EASYSTAT 1.0 – Uživatelský manuál

Josef Novotný, Vojtěch Nosek, Karel Jelínek
Kontakt: pepino@natur.cuni.cz
Přírodovědecká fakulta Univerzity Karlovy v Praze

OBSAH

1. Úvod

EasyStat je volně dostupný program fungující v operačním systému Windows. Umožňuje výpočty vybraných statistických indikátorů, které nejsou zahrnuty do nejběžněji používaných tabulkových procesorů. Verze 1.0 obsahuje výpočty:

- Vybraných měr variability (Variačního koeficientu, Giniho koeficientu a Theilova koeficientu) v jejich nevážené (obvyklé), ale i vážené formě.
- Rozkladu Theilova koeficientu na jeho mezi-skupinovou a vnitro-skupinovou složku včetně očištění o stochastickou (náhodnou) složku.
- Lokalizačních kvocientů z matice dat.
- Upraveného Jaccardova indexu a Upraveného Diceho indexu, jakožto indikátorů prostorové příbuznosti jevů odvozované na základě sledování jejich společných výskytů v regionech určitého územního systému.

Při použití aplikace prosím uvádějte následující odkaz:

Novotný, J., Nosek, V., Jelínek, K. (2014): EasyStat. Přírodovědecká fakulta UK, Praha, dostupné na: http://web.natur.cuni.cz/~pepino/EasyStat.zip

2. Spuštění programu, načtení dat, volba počítaných indikátorů

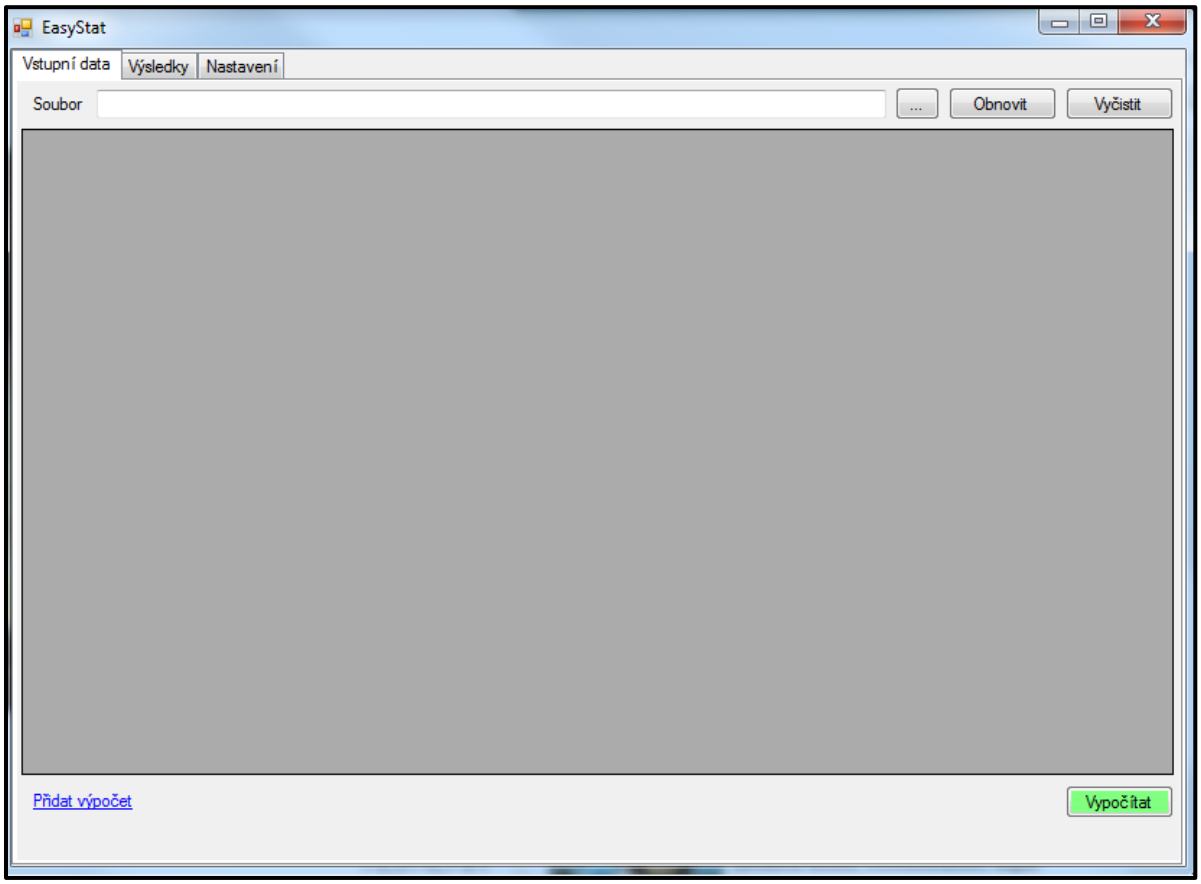
Program je k dispozici ke stažení na adrese: http://web.natur.cuni.cz/~pepino/EasyStat.zip.

Staženou zip složku je nutno rozbalit na disk počítače. Obsahuje několik souborů (včetně manuálu), které musí pro správné fungování programu zůstat uloženy ve společném adresáři. Zde je též spouštěcí soubor **EasyStat.exe**. Po kliknutí na něj se objeví výchozí okno programu (Obrázek 1). Další instalace programu do počítače není potřebná.

Program EasyStat **načítá vstupní data z CSV souborů** („csv – oddělený středníkem“ při ukládání z MS Excel). Ve vstupním souboru je zapotřebí mít **jednotlivé proměnné ve sloupcích** (při výpočtech indikátorů podobnosti v matici, viz dále). První řádek může obsahovat popisky proměnných, v opačném případě jsou proměnné po načtení dat pojmenovány automaticky písmeny dle abecedy.

Načtení vstupního csv souboru z daného adresáře počítače se provádí zadáním cesty k tomuto souboru kliknutím na tlačítko označené třemi tečkami [...] vpravo nahoře vedle řádku, kde se poté zobrazí cesta k danému souboru. Vedlejší tlačítka *Obnovit* a *Vyčistit* slouží k aktualizaci zdrojového souboru, resp. k jeho odstranění z programu EasyStat (nikoliv z příslušného adresáře počítače).

Obrázek 1: Výchozí okno programu EasyStat po spuštění (záložka Vstupní data)

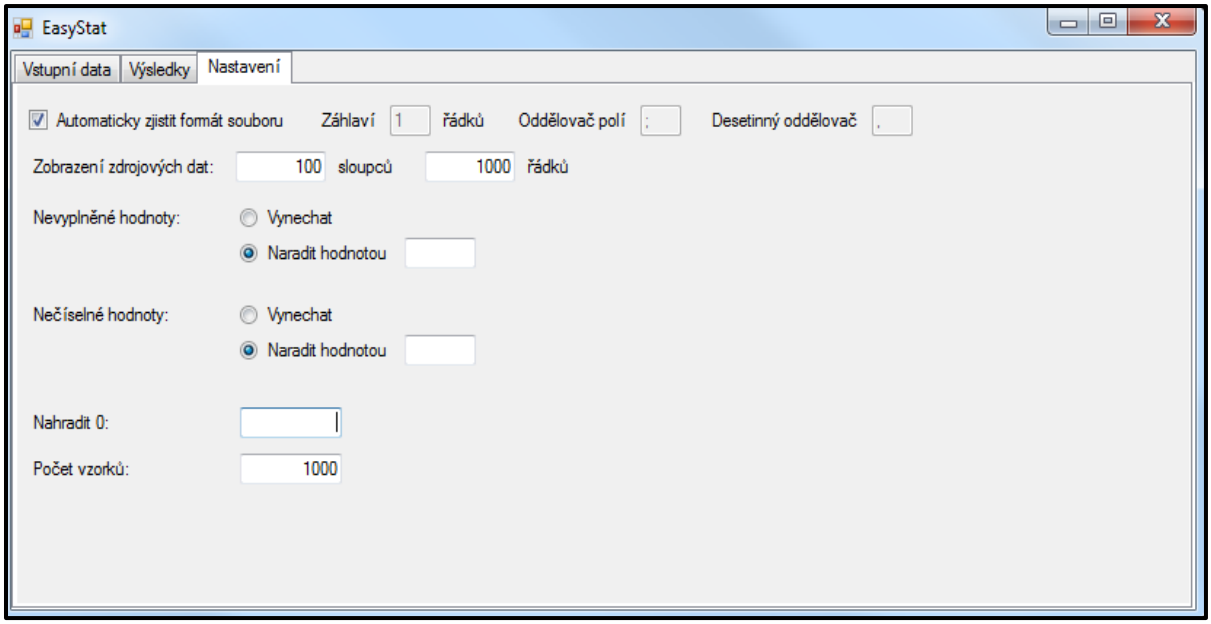


Před načtením dat je možné upravit formát načítaného souboru v záložce *Nastavení* (viz Obrázek 2). Mimo jiné je zde možno specifikovat způsob nahrazování nevyplněných a nečíselných hodnot vstupních dat. Základní nastavení nahrazuje nevyplněné a nečíselné hodnoty nulami.

Pokud mají proměnné (sloupce dat) v jednom souboru různý počet pozorování (řádků), dojde proto při zachování základního nastavení k automatickému doplnění nul do každého sloupce dle rozsahu proměnné s největším počtem pozorování. Pokud chceme, aby vypočtené indikátory reflektovaly skutečný počet pozorování u jednotlivých proměnných, je třeba v záložce Nastavení zvolit možnost vynechání nevyplněných hodnot.

V záložce *Nastavení* je také možno zvolit rozsah zobrazených dat načteného souboru (základní nastavení je 100 sloupců a 1 000 řádků). Je možné a v případech větších souborů dat i vhodné zobrazovat pouze část výchozího souboru dat. Toto nemá vliv na výpočty, neboť program pracuje vždy s úplným souborem výchozích dat, bez ohledu na zobrazený rozsah.

Obrázek 2: Záložka „Nastavení“



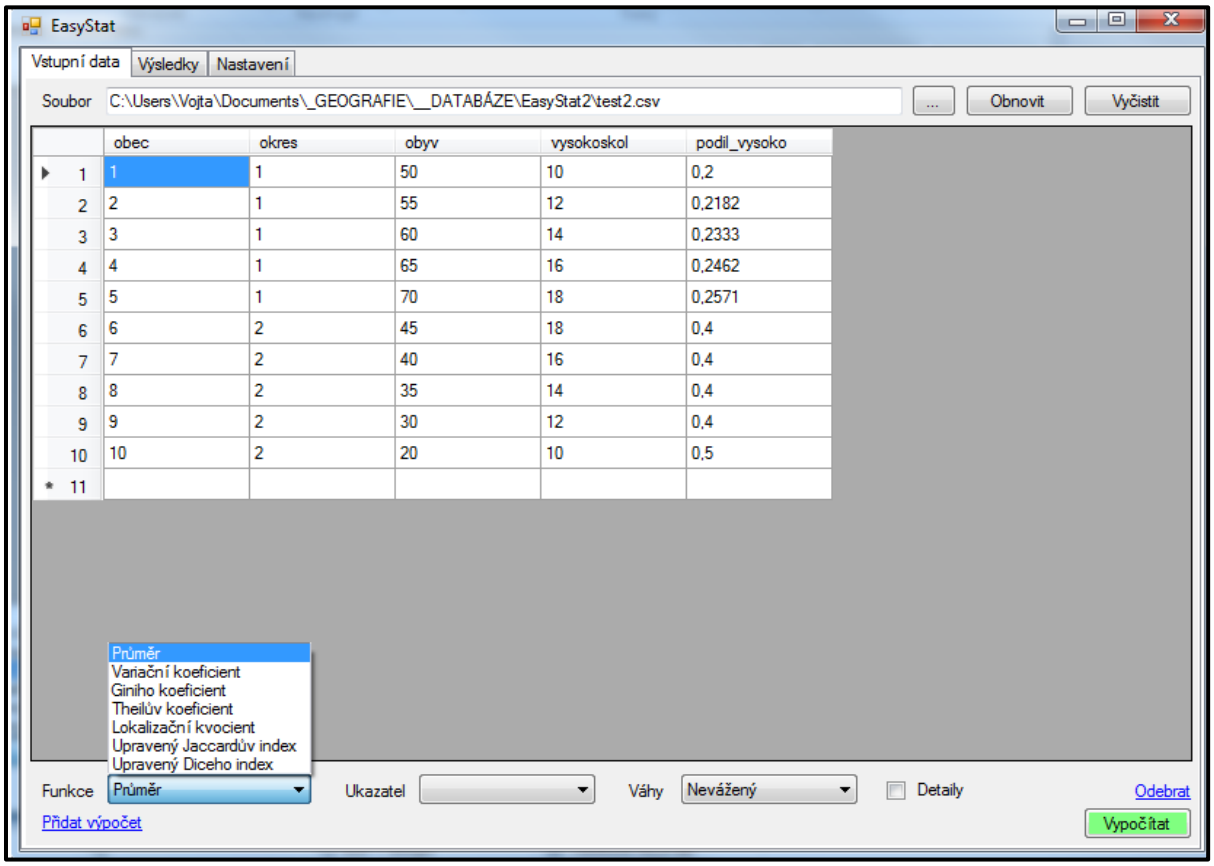
Volba indikátorů (funkcí) pro výpočty se provádí výběrem z nabídky indikátorů ve spodní části titulní záložky *Vstupní data* u popisku *Funkce* (Obrázek 3).

Program umožňuje současné výpočty více indikátorů – pro každou další volbu je třeba kliknout na text *Přidat výpočet* (vlevo dole).

Vysvětlení parametrů, zadávaných k jednotlivým indikátorům, je níže.

Volba indikátorů (funkcí) pro výpočty se provádí výběrem z nabídky indikátorů ve spodní části titulní záložky *Vstupní data* u popisku *Funkce* (). Program umožňuje současné výpočty více indikátorů – pro každou další volbu je třeba kliknout na text *Přidat výpočet* (vlevo dole). Vysvětlení parametrů, zadávaných k jednotlivým indikátorům, je níže.

Obrázek 3: Volba počítaných indikátorů



3. Dostupné statistiky

3.1. Průměr a míry variability - Variační koeficient, Giniho koeficient, Theilův koeficient v jejich nevážené a vážené formě

Popis indikátoru: Z popisných statistik je k dispozici výpočet průměru, ve vážené i nevážené podobě. Pro ***měření míry nerovnoměrnosti*** jsou k dispozici variační koeficient, Giniho koeficient a Theilův koeficient.

Vstupní data: Matice dat, ve které sloupce označují sledované proměnné, u nichž sledujeme míru nerovnoměrnosti nebo průměr, a řádky jednotlivá pozorování. Theilův index neumí pracovat s nulovými hodnotami, v Nastavení je proto nutné nastavit *Nahradit 0* hodnotou blízkou nule.

Výstupy: Výstupem jsou jednočíselné hodnoty zvolených statistik.

Volba parametrů: Kromě zvoleného ukazatele lze volit *Váhy*. Je možné volit mezi *Neváženy* (všechna pozorování mají stejnou váhu) nebo jakýmkoliv jiným číselným ukazatelem ze vstupních dat (jednotlivá pozorování vážena podle zvoleného ukazatele).

Příklad použití: Průměr je základní popisnou statistikou, variační, Giniho a Theilův koeficient měří míru nerovnoměrnosti sledovaného ukazatele. Buď sledujeme rozdíly mezi prostými hodnotami (bez použití vah, jako *Váhy* volíme *Neváženy*), nebo pracujeme s populačně či

jinak váženými hodnotami (jako *Váhy* volíme libovolný číselný ukazatel, typicky počet obyvatel). S neváženou mírou nerovnoměrnosti se setkáme nejčastěji při studiu výkonnosti regionů jako ekonomických jednotek, vážená míra nerovnoměrnosti zachycuje lépe dopad nerovnoměrnosti na obyvatele těchto regionů.

Formální zápis výpočtu (v nevážené i vážené formě):

$$V = \frac{\sqrt{\sum_{i=1}^{k} |y_i - y|^2}}{y};\ V_v = \frac{\sqrt{\sum_{i=1}^{k} \frac{n_i}{n} |y_i - y|^2}}{y}$$

$$G = \frac{1}{k^2} \frac{1}{2y} \sum_{i=1}^{k} \sum_{j=1}^{k} \left(|y_i - y_j| \right);\ G_v = \frac{1}{2y} \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\frac{n_i}{n} \frac{n_j}{n} |y_i - y_j| \right)$$

$$T = \frac{1}{k} \sum_{i=1}^{k} \frac{y_i}{y} \ln \frac{y_i}{y};\ T_v = \left(\sum_{i=1}^{k} \frac{n_i}{n} \frac{y_i}{y} \ln \frac{y_i}{y} \right)$$

Kde V značí variační koeficient, G Giniho koeficient a T Theilův koeficient. Spodní index v značí váženou formu výpočtu; y značí průměr sledovaného ukazatele (při výpočtech vážených forem jde o vážený průměr), y_i a y_j hodnoty jednotlivých pozorování a n_i a n_j jejich váhy n.

3.2. Rozklad Theilova koeficientu a testování statistické významnosti

Popis indikátoru: Rozklad Theilova koeficientu umožňuje ***rozdělit míru nerovnoměrnosti*** (změřenou pomocí Theilova koeficientu) beze zbytku ***na mezi-skupinovou a vnitro-skupinovou složku***. Rozklad tak umožňuje odlišit, jak velká část nerovnoměrnosti může být přisouzena rozdílům uvnitř skupin nebo mezi skupinovými průměry (např. Novotný 2007). Výpočty lze očistit o stochastickou (náhodnou) složku nerovnoměrnosti (viz Novotný a Nosek 2012).

Vstupní data: Matice dat, ve které sloupce označují sledované proměnné, u nichž sledujeme míru nerovnoměrnosti, a řádky jednotlivá pozorování. Minimálně jeden ze sloupců označuje rozřazení jednotlivých pozorování do skupin. Skupiny musí být definované číselnými hodnotami. Theilův index neumí pracovat s nulovými hodnotami, v Nastavení je proto nutné nastavit *Nahradit 0* hodnotou blízkou nule.

Výstupy: Program vypočte celkový Theilův koeficient, jeho mezi-skupinovou a vnitro-skupinovou složku a podíly jednotlivých částí. Při použití *Převzorkování* je vedle hodnoty mezi-skupinové složku uváděna také hodnota očištěná o stochastickou (náhodnou) složku (viz Novotný a Nosek 2012). Součet mezi-skupinové a vnitro-skupinové složky je vždy roven 100 %.

Volba parametrů: Zvolen musí být *Ukazatel*, proměnná, pro kterou je vypočítávána míra nerovnoměrnosti. Pro výpočet rozkladu je nutné zvolit *Region*, ukazatel, podle kterého je hodnota Theilova koeficientu rozdělena na mezi-skupinovou a vnitro-skupinovou složku. Dále lze vybrat *Váhy*. Je možné volit mezi *Neváženým* (všechna pozorování mají stejnou váhu) nebo jakýmkoliv jiným číselným ukazatelem ze vstupních dat (jednotlivá pozorování vážena

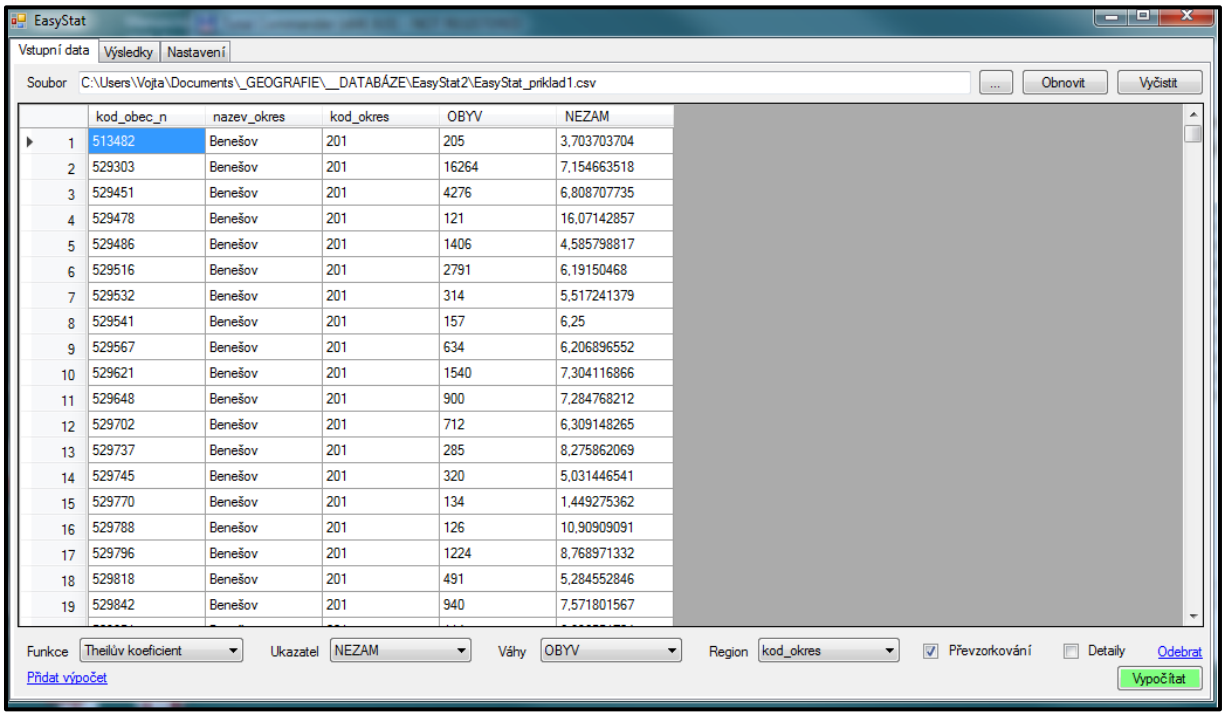
podle zvoleného ukazatele). Výsledky lze očistit o stochastickou složku pomocí zaškrtnutí pole *Převzorkování*. Stochastická složka je vypočtena po náhodném přeskupení jednotlivých pozorování jako průměr z n-opakování. Počet opakování lze nastavit libovolně v záložce *Nastavení – Počet vzorků*. Jako postačující hodnota se většinou uvádí 1000 opakování. Porovnáním hodnot bez převzorkování a s převzorkováním lze také testovat statistickou významnost.

Příklad použití: Výpočet rozkladu míry nerovnoměrnosti obcí Středočeského kraje s rozkladem míry nerovnoměrnosti na mezi-okresní (mezi okresními průměry) a vnitro-okresní složku (mezi obcemi uvnitř okresů). Vstupem jsou data za jednotlivé obce, okres, ve kterém se nachází, počet obyvatel a míra nezaměstnanosti. Jako ukazatel zvolíme NEZAM (míru nezaměstnanosti), budeme vážit podle počtu obyvatel (OBYV) a za skupinu zvolíme okres (kod_okres). Výsledky očistíme o stochastickou složku zaškrtnutím *Převzorkování*. Viz Obrázek 4.

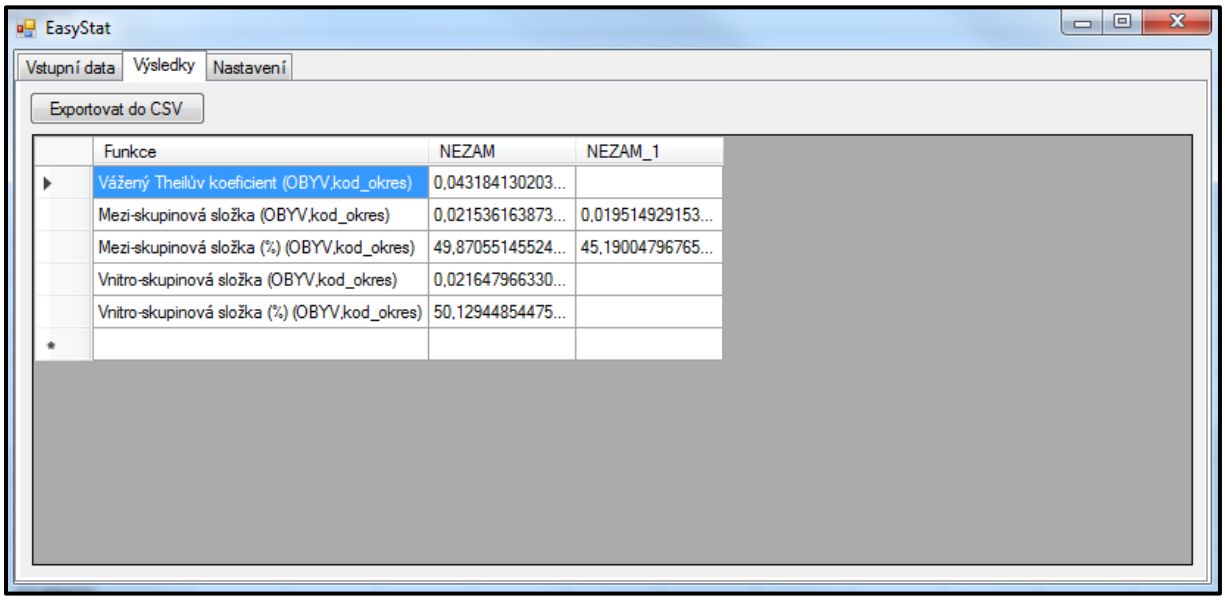
Výsledný podíl mezi-skupinové složky je roven cca 49,9 %, po odečtení stochastické složky cca 45,2 %. Většina nerovnoměrnosti (cca 55 %) mezi obcemi Středočeského kraje se tak nachází uvnitř okresů. Výsledky viz Obrázek 5.

Pro konkrétní příklady použití viz také Nosek a Netrdová (2014) nebo Netrdová a Nosek (2015).

Obrázek 4: Zadání dat pro výpočet rozkladu Theilova indexu



Obrázek 5: Zobrazení dat výsledků rozkladu Theilova indexu



Formální zápis výpočtu: $T_C = \left(\sum_{j=1}^{k} \frac{n_j}{n} \frac{y_j}{y} \ln \frac{y_j}{y} \right) + \left(\sum_{j=1}^{k} \frac{n_j}{n} \frac{y_j}{y} \sum_{i=1}^{n_j} \frac{y_{ij}}{y_j} \ln \frac{y_{ij}}{y_j} \right) = T_B + T_W$

Kde T_B je mezi-skupinová složka celkové míry nerovnoměrnosti T_C a T_W vnitro-skupinová složka. Podíl mezi-skupinové složky odpovídá podílu T_B/T_C; y značí průměr sledovaného ukazatele, y_i a y_j hodnoty jednotlivých pozorování a n_i a n_j jejich váhy n; y_{ij} značí průměrnou hodnotu sledovaného jevu v *i-té* jednotce ve skupině *j*.

3.3. Lokalizační kvocient

Popis indikátoru: Lokalizační kvocient ukazuje ***míru koncentrace*** určitého jevu v určitém regionu. Poměřuje relativní četnost výskytu tohoto jevu v daném regionu s relativní četností tohoto jevu v územním systému vyššího řádu.

Vstupní data: Matice zdrojových dat, ve které sloupce označují jevy, u nichž sledujeme míru koncentrace, a řádky označují regiony.

Výstupy: Program vypočte lokalizační kvocient pro každou buňku dané matice. Výsledkem je proto matice shodného rozsahu jako zdrojová data, přičemž v každé z buněk je vypočten lokalizační kvocient pro daný jev a region.

Volba parametrů: Tento indikátor nemá žádné další parametry.

Příklad: Výpočet lokalizačních kvocientů pro skupiny imigrantů v regionech Česka. Vstupem je matice dat, kde sloupce označují jednotlivé skupiny imigrantů a řádky regiony Česka. Pokud žije v populaci určitého regionu např. 1 % imigrantů z Ukrajiny, zatímco v celé populaci dané země žije pouze 0,5 % imigrantů z Ukrajiny, je lokalizační kvocient koncentrace Ukrajinců v daném regionu roven 2. Obdobně pro další skupiny imigrantů a regiony.

Formální zápis výpočtu:

$$LQ_{i,r} = \frac{F_{i,r} \Big/ \sum_i F_{i,r}}{\sum_r F_{i,r} \Big/ \sum_i \sum_r F_{i,r}}$$

Kde $LQ_{i,r}$ je lokalizační kvocient vypočtený pro určitou skupinu (jev) i a region r a $F_{i,r}$ je relativní četnost této skupiny i v regionu r.

3.4. Míry prostorové příbuznosti – Upravený Jaccardův index, Upravený Diceho index

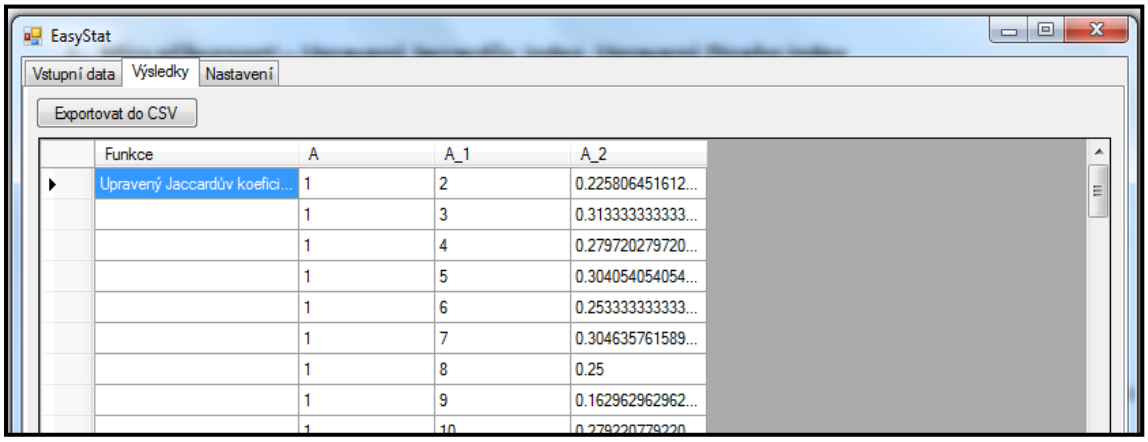
Popis indikátorů: Oba tyto indikátory měří tzv. ***prostorovou příbuznost*** dvou jevů na základě vyhodnocení frekvence společných koncentrací těchto jevů v regionech (každý z indikátorů provádí toto vyhodnocení trochu odlišným způsobem). Existence či neexistence koncentrace daného jevu v daném regionu je přitom měřena pomocí lokalizačního koeficientu.

Vstupní data: Matice zdrojových dat, ve které sloupce označují jevy, jejichž vzájemnou prostorovou příbuznost chceme vyhodnotit a řádky označují regiony.

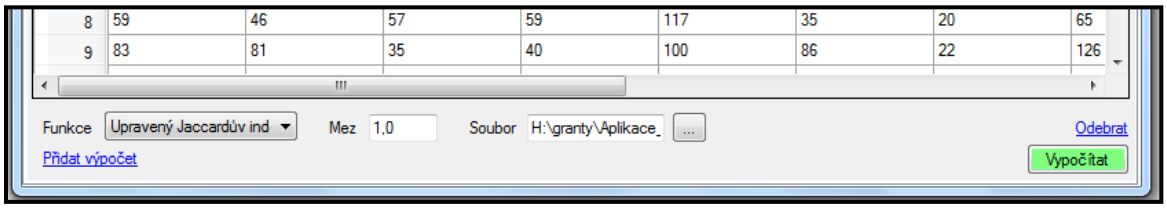
Výstupy: Program vypočítá míry daného indikátoru prostorové příbuznosti (*Upravený Jaccardův index* nebo *Upravený Diceho index*) pro všechny dvojice jevů (tzn. sloupců) ve vstupní matici. Výstupem je tedy vždy soubor n*(n-1)/2 výsledků, kde n = počet sloupců dat. U souboru s méně sloupci je výsledky možno zobrazit přímo v záložce *Výsledky* (Obrázek 6). V takovém případě je ale nutno nezadávat výstupní *Soubor* při volbě parametrů indikátoru (Obrázek 7). Pokud cestu k výstupnímu csv souboru zadáme, uloží se výsledky do něj.

Zobrazený výstup na Obrázku 6 má popisek indikátoru a tři sloupce dat, přičemž první dva označují pomocí pořadových čísel sloupců vstupního souboru dvojici sledovaných jevů a ve třetím sloupci je hodnota zvoleného indikátoru prostorové příbuznosti. Výsledky zde jsou zapsány obdobným způsobem (tj. v jednotlivých řádcích jsou pořadová čísla dané dvojice proměnných a hodnota zvoleného indikátoru), přičemž položky v jednotlivých řádcích jsou odděleny čárkou (z tohoto souboru lze jednoduše načíst výsledná data např. do tabulky MS Excel pomocí funkce „import dat z textu“)

Obrázek 6: Zobrazení výsledků při výpočtech indikátorů prostorové příbuznosti



Obrázek 7: Volba parametrů při zadávání výpočtů indikátorů prostorové příbuznosti



Volba parametrů: Při zadání indikátorů prostorové příbuznosti je vedle parametru *Soubor* (viz výše) možno stanovit parametr *Mez*. Ten udává hodnotu lokalizačního kvocientu, kterou je definována koncentrace sledovaných jevů v jednotlivých regionech. Defaultně je nastavena mez rovna 1,0.

Příklad použití: Výpočet indikátorů prostorové příbuznosti skupin imigrantů v Česku. Vstupem je matice dat, kde sloupce označují jednotlivé skupiny imigrantů (např. Ukrajince, Vietnamce, Slováky, Němce) a řádky regiony Česka. Indikátory prostorové příbuznosti (Upravený Jaccardův index nebo Upravený Diceho index) je možno použít k výpočtům vzájemné prostorové příbuznosti mezi dvojicemi těchto skupin. Výpočty prostorové příbuznosti mohou být z různých důvodů užitečné, neboť prostorová příbuznost velmi často odráží i jiné aspekty příbuznosti sledovaných jevů. Pro konkrétní příklady použití těchto indikátorů viz Novotný, Cheshire (2012); Novotný, Hasman (2015).

Formální zápis výpočtu: Vstupní data jsou popsány maticí udávající počty zástupců určitých skupin *i, j*... (sloupce) v *r* regionech (řádky). Nejprve je pro všechny skupiny a regiony vypočtena matice lokalizačních kvocientů *LQ* (viz výše). Na základě stanovené meze *LQ* (defaultně 1,0) je stanoveno, zda se daná skupina koncentruje v regionu r (tj. zda $LQ > 1$) nebo ne. Pro výpočet indikátorů prostorové příbuznosti mezi skupinami *i* a *j* je pak porovnávána množina regionů, ve kterých se tyto skupiny koncentrují – formálně označujeme tyto množiny regionů jako $\{r: LQ_{i,r} > 1\}$ a $\{r: LQ_{j,r} > 1\}$.

Upravený Jaccardův index prostorové příbuznosti pak vypočteme jako:

$$J_{i,j} = \frac{\left|\{r : LQ_{i,r} > 1\} \cap \{r : LQ_{j,r} > 1\}\right|}{\left|\{r : LQ_{i,r} > 1\} \cup \{r : LQ_{j,r} > 1\}\right|}$$

Pro výpočet upraveného Diceho indexu je třeba nejdříve vypočíst jeho „nesymetrické" varianty. První nesymetrický Diceho index prostorové příbuznosti dvou skupin *i* a *j* vypočteme jako podmíněnou pravděpodobnost, že skupina *i* je koncentrována v regionu *r*, je-li v daném regionu koncentrována také skupina *j*:

$$D^1_{i|j} = P\left(LQ_{i,r} > 1 \mid LQ_{j,r} > 1\right) = \frac{\left|\{r : LQ_{i,r} > 1\} \cap \{r : LQ_{j,r} > 1\}\right|}{\left|\{r : LQ_{j,r} > 1\}\right|}$$

Druhý nesymetrický Diceho index pak analogicky zachycuje podmíněnou pravděpodobnost, že skupina *j* je koncentrována v regionu *r*, je-li v daném regionu koncentrována také skupina *i*:

$$D^2_{j|i} = P\left(LQ_{j,r} > 1 \mid LQ_{i,r} > 1\right) = \frac{\left|\left\{r : LQ_{i,r} > 1\right\} \cap \left\{r : LQ_{j,r} > 1\right\}\right|}{\left|\left\{r : LQ_{i,r} > 1\right\}\right|}$$

Upravený symetrický Diceho index prostorové příbuznosti odpovídá menší z jeho výše uvedených asymetrických variant:

$$D_{i,j} = \min\left(D^1_{i|j}; D^2_{j|i}\right)$$

4. Použitá literatura

NETRDOVÁ, P., NOSEK, V. (2015): Spatial patterns of unemployment in Central Europe: emerging development axes beyond the blue banana, *Journal of Maps*, přijato k publikaci.

NOSEK, V., NETRDOVÁ, P. (2014): Measuring spatial aspects of variability. Comparing spatial autocorrelation with regional decomposition in international unemployment research. *Historical Social Research*, 39, 2, 292-314.

NOVOTNÝ, J. (2007): On the measurement of regional inequality: does spatial dimension of income inequality matter? *The Annals of Regional Science*, 41, 3, 563-580.

NOVOTNÝ, J., HASMAN, J. (2015): The emergence of regional immigrant concentrations in USA and Australia: a spatial relatedness approach. *PLoS ONE*, 10(5): e0126793.

NOVOTNÝ, J., CHESHIRE, J. (2012): The surname space of the Czech Republic: examining population structure by network analysis of spatial co-occurrence of surnames. *PLoS ONE*, 7(10), e48568.

NOVOTNÝ, J., NOSEK, V. (2012): Comparison of regional inequality in unemployment among four Central European countries: an inferential approach. *Letters in Spatial and Resource Sciences*, 5, 2, 95-101.