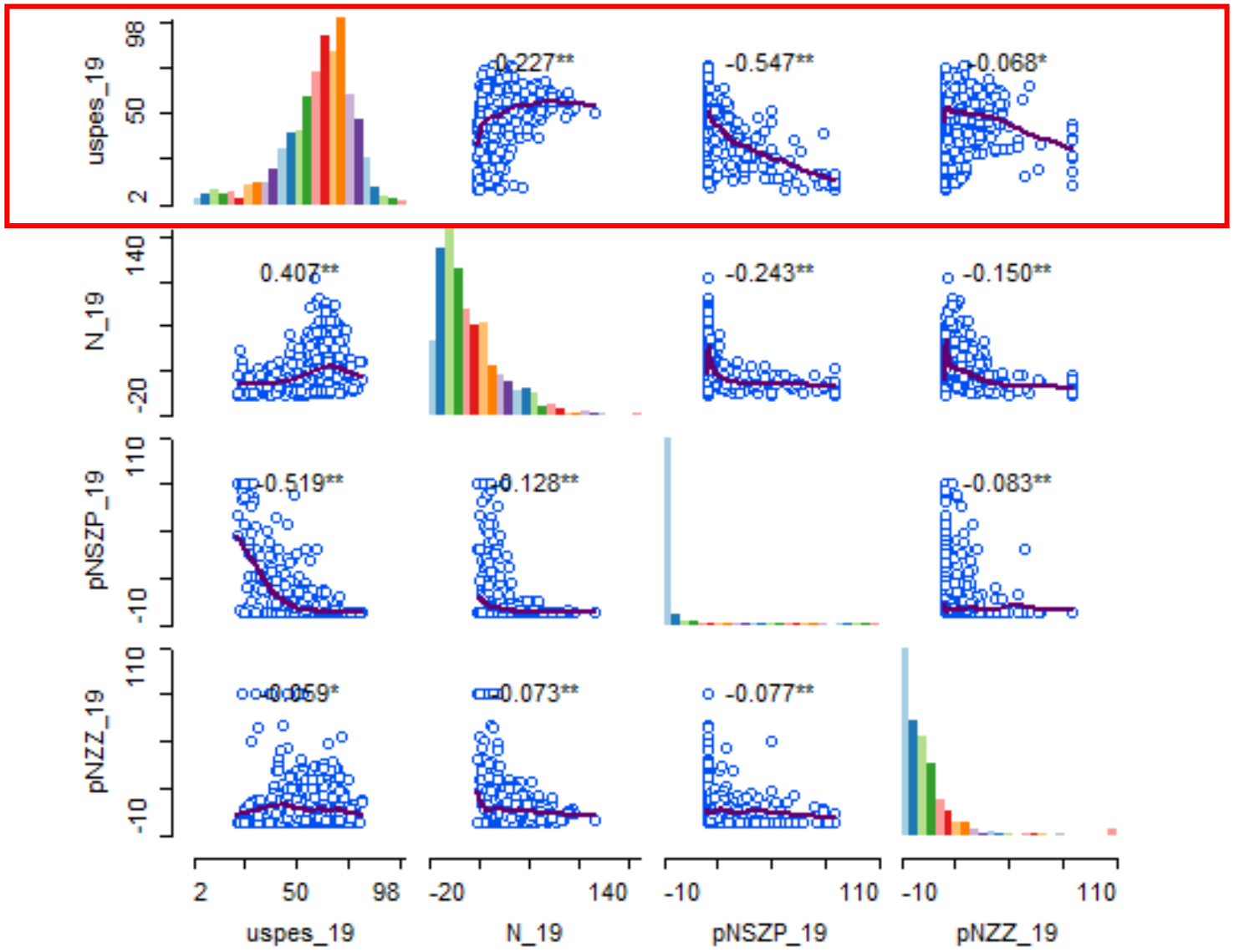
Chow test for sel/unsel regression subsets: distrib=F(2,1374), ratio=21.7363, p-val=0.0000



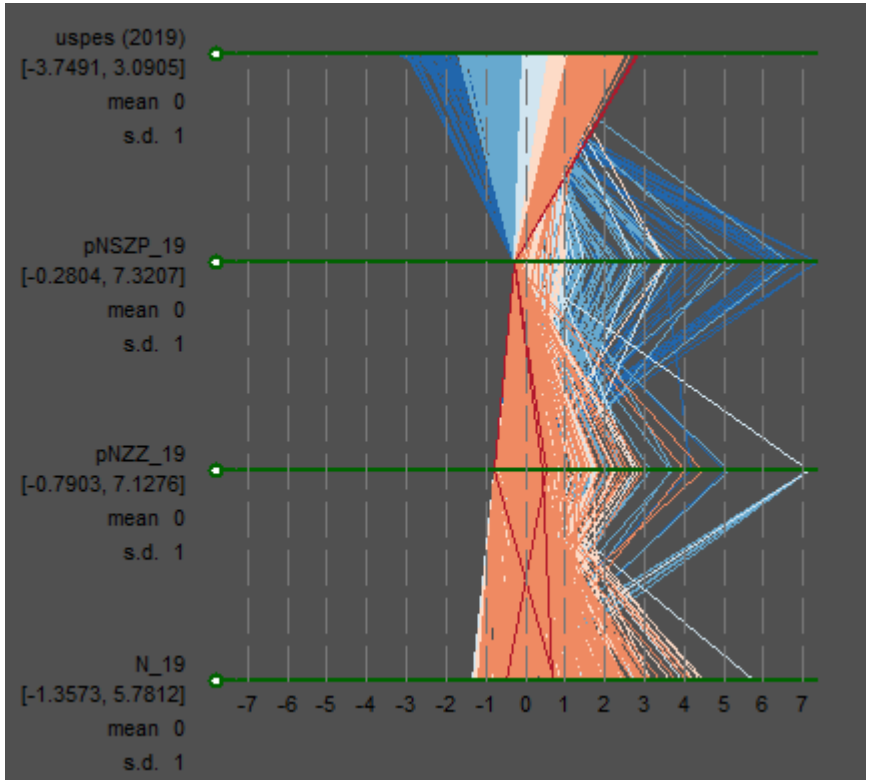
#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
1370	0.296	28.939	1.293	22.379	0	0.568	0.024	23.978	0
414	0.000	65.598	5.607	11.698	0	0.005	0.083	0.060	0.952
956	0.313	20.215	1.759	11.493	0	0.764	0.037	20.847	0

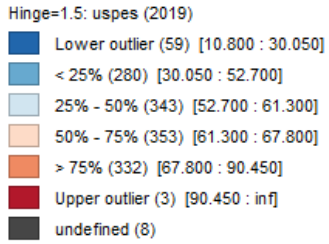
Chow test for sel/unsel regression subsets: distrib=F(2,1374), ratio=39.0563, p-val=0

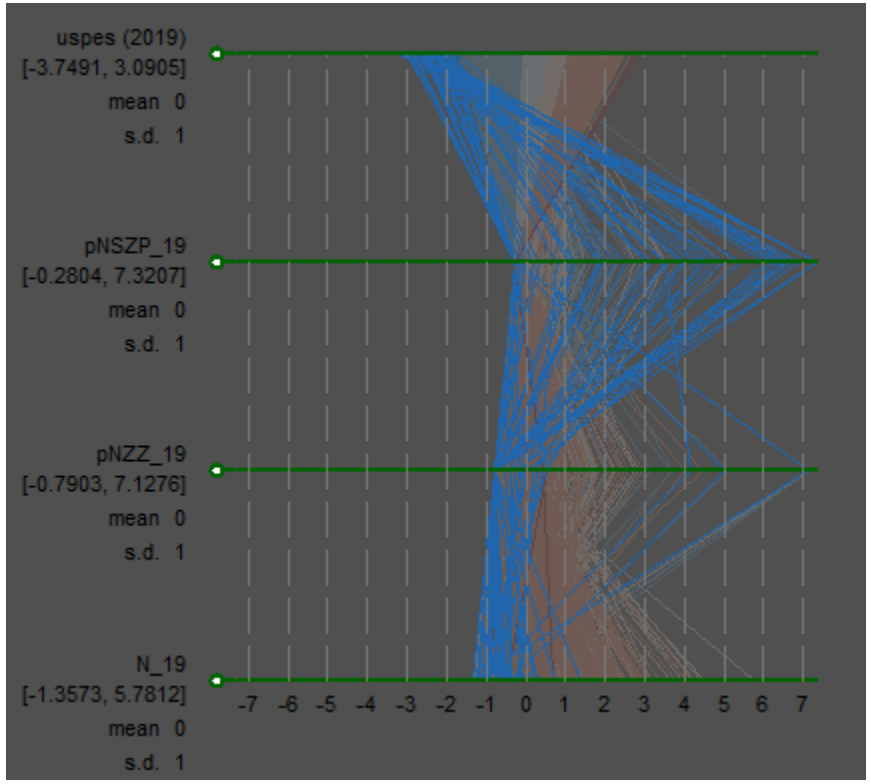
Analýza závislostí - Scatter plot matrix



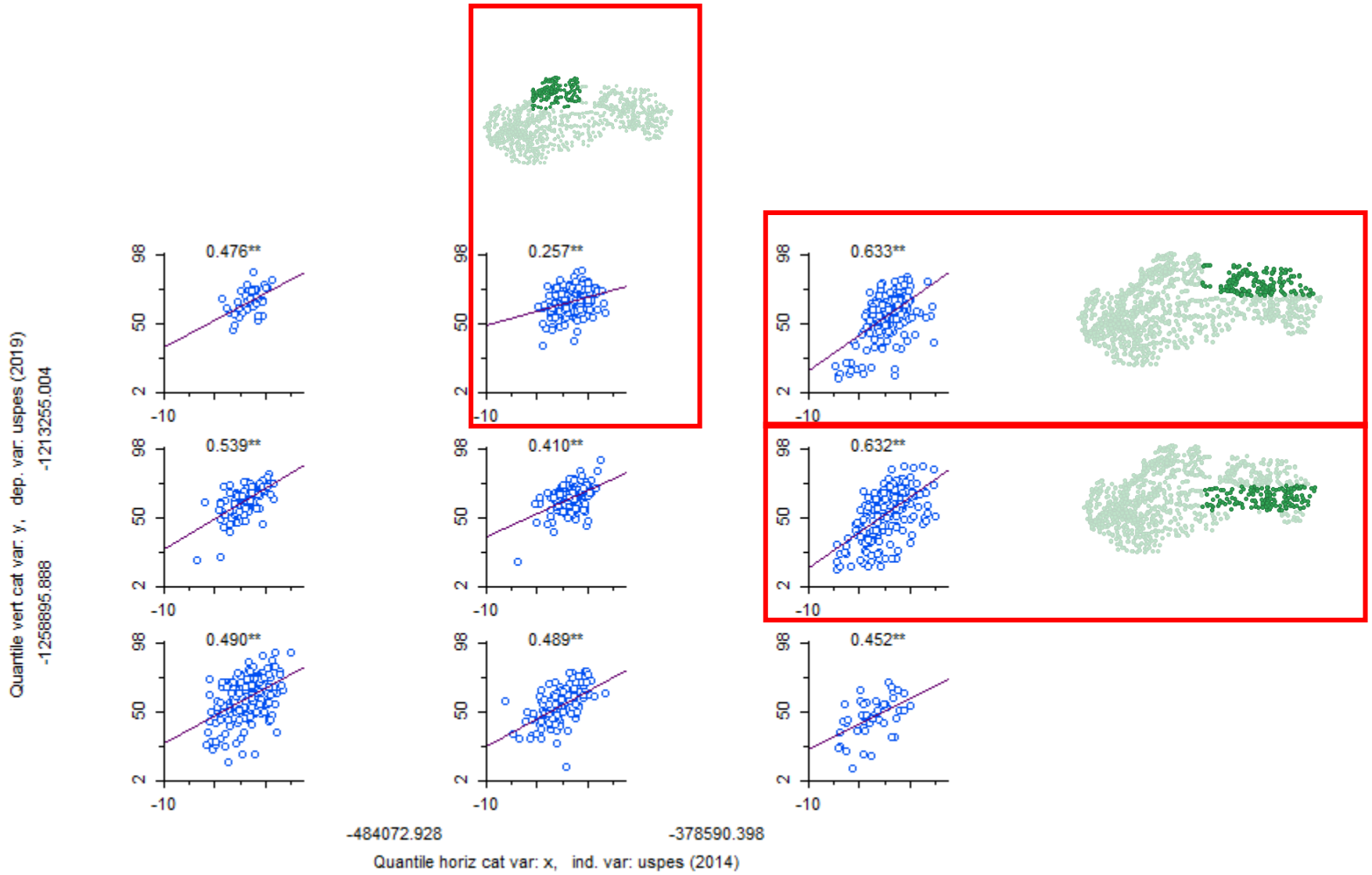
Analýza závislostí – conditional plot



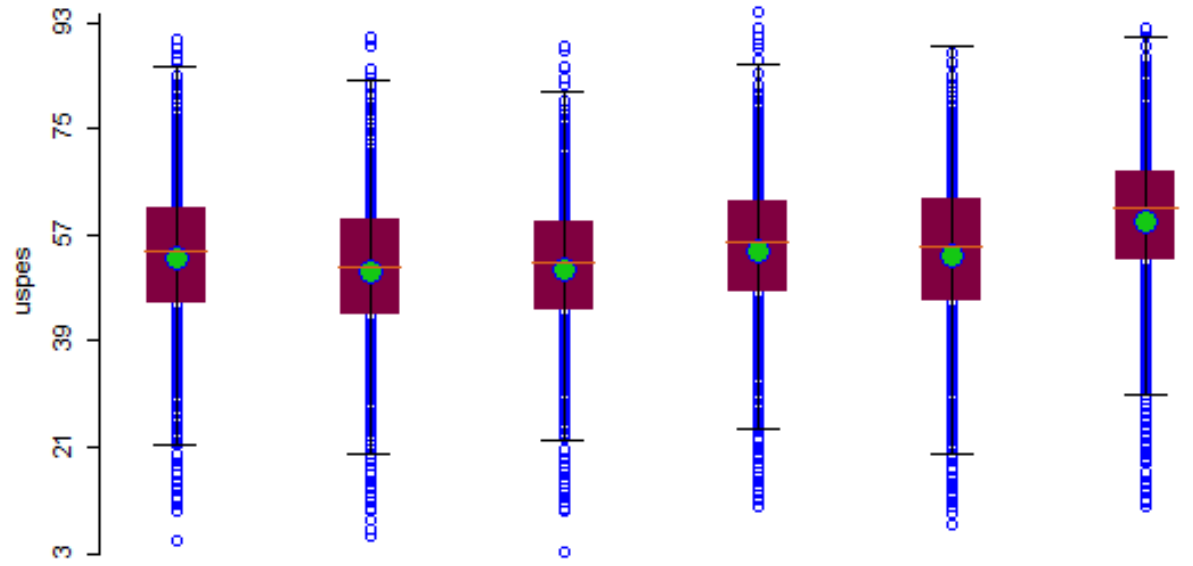




Analýza závislostí - Conditional plot



Časová analýza (porovnanie distribúcie)



	2014	2015	2016	2017	2018	2019
min	5	5.8333	3.3000	10.8000	7.8000	10.8000
max	90	90.5128	89	95	87.7000	92.1000
Q1	45.5000	43.5208	44.3000	47.1500	45.7000	52.7000
median	53.8516	51.4634	52.2000	55.7000	55	61.3000
Q3	61.5789	59.4444	59.1000	62.6500	63	67.8000
IQR	16.0789	15.9236	14.8000	15.5000	17.3000	15.1000
mean	53.0307	50.6719	50.9842	54.1302	53.4731	59.0624
s.d.	12.9914	13.0354	12.3791	13.1066	13.7264	13.5049

Časová analýza outlierov (averages chart)

Variable: uspes (2014-2019)

Groups: Selected vs. Unselected

Difference-in-Means Test:

Group 1: Selected Period 1: 2014

Group 2: Unselected Period 2: 2014

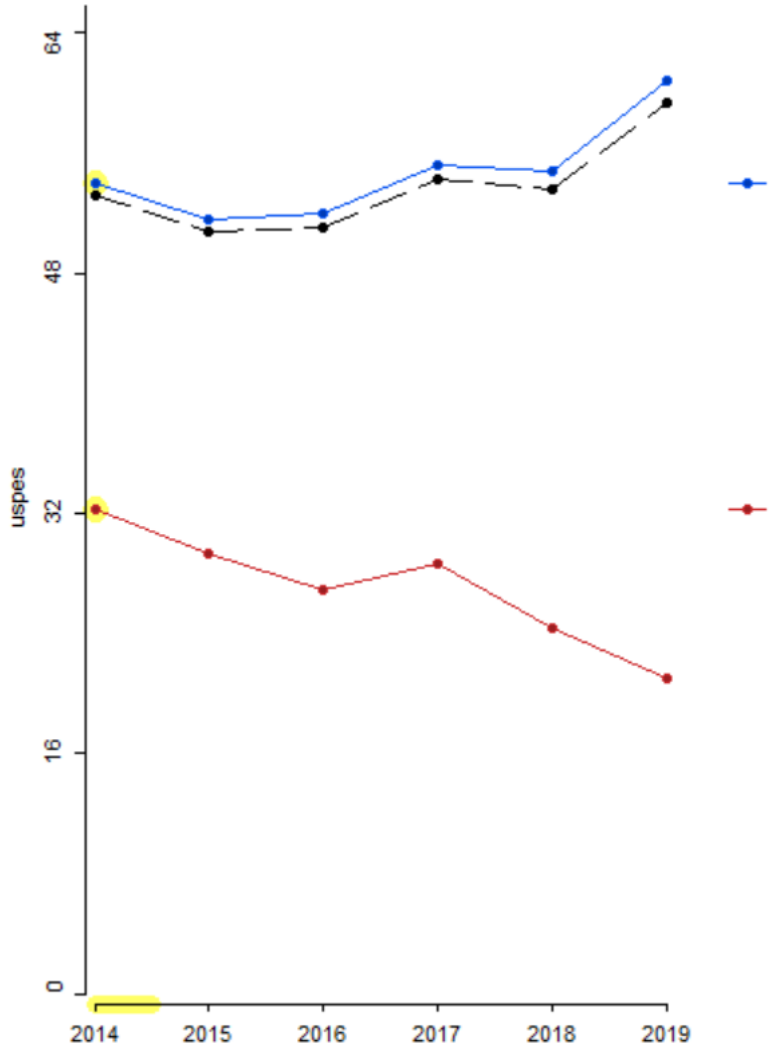
Run Diff-in-Diff Test | Save Dummy

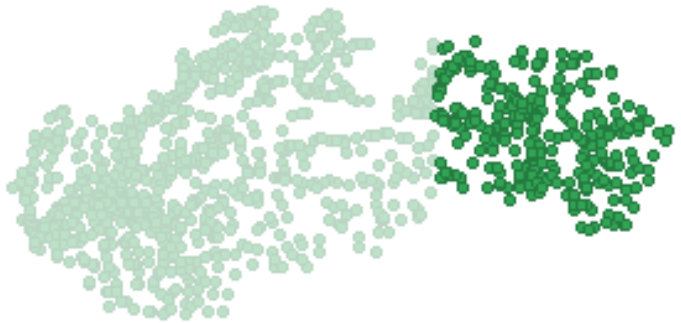
☐ Save Test Results

Group	Obs.	Mean	S.D.
Selected	51	32.22	15.04
Unselected	1300	54.00	12.06

Do Means Differ? (ANOVA)

D.F.	1349
F-val	156.79
p-val	0.000





Variable: uspes (2014-2019)

Groups: Selected vs. Unselected

Difference-in-Means Test:

Group 1: Selected Period 1: 2014

Group 2: Unselected Period 2: 2014

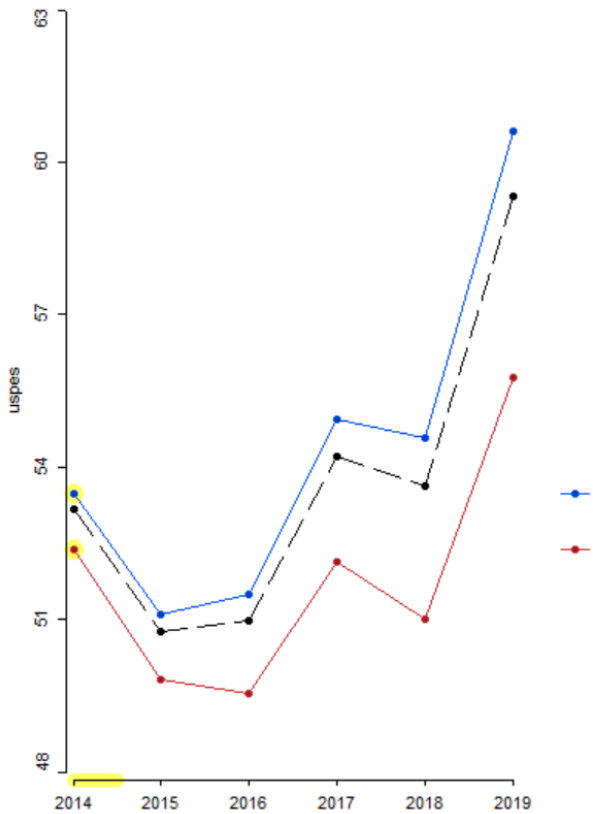
Run Diff-in-Diff Test | Save Dummy

☐ Save Test Results

Group	Obs.	Mean	S.D.
Selected	352	52.36	14.67
Unselected	999	53.47	12.17

Do Means Differ? (ANOVA)

D.F.	1349
F-val	1.93
p-val	0.166



Variable: uspes (2014-2019)

Groups: Selected vs. Unselected

Difference-in-Means Test:

Group 1: Selected Period 1: 2019

Group 2: Unselected Period 2: 2019

Run Diff-in-Diff Test | Save Dummy

☐ Save Test Results

Group	Obs.	Mean	S.D.
Selected	352	55.76	15.41
Unselected	999	60.59	11.95

Do Means Differ? (ANOVA)

D.F.	1349
F-val	36.24
p-val	0.000

Klastrová analýza (K means clusters)

Method: KMeans
Number of clusters: 4
Initialization method: KMeans++
Initialization re-runs: 150
Maximum iterations: 1000
Transformation: Standardize (Z)
Distance function: Euclidean

Cluster centers:

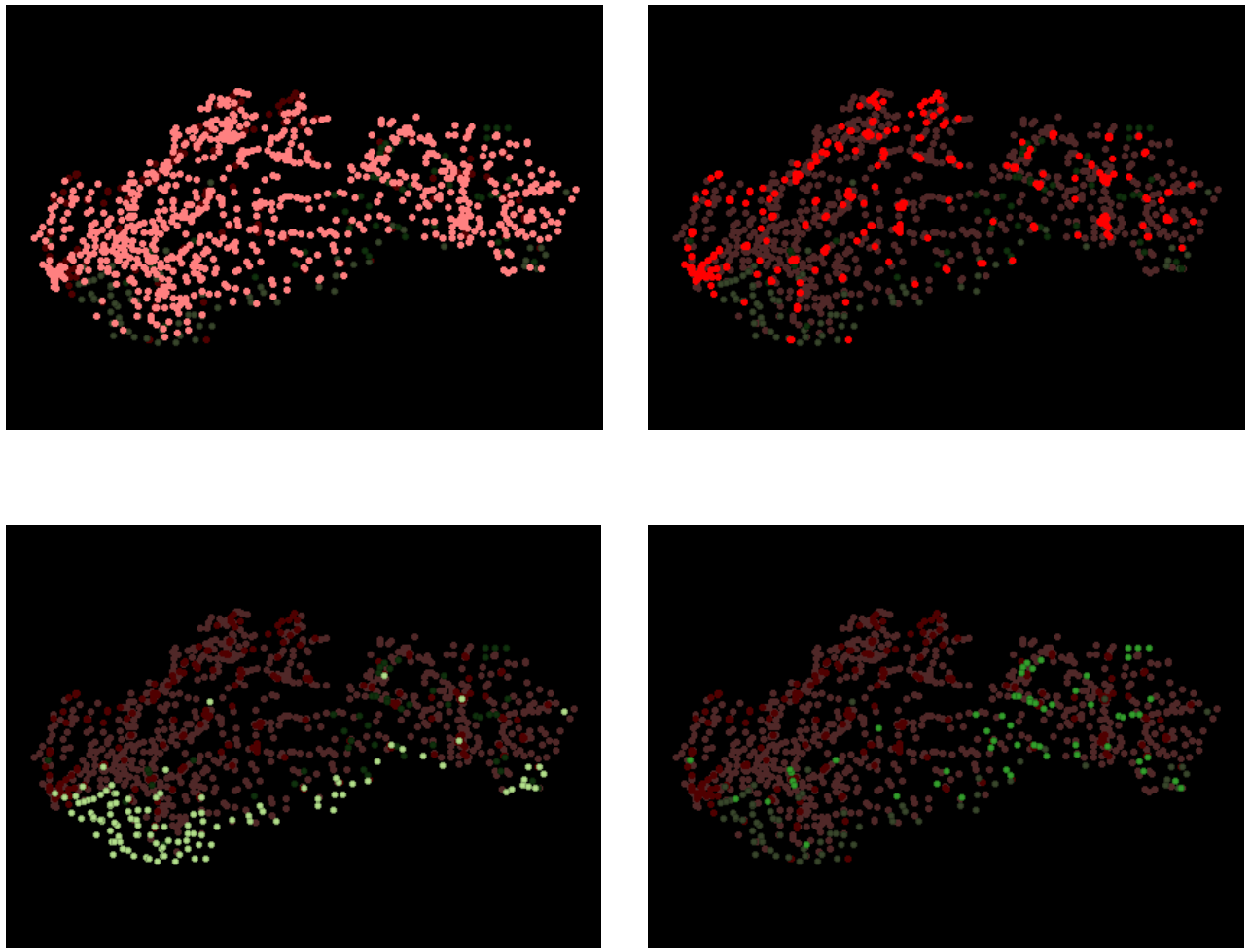
	jazyk_iny_	N_19	uspes (2019)	pNSZP_19
C1	0	18.4842	58.4874	1.39154
C2	0	51.044	67.0396	0.278081
C3	1	19.1066	51.3959	3.11224
C4	0.155172	13.2241	28.6379	58.396

The total sum of squares: 5508

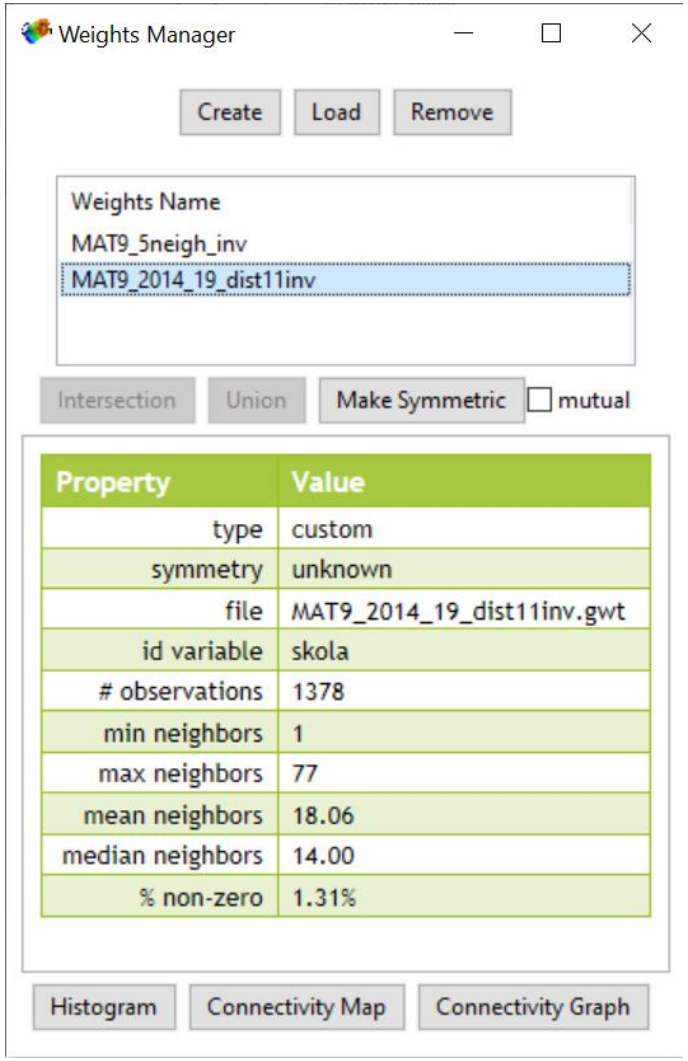
Within-cluster sum of squares:

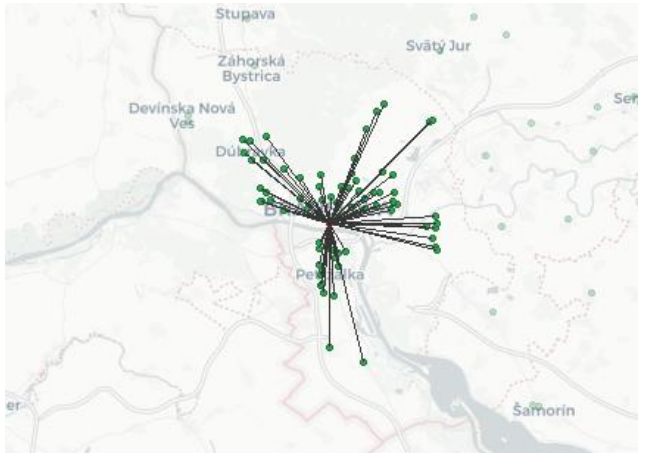
	Within cluster S.S.
C1	982.85
C2	344.775
C3	236.143
C4	343.404

The total within-cluster sum of squares: 1907.17
The between-cluster sum of squares: 3600.83
The ratio of between to total sum of squares: 0.653745

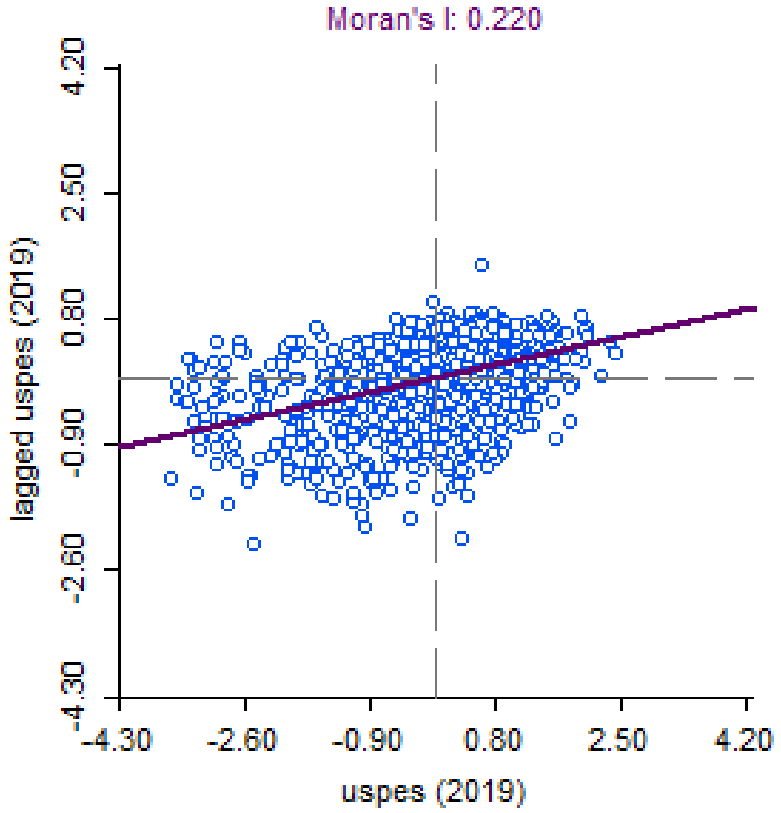


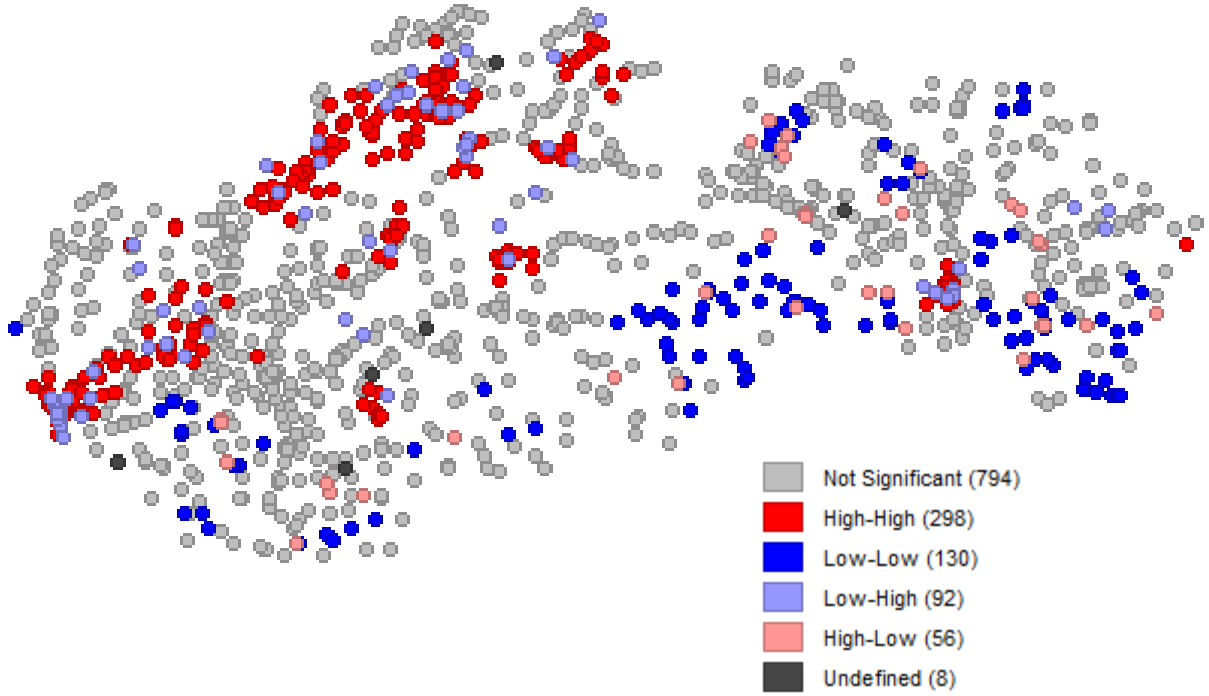
Priestorové váhy



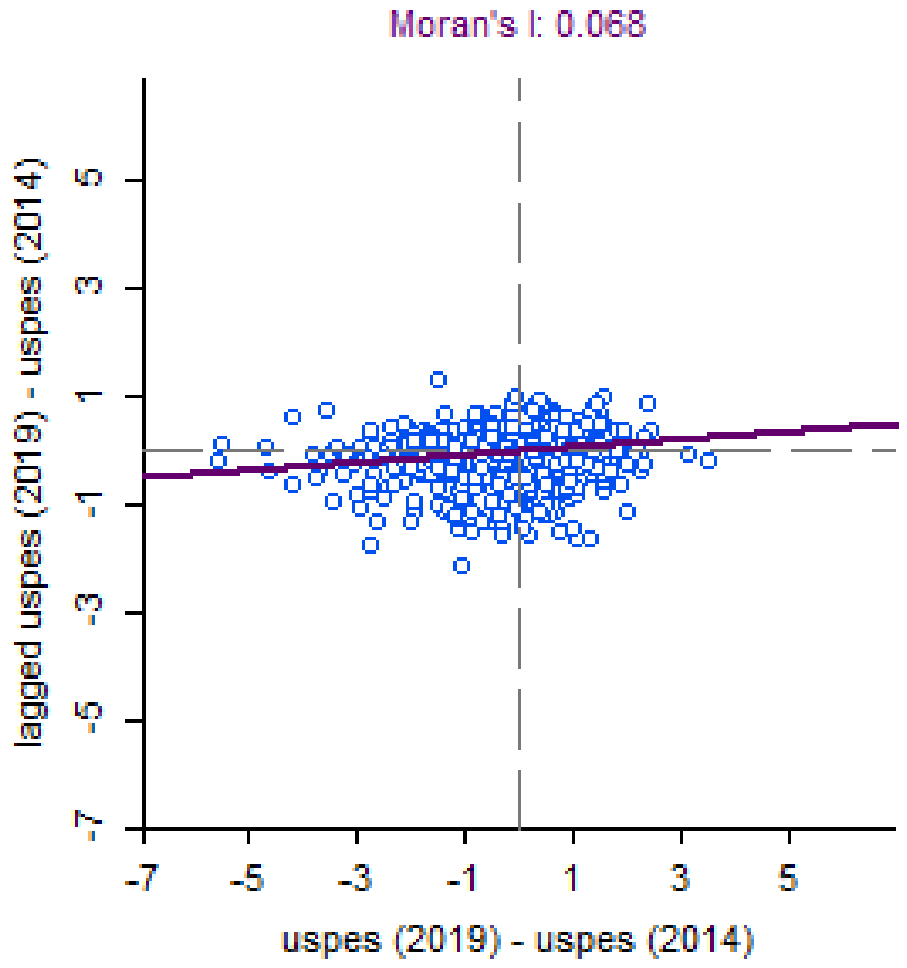


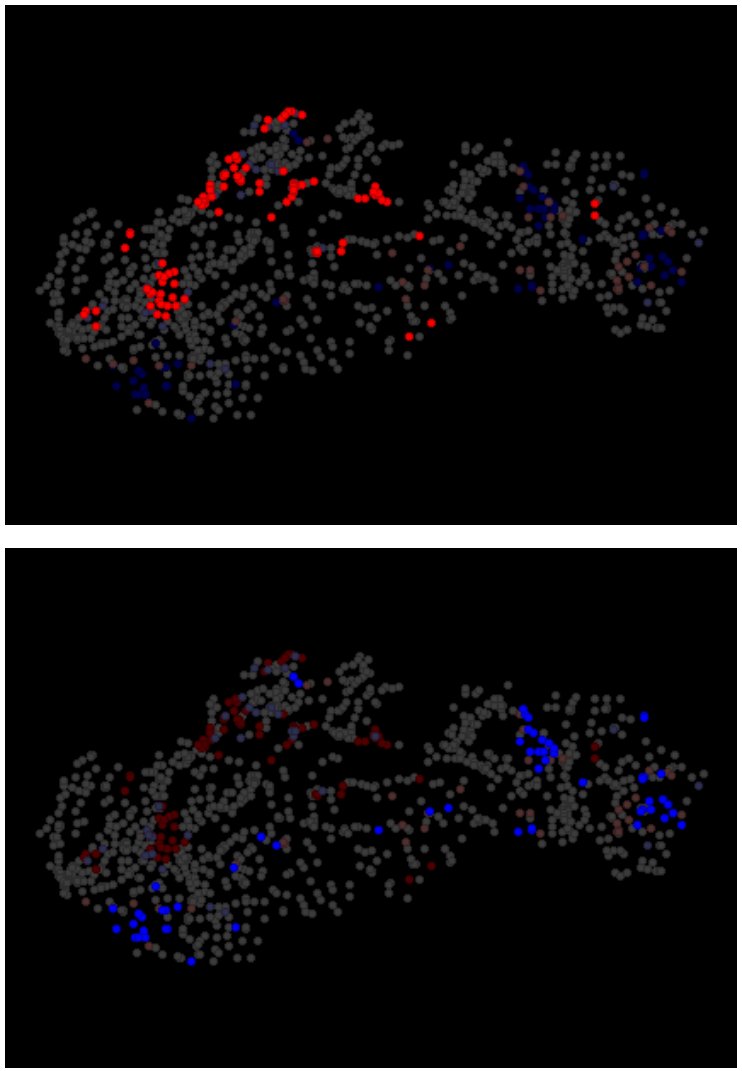
Priestorová autokorelácia – Moranovo I a LISA





Priestorová autokorelácia – Differential LISA





Regresia

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  MAT9_2014_19_FINAL
Dependent Variable  :uspes (2019)  Number of Observations: 1370
Mean dependent var  :     59.0624  Number of Variables   :    8
S.D. dependent var  :        13.5  Degrees of Freedom    : 1362

R-squared           :    0.487527  F-statistic           :       185.1
Adjusted R-squared  :    0.484893  Prob(F-statistic)     :           0
Sum squared residual:      127955  Log likelihood        :     -5051.7
Sigma-square        :     93.9461  Akaike info criterion :     10119.4
S.E. of regression  :     9.69258  Schwarz criterion     :     10161.2
Sigma-square ML     :     93.3975
S.E of regression ML:     9.66424

--------------------------------------------------------------------------
      Variable      Coefficient      Std.Error    t-Statistic   Probability
--------------------------------------------------------------------------
          CONSTANT       38.1917        1.27564        29.9394     0.00000
              N_19      0.101498      0.0151489        6.70001     0.00000
           pNSZP_19     -0.404022      0.0212796       -18.9863     0.00000
             pNC_19    -0.0994534       0.169592      -0.586426     0.55769
            pNZZ_19    -0.0884409      0.0214993       -4.11366     0.00004
         jazyk_iny_       -5.33256       0.915046       -5.82764     0.00000
           sukromna        6.42404        2.12481        3.02335     0.00255
      uspes (2014)       0.395372      0.0216829        18.2343     0.00000
--------------------------------------------------------------------------
```

Priestorová regresia

```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : MAT9_2014_19_FINAL
Spatial Weight      : MAT9_2014_19_dist1linv
Dependent Variable  :uspes (2019)  Number of Observations: 1370
Mean dependent var  :     59.0624  Number of Variables   :    9
S.D. dependent var  :        13.5  Degrees of Freedom    : 1361
Lag coeff.   (Rho)  :    0.297815

R-squared           :    0.517180  Log likelihood        :    -5017.06
Sq. Correlation     : -            Akaike info criterion :     10052.1
Sigma-square        :     87.9934  Schwarz criterion     :     10099.1
S.E of regression   :     9.38048
```

Variable	Coefficient	Std.Error	z-value	Probability
W_uspes (201	0.297815	0.0344023	8.65682	0.00000
CONSTANT	22.3344	2.20587	10.125	0.00000
N_19	0.0995038	0.0146626	6.78622	0.00000
pNSZP_19	-0.376682	0.0209277	-17.9991	0.00000
pNC_19	-0.129946	0.164261	-0.791095	0.42889
pNZZ_19	-0.103014	0.0208095	-4.95031	0.00000
jazyk_iny_	-3.5053	0.904342	-3.87607	0.00011
sukromna	5.62912	2.05752	2.73588	0.00622
uspes (2014)	0.360711	0.0214154	16.8436	0.00000

Záver z ilustratívneho príkladu

- existujú významné rozdiely v kvalite škôl (žiakov), problémom je najmä veľký počet extrémne nekvalitných škôl
- nekvalitné školy sú najmä na juhu stredného a na východnom Slovensku
- tento problém je dlhodobý – najmä menej kvalitné školy ostávajú pozadu dlhodobo a problém sa zväčšuje v čase
- nekvalita je spojená s vyšším podielom detí zo sociálne znevýhodneného prostredia, zdravotne znevýhodnení, s iným jazykom, a štátna menšou veľkosťou tried, na východnom Slovensku
- Identifikovali sme školy, ktoré majú nízku úspešnosť v MAT sú malé a majú veľký podiel NZSP
- Identifikovali sme regióny kde sú koncentrované dobré a zlé školy a regióny, kde sa situácia zhoršuje a zlepšuje

Ďakujem za pozornosť

stefan.rehak@euba.sk